



HAL
open science

Détection de l'activité des éléments transposables chez les plantes cultivées : étude du mobilome par la caractérisation du compartiment extrachromosomique

Sophie Lanciano

► **To cite this version:**

Sophie Lanciano. Détection de l'activité des éléments transposables chez les plantes cultivées : étude du mobilome par la caractérisation du compartiment extrachromosomique. Génétique des plantes. Université Montpellier, 2017. Français. NNT : 2017MONTT129 . tel-01707480

HAL Id: tel-01707480

<https://theses.hal.science/tel-01707480>

Submitted on 12 Feb 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



THÈSE

Pour obtenir le grade de
Docteur

Délivré par
Université de Montpellier

Préparée au sein de l'école doctorale GAIA
Et de l'unité de recherche Institut de Recherche pour le
développement – UMR DIADE

Spécialité : **BIDAP – Biologie, Interactions, Diversité
Adaptative des Plantes**

Présentée par **Sophie LANCIANO**

**Détection De L'Activité Des Éléments
Transposables Chez Les Plantes Cultivées :
Étude Du Mobilome Par La Caractérisation
Du Compartiment Extrachromosomique**

Soutenue le 10 novembre 2017 devant le jury composé de



Marie Angèle Grandbastien – Directeur de recherche, INRA de Versailles	Rapporteur
Gaël Cristofari – Directeur de recherche, INSERM de Nice	Rapporteur
Anne Roulin – Chargée de recherche, Université de Zurich	Examineur
Claude Becker – Directeur de recherche, Institut Gregor Mendel	Examineur
Etienne Bucher – Directeur de recherche, INRA d'Angers	Examineur
Pierre Czernic – Professeur, Université de Montpellier	Examineur
Marie Mirouze – Chargée de recherche, IRD Montpellier	Directrice de thèse
Alain Ghesquière – Directeur de recherche, IRD Montpellier	Directeur de thèse

Résumé

Les éléments transposables (ET) sont des éléments génétiques ubiquitaires et potentiellement mobiles dans les génomes eucaryotes. Les génomes hôtes ont développé des mécanismes épigénétiques pour contrôler et prévenir la prolifération des ET. Néanmoins, certains ET semblent capables de s'activer en réponses à des stress ou à des facteurs développementaux. Les méthodes disponibles pour détecter l'activité transpositionnelle d'un ET sont souvent limitées au stade transcriptionnel ou sont adaptées à des génomes de petite taille. Relativement peu d'ET sont actuellement connus pour être actifs et les mécanismes spécifiques qui les contrôlent ne sont pas clairement identifiés.

Durant mes travaux de thèse, nous avons développé une stratégie de séquençage à haut débit qui permet la détection d'ADN extrachromosomique circulaire (ADNecc) témoignant notamment de l'activité des ET et de la stabilité d'un génome. Ainsi nous avons pu caractériser chez plusieurs espèces le mobilome, défini comme l'ensemble des ADNecc présents dans un tissu.

La technique du mobilome-seq s'est avérée être un outil puissant pour la détection des ET actifs notamment chez le riz asiatique *Oryza sativa*. Notre analyse du mobilome a permis l'identification d'un rétrotransposon *PopRice* actif dans l'albumen (tissu nourricier du grain) chez différentes variétés de riz. Pour la première fois chez les plantes, nous avons également détecté des insertions somatiques d'ET par re-séquençage de génome entier. À partir de nos résultats, nous avons combiné nos données mobilomiques avec une analyse GWAS pour proposer des pistes afin d'identifier de nouveaux mécanismes de régulation de cet élément.

En parallèle, nous avons appliqué la technique du mobilome-seq à différents organismes animaux et végétaux révélant ainsi des spécificités de mobilome propre à chaque espèce. Nos travaux en collaboration avec d'autres équipes ont notamment contribué à préciser le rôle de l'ARN polymérase II dans le contrôle des ET chez *O. sativa* et à mettre en évidence le lien entre la présence d'ADNecc viral et la réponse immunitaire chez *Drosophila melanogaster*.

Mes travaux de thèse ouvrent des perspectives pour l'étude du mobilome, ce répertoire génomique encore largement inexploré et qui se révèle être à la fois une source d'information au niveau des mouvements des ET mais aussi de la stabilité des génomes. L'étude future des mobilomes promet d'apporter des réponses sur notre compréhension de la dynamique des génomes.

Mots clés : ADN extrachromosomique, élément transposable, mobilome, *Oryza sativa*, épigénétique.

Abstract

Transposable elements (TEs) are mobile genetic elements that constitute a major part of eukaryotic genomes. Host genomes have developed epigenetic mechanisms to control and prevent their proliferation. While efficiently silenced by the epigenetic machinery, they can be reactivated upon stress or at precise developmental stages. However, available methods to detect TE activity are often limited to transcriptional level or more adapted to small genomes. Today, only few TEs are known to be active and specific mechanisms controlling TEs are not well defined.

To address this question during my PhD, we developed a strategy of high throughput sequencing that detects extrachromosomal circular DNA (eccDNA) forms which reflect TE activity and genome stability. We characterised mobilomes from different organisms defined as all eccDNA in a cell.

Our mobilome-seq technique successfully identified active TEs especially in asian rice *Oryza sativa*. We identified an active retrotransposon *PopRice* in endosperm tissue from different rice varieties. Interestingly and for the first time in plants, we detected somatic insertions from genome-wide resequencing. We combined our mobilome-seq results with a GWAS analysis to propose new *PopRice* regulation mechanisms.

In a second step, we applied our mobilome seq technique to different animal and plant organisms showing mobilome specificities from each species. Our work in collaboration with different labs help contributed to define role of RNA polymerase II in the control of TEs in *O. sativa* and have revealed a link between presence of eccDNA from virus and immune response in *Drosophila melanogaster*.

Altogether, our mobilome-sequencing method opens the possibility to explore unexplored genomic compartment. Future mobilome analysis represents new possibilities to improve our understanding of dynamics of genomes.

Keywords : extrachromosomal DNA, transposable element, mobilome, *Oryza sativa*, epigenetics.

REMERCIEMENTS

« Jamais auparavant le mot « merci » ne m'a semblé aussi inadéquat. Mais en attendant que quelqu'un trouve un terme plus approprié, cette simple platitude devra faire l'affaire » Lori Nelson Spielman.

Je voudrais tout d'abord remercier tous les membres du jury qui ont accepté d'évaluer mes travaux de thèse. Merci à **Marie-Angèle Grandbastien** et **Gaël Cristofari** de rapporter ce travail. Merci à **Anne Roulin**, **Claude Becker**, **Etienne Bucher** et **Pierre Czernic** d'avoir accepté de faire partie de mon jury.

Je voudrais te remercier tout particulièrement **Etienne** pour avoir suivi cette thèse tout du long, pour tous tes conseils et tes encouragements. J'ai également pris beaucoup de plaisir à travailler sur votre projet avec **Michi**. Et encore merci d'avoir accepté de boucler la boucle et d'examiner pour la dernière fois mes travaux de thèse.

Je tiens à remercier **Emmanuel Guiderdoni** pour avoir suivi mes travaux, pour nos riches discussions et pour ses nombreux conseils et encouragements lors de mes comités de thèse.

Je tiens également à remercier tout particulièrement **Jean-Marc Deragon** pour m'avoir accueillie au sein du laboratoire et pour toutes ces discussions passionnées et passionnantes des repas du midi.

Une thèse ça vous change ~~un homme~~ une femme, non ? Nous y voilà **Marie**, le PDF bientôt dans les bacs, les dernières relectures, les dernières modifications, ... je crois qu'il faut se rendre à l'évidence, ça marque la fin d'une sacrée aventure ! En vrai, je ne sais même pas par où commencer tellement j'ai de raisons de te remercier aujourd'hui. De Cambridge à San Diego en passant par Hanoï, de Nice à Saint Malo en passant par Strasbourg, tu m'as offert tellement d'opportunités, de libertés dans mes recherches, du temps pour m'enseigner et discuter que j'ai beau tourner mes phrases dans tous les sens, quoi que j'écrive, les mots ne suffiront pas à exprimer le plaisir immense que cela a été de réaliser ma thèse à tes côtés. Merci pour m'avoir accordé plus de confiance que je ne m'en accorde moi-même. Merci pour tout le temps que tu m'as consacré. Et enfin, merci d'avoir fait de ton projet, notre projet ! Quand je suis arrivée au labo j'étais persuadée qu'un jour je serais chercheuse. J'ai eu la chance et le plaisir de faire mes premiers pas dans ce monde avec une personne qui représente tout ce que j'aurais voulu que la recherche soit. Malgré mes déceptions face à certaines réalités et la difficulté de certains échecs, j'ai énormément appris de cette expérience et j'en repars grandie. Aujourd'hui je ne sais pas si un jour je serai chercheuse mais si c'est le cas, j'espère que ma recherche sera à l'image de la tienne. Finalement, c'est avec ce mot si « platonique » mais terriblement sincère que je terminerai : MERCI pour tout Marie.

Olivier, quel plaisir cela a été de travailler dans ton équipe ! Merci de m'avoir aussi bien accueillie, écoutée, appris et conseillée. Tu as su répondre présent à chaque fois que j'en ai eu besoin et j'ai pris énormément de plaisir à discuter avec toi. Tu as toujours réussi à me montrer les bons côtés de la recherche quand je ne voyais plus que les mauvais et je t'en remercie. Tu avais raison. D'un

point de vue beaucoup moins scientifique, merci de nous avoir fait naviguer dans les eaux méditerranéennes, marcher sur les terres de la Massane, courir sur le sable de Saint-Malo ou rouler sur les routes californiennes. Joindre la science à l'agréable... les aventures dans l'équipe furent nombreuses et je réalise la chance que j'ai eue grâce à vous. Merci, merci, merci.

Je tiens également à te remercier **Alain** pour avoir suivi mes travaux de thèse, pour tous tes conseils, tes encouragements et ton soutien. Merci.

Merci à tous les membres des deux équipes auxquelles j'ai appartenu. Merci à tous pour vos conseils et astuces pour venir à bout de ces trois années. Je remercie tout particulièrement **Christel, Eric, Marie-Christine & Nathalie** avec qui j'ai eu le plaisir de travailler et avec une pensée toute particulière pour **Richard** qui coule des jours tranquilles à la retraite et qui a bien raison. Je voudrais également remercier mes collègues montpelliérains qui ont toujours su m'accueillir et m'entourer lors de mes visites iridiennes.

M'enseigner les ficelles de la bio-informatique n'était sûrement pas une mince affaire lorsque je suis arrivée ici en master. Pourtant **Marie-Christine** tu n'as jamais baissé les bras même lorsque je t'assurais que mon ordinateur était vivant et qu'il s'amusait à me faire des blagues. Depuis, les awk ont fusé, les programmes ont *runé*, les jobs ont *jobé*... le plus fou dans l'histoire c'est qu'aujourd'hui j'adore ça ! Je te remercie Marie-Christine pour avoir eu la patience de m'enseigner et de me conseiller. Je sais la chance que j'ai eue de t'avoir à mes côtés depuis le début et si cette thèse s'est réalisée avec beaucoup de sérénité, c'est en grande partie grâce à toi. Nos aventures anglaises et californiennes, nos selfies touristiques, nos discussions interminables et nos entraînements semi-marathoniens (*résolus*) y sont également pour beaucoup.

Lorsque je n'étais pas devant mon ordinateur, j'ai eu à de très nombreuses fois besoin d'aide à la paillasse. Fidèle au poste et toujours avec le sourire et des petits mots encourageants, **Christel** tu aurais déplacé des montagnes pour m'aider et ton aide a été si précieuse depuis trois ans que c'est difficile de trouver les mots pour la qualifier. Tu as été ma maman de labo, ma confidente et souvent ma co-équipière de Southern, de broyage de grains et de manips en tous genre. Je t'en remercie très chaleureusement.

Christophe tes origines géographiques ~~ont été~~ auraient pu être gênantes mais nous avons su aller au-delà des frontières et dépasser nos différences... Merci pour tous nos moments partagés, nos discussions passionnées et passionnantes, tes conseils et tes remarques parfois pertinentes qui me feraient presque oublier que tu es bordelais ! J'ai toujours trouvé auprès de toi du réconfort et des encouragements lorsque c'était nécessaire et je te remercie du fond du cœur d'avoir été un si bon ami durant cette thèse.

Edouard, tu sais que mes petits mots oranges vont me manquer ! Je tiens très sincèrement à te remercier pour nos moments passés ensemble, nos discussions passionnées et enragées et nos tours de stades essoufflés. Merci pour tout ça et pour tous les autres trucs. Merci également à **Julie & Nathalie** pour tous nos moments sportifs et caféinés qui m'ont permis de réaliser cette thèse avec beaucoup de sérénité.

Je tiens également à remercier le meilleur colocataire et coéquipier de thèse du monde **Jérémy**. Nous sommes venus à bout de ces trois années, non pas sans douleur, sans remise en questions et sans développement musculaire, mais nous y sommes arrivés. Je suis très heureuse d'avoir partagé cette expérience avec toi alors merci.

Rémy je tiens particulièrement à te remercier pour avoir un si grand nombre de fois partagé ton expérience et tes conseils en recherche et en randonnées, pour tes encouragements et tes idées arrêtées. J'ai toujours pu compter sur toi (même pour écrire une préface) et je t'en suis particulièrement reconnaissante.

Merci **Fred** pour avoir partagé ton expérience de chercheur expérimenté mais encore jeune, merci pour tous les conseils et encouragements que tu m'as apportés et merci pour toutes nos discussions scientifiques ou pas. J'ai pris énormément de plaisir à discuter, argumenter et enrager avec (ou contre) toi.

Merci à **Emilie & Elodie** pour tous les jolis moments partagés ensemble (failles temporelles, discussions nocturnes, partages de recettes citronnées ou lavage d'enceintes, et j'en passe...) et qui j'espère ne seront pas les derniers. Et une pensée toute particulière pour tous les thésards du laboratoire.

J'ai l'habitude de dire qu'une thèse c'est comme des montagnes russes, il y a des hauts et il y a des bas. Lorsque je fais le bilan de ces trois années et demi, je me rends compte qu'il y a eu beaucoup plus de hauts que de bas et cela grâce à tous les **LGDPiennes & LGDPiens** avec une pensée toute particulière pour **Jean-Jacques** (et **Emmanuel**), **Michèle**, **Sylvie**, **Jean-René**, **Viviane**, **Laura**, **Ariadna**, **Guillaume**, **Thierry**, **Dom**, **Sophie**, **Myriam**, **Elisabeth**, **Odile**,... Merci pour votre aide quotidienne, scientifique ou pas. Merci pour tous ces moments partagés entre deux manip ou deux tours de stade, à la cafet' ou dans la salle PCR, un café (~~ou une bière~~) à la main ou des baskets aux pieds. J'ai eu beaucoup de chance et de plaisir à travailler avec vous et je vous en remercie très sincèrement.

PopRice, merci d'avoir alimenté en mystères ma thèse et d'avoir illuminé mes journées et sombré certaines de mes nuits. Si ça n'a pas été facile de t'accepter dans le monde des publications, on aura vécu de beaux moments scientifiques tous les deux. Tous les mystères qui t'entourent ne sont pas résolus et je suis certaine que tu as encore de beaux jours devant toi. Merci de m'avoir tout de même livré quelques-uns de tes secrets. Si c'était à refaire, je miserais sur toi les yeux fermés. Et merci petit **Cluster** d'avoir suivi le rythme de mes analyses et de m'avoir (parfois) montré le chemin de mes erreurs.

Je ne pouvais pas imaginer mieux que d'arriver au LMI en scooter avec le boss !! **Michel** je tiens tout particulièrement à te remercier de m'avoir si bien accueillie et entourée à Hanoi. Ce fut une expérience très enrichissante et je ne l'oublierai pas. I would also like to thank **Jurek** for his welcome at Cambridge and for this great week.

Je tiens également à remercier tous les collaborateurs avec qui j'ai eu le plaisir de travailler. Merci pour ces étroites et enrichissantes (et parfois exotiques) collaborations. And a special thanks to **Bence, Ernandes & Salvatore**.

Virginie tu m'as sauvée la mise administrativement parlant un paquet de fois ! Alors même si nos échanges n'ont été qu'électroniques, je tenais très sincèrement à te remercier pour ta si précieuse aide.

Laurine, Léo, Pierre, Mélanie, Gaëtan, mon **Binôme** et tous les autres **BFPiens**. Depuis notre rencontre, les aventures n'ont pas cessé. Que ça soit aux quatre coins de l'Europe, sur les pistes de ski, dans une salle de la BU ou assis à une table devant un verre, chaque moment passé ensemble fut inestimable. N'arrêtons jamais les aventures BFPiennes ! Un merci tout particulier à tous nos enseignants de master qui ont rendu ces deux années exceptionnelles.

Julie, Joana, Paul, Nico, Emilie, Mathilde & Laurent et tous les autres que j'oublie, en trois ans je n'ai sûrement pas été l'amie la plus disponible physiquement ou téléphoniquement parlant, je vais tâcher de me rattraper. Je réalise la chance que c'est de vous compter parmi mes amis. Merci.

Papa, Maman, Caroline même lorsque vous ne compreniez pas ce que je voulais faire, vous n'avez jamais cessé de me soutenir et de m'encourager. Si je suis arrivée jusque-là aujourd'hui c'est seulement grâce à vous. Je ne vous en remercierai jamais assez. Et un merci tout particulier à ma mamie et à tous mes si nombreux tontons et taties, cousins et cousines d'être cette famille si formidable.

Trois ans c'est long et court à la fois. Long parce qu'il y a eu des moments de stress, de paniques et de plaques d'eczéma et court parce que l'histoire ne fait que commencer. Merci **Jordan** d'avoir su gérer mon stress et mes inquiétudes avec tant de bravoure (et le mot est faible) de patience et d'amour. Merci de m'avoir montré que l'aventure d'une vie ne se résume pas à une thèse mais plutôt à toutes les aventures qui nous attendent dans notre camion. Les routes et les chemins de montagne n'attendent que nous et j'ai hâte de parcourir tout ce chemin avec toi. Un lourd, sincère et amoureux merci.

La biologie vous gagne et vous rattrape. **Papi**, j'aurais préféré te montrer cette thèse terminée mais c'est ainsi. *Gracias papi y mamie por esta familia si formidable.* **Laurent** tu m'as injustement rappelé que la vie était bien trop courte et qu'il fallait en savourer chaque instant alors promis je ne l'oublierai pas.

PREFACE

***B**arbara McClintock aurait-elle imaginé des décennies après sa découverte des éléments transposables, qu'aujourd'hui nous serions capables de les identifier de manière massive à l'échelle des génomes par des approches de biologie moléculaire ? Aujourd'hui, appelée mobilome-Seq, cette approche permet de faire l'inventaire des éléments transposables actifs au sein des génomes. Cet ouvrage retrace donc la mise en place de cette approche pour notamment l'étude des génomes de l'arabette et du riz. Dans un premier temps, l'auteur y dresse de manière élégante le répertoire des éléments transposables actifs. Grâce à cette approche innovante, l'auteur y découvre un nouvel élément, surnommé PopRice, actif dans l'albumen du riz. Grâce à cette approche, l'auteur a développé d'étroites collaborations qui ont permis l'application de cette méthode à d'autres contextes biologiques. L'auteur y décrit notamment, le lien étroit qui existe entre inhibition de l'ARN polymérase II et la mobilisation de ces éléments actifs. Par ces travaux, l'auteur fait donc un nouveau saut en avant dans l'analyse des éléments transposables. De manière certaine, ces travaux laisseront une trace dans le monde des éléments transposables et ceci par un mécanisme de copier-coller et non de couper-coller.*

L'écrivain transposable

Abréviations

ABA	Acide abscissique
ADN	Acide désoxyribonucléique
ADNec	ADN extrachromosomique
ADNecc	ADN extrachromosomique circulaire
AFLP	<i>Amplified fragment length polymorphism</i>
AP	Protéase aspartique
ARNies	ARN spécifiques des IES
ARNm	ARN messenger
ARNt	ARN de transfert
Bp / Kb	Paire de base / Kilobases
CRISPR	<i>Clustered regularly interspaced short palindromic repeats</i>
DAP	Jours après la pollinisation
DDM1	<i>Decreased in DNA methylation</i>
DME	Demeter
ET	Élément transposable
FISH	<i>Fluorescence in situ hybridation</i>
FON1	<i>Floral organ number 1</i>
FT	Facteur de transcription
GBS	<i>Genotyping by sequencing</i>
GWAS	<i>Genome-wide association study</i>
HS	Stress thermique
HR	Recombinaison homologue
HRE	<i>Heat shock element</i>
IES	<i>Internal eliminated sequences</i>
IGV	<i>Integrative genomics viewer</i>
IN	Intégrase
IRGSP	<i>International rice genome sequencing project</i>
LINE	<i>Long interspread nuclear element</i>
LTR	<i>Long terminal repeat</i>
Mb / Gb	Mégabases / Gigabases
MITE	<i>Miniature inverted-repeat transposable element</i>
NGS	<i>Next generation sequencing</i>
NHEJ	<i>Non-homologous end-joining</i>
Nt	Nucléotides
ORF	<i>Open reading frame</i>
PLE	<i>PopRice-like element</i>
POL	Polyprotéine
Pol II	ARN Polymérase II
polyA	Polyadénylé
Q-Q plot	<i>Quantile-quantile plot</i>
RdDM	<i>RNA dependent DNA methylation</i>
RT	-Transcriptase réverse
siARN	petits ARN interférents
SINE	<i>Short interspread nuclear element</i>
SNP	<i>Single nucleotide polymorphism</i>
spcDNA	<i>Small polydisperse circular DNA</i>
TIR	<i>Terminal inverted repeat</i>
TSS	Site d'initiation à la transcription
VLP	<i>Virus-like particule</i>
WT	<i>Wild-type</i>

SOMMAIRE

INTRODUCTION GÉNÉRALE	1
CHAPITRE 1 : L'ACTIVITÉ DES ÉLÉMENTS TRANSPOSABLES AU SEIN DES GÉNOMES.....	6
1. Caractérisation des ET	7
1.1 Classification des ET.....	7
1.1.1 Classe I : les rétrotransposons.....	7
1.1.2 Classe II : les transposons à ADN.....	8
1.2 Les mécanismes de transposition.....	9
1.2.1 Mécanisme de « copier-coller ».....	9
1.2.2 Mécanisme de « couper-coller ».....	10
1.3 Distribution des ET au sein des génomes.....	11
2. Mécanismes de contrôle des ET	13
2.1 Maintenance et mise en place de la méthylation des ET.....	13
<i>Revue : DNA Methylation in Rice and Relevance for Breeding</i>	
2.2 Réactivation des ET au cours du développement.....	14
2.3 Réactivation des ET en conditions de stress.....	16
3. Rôle des ET au sein des génomes eucaryotes	17
3.1 Les ET : une source significative d'éléments régulateurs	18
3.2 L'impact des ET sur les phénotypes des organismes hôtes.....	20
3.3 Bénéfiques ou néfastes ?	21
4. Méthodes de détection des ET.....	22
4.1 Techniques moléculaires de détection des ET.....	23
4.2 Techniques de détection des ET par séquençage.....	24
4.2.1 Analyses transcriptionnelles par RNA-seq.....	24
4.2.2 Capture d'ADN de rétrotransposon (RC-seq, ATLAS-seq).....	25
4.2.3 Reséquençage de génome.....	25
4.2.3.1 Séquençage de 2 ^{ème} génération et détection des néo-insertions.....	25
4.2.3.2 Séquençage de 3 ^{ème} génération et détection des néo-insertions.....	27
CHAPITRE 2 : ÉTUDE DES ADN CIRCULAIRES EXTRACHROMOSOMIQUES.....	29
1. Définition du mobilome et caractérisation des ADNecc.....	30
1.1 Définition et origine du mobilome.....	30
1.2 Rôles émergents des ADNecc dans les cellules eucaryotes.....	31
1.2.1 L'ADNecc comme biomarqueurs de cancers ?.....	31
1.2.2 Les ADNecc participent à la production des petits ARN chez la paramécie... ..	33
1.3 L'ADNecc : un trésor génomique inexploité ?.....	34

2. Méthodes de détection des ADNec	35
2.1 Méthodes moléculaires pour la détection d'ADNec.....	35
2.2 Méthodes génome entier pour la détection d'ADNec.....	35
CHAPITRE 3 : RÉSULTATS	37
PARTIE 1. LE SÉQUENÇAGE DE L'ADN CIRCULAIRE EXTRACHROMOSOMIQUE RÉVÈLE L'ACTIVITÉ DES	
RÉTROTRANSPOSONS CHEZ LES PLANTES	38
<i>1^{ère} publication : Sequencing the extrachromosomal circular mobilome reveals retrotransposon activity in plants.</i>	
PARTIE 2. CARACTÉRISATION DE L'ACTIVITÉ DE POPRICE AU SEIN DES GRAINS DE CÉRÉALES	40
Introduction	40
Résultats	41
2.1 Régulations génétiques et épigénétiques de <i>PopRice</i>	41
2.2 Activité transpositionnelle de <i>PopRice</i> chez Nipponbare.....	43
2.3 Analyse comparative de <i>PopRice</i> au sein des céréales	45
2.3.1 Histoire évolutive de <i>PopRice</i> au sein des céréales.....	45
2.3.2 Analyse de <i>PopRice</i> chez différentes variétés de riz.....	47
2.3.3 Analyse de <i>PopRice</i> chez différentes espèces de riz.....	48
2.3.4 Étude comparative de l'activité de <i>PopRice</i> chez les céréales.....	50
2.4 Activité de <i>PopRice</i> en conditions de stress chez Nipponbare.....	51
Discussion	52
PARTIE 3. CARACTÉRISATION DU MOBILOME-SEQ CHEZ DIFFÉRENTS ORGANISMES	57
3.1 Les cousins de <i>PopRice</i> (<i>PopRice-like elements</i>) sont-ils actifs chez le maïs ?.....	57
3.2 La déstabilisation de l'épigénome chez les plantes induit-elle une réactivation massive	
du mobilome ?.....	59
3.3 Les très gros génomes ont-ils un mobilome plus actifs ?.....	60
3.4 L'ARN polymérase II est-elle impliquée dans le contrôle des ET.....	61
<i>2^{ème} publication (collaboration) : Inhibition of RNA polymerase II allows controlled mobilisation of retrotransposons for plant breeding.</i>	
3.5 Quel est le rôle des ADNec viraux chez la drosophile ?.....	63
3.6 Conclusion.....	63
DISCUSSION GÉNÉRALE	65
MATÉRIEL & MÉTHODES	70
BIBLIOGRAPHIE	79
ANNEXES	92

- INTRODUCTION GÉNÉRALE -

Les domaines de la biologie des organismes et de la génétique ont été marqués par plusieurs grands noms de la science, de Gregor Mendel à Charles Darwin en passant par Oswald Avery, Rosalind Franklin, Conrad Hal Waddington, Barbara McClintock, ... Il serait difficile de n'en citer qu'un mais il est aussi difficile de les citer tous et le choix peut vite s'avérer cornélien. L'histoire de la science a été et est sans cesse marquée par la remise en cause de croyances longtemps transmises, par l'apparition de nouvelles idées et par l'émergence de nouvelles avancées technologiques. Barbara McClintock disait en 1973 : « *Il faut attendre le moment propice pour les changements de paradigme* ». Quelques années plus tard, en 1983 elle recevait le prix Nobel pour la découverte des éléments transposables (ET) dans les années 50. Trente années, c'est le temps qu'elle aura attendu pour que ses travaux de recherche soient valorisés et enfin reconnus par la communauté scientifique. Mais pourquoi tant d'années ont été nécessaires ? Le contexte de l'époque peut expliquer bien des choses.

Durant les années 40, les scientifiques s'interrogent sur la molécule ou les molécules qui renferment l'information génétique d'un organisme. C'est Oswald Avery et son équipe qui identifièrent pour la première fois en 1944 l'acide désoxyribonucléique, autrement dit l'ADN chez la bactérie (Avery et *al.* 1944; Ghose 2004). Ils ont ainsi montré que l'information génétique d'un organisme était renfermé dans l'ADN et non dans des protéines, comme longtemps supposé. Six années plus tard, à son tour, Edwin Chargaff décrypte la composition de la macromolécule d'ADN et de ses quatre bases azotées (Chargaff 1950) et l'année suivante, à l'aide de rayons X, Rosalind Franklin obtient la première photographie de l'ADN. Ce sont les travaux de ces trois scientifiques oubliés qui permettront à Francis Crick et James Watson d'identifier la structure en double hélice de l'ADN (Watson et Crick 1953) et qui permettront le décodage du code génétique par Har Gobind Khorana, Robert W. Holley, Marshall W. Nirenberg. Malgré le travail considérable de certains scientifiques dans la découverte de l'ADN, c'est F. Crick et J. Watson qui ont obtenu le prix Nobel de médecine en 1963.

Nous pouvons alors nous demander comment les travaux sur l'ADN peuvent expliquer les réticences à l'encontre de l'existence des ET ? En 1950, lorsque Barbara McClintock souleva l'hypothèse de la présence de gènes capables de se déplacer dans l'ADN chez le maïs, elle fut face à de nombreuses et violentes critiques. La molécule d'ADN tout juste identifiée, il était alors impensable à l'époque de croire en l'existence d'éléments capables de créer de l'instabilité au sein de ce matériel génétique. Le « dogme central de la biologie moléculaire » consistait en trois niveaux strictement régulés : l'ADN transcrit en ARN lui-même traduit en protéines (Crick

1970). Ce dogme laissait peu de place à l'existence et l'intervention d'autres éléments génétiques et à une possible dynamique des génomes. Les travaux effectués chez la levure et sur l'élément P chez la drosophile (Britten et Kohne 1968) confirmeront l'existence des ET dans les génomes et suivront quelques années plus tard la reconnaissance des travaux de Barbara McClintock. Néanmoins, le dogme central toujours présent dans la communauté scientifique, a longtemps réfuté l'intérêt des ET dans les génomes hôtes et durant des décennies, les ET ont été considérés comme de simples parasites du génome, de l'« ADN poubelle » sans fonction biologique.

En parallèle, la naissance de l'épigénétique semble une fois de plus en contradiction avec le dogme central. Dès les années 40, Conrad Hal Waddington introduit une nouvelle notion, l'assimilation génétique et parle de « paysage épigénétique » pour expliquer les relations complexes et dynamiques entre l'organisme et les facteurs développementaux et environnementaux (Waddington 1940; 1942). Ces travaux sur l'impact des chocs thermiques sur les œufs de drosophile lui permirent d'observer l'effet de l'environnement sur le phénotype de l'organisme et il proposa l'assimilation génétique comme un nouveau concept de régulation génétique. En réponses à des facteurs environnementaux, le génome peut assimiler une réponse. Les théories de C. H. Waddington ont eu un grand impact dans la communauté scientifique, en grande partie parce que ses travaux allaient dans le sens des hypothèses de Lamarck sur l'hérédité des caractères acquis et parce qu'il a proposé le concept d'épigénétique, sans en connaître les mécanismes moléculaires. Plus tard, l'épigénétique sera définie comme l'étude des changements d'expression de gènes, stables et héréditaires sans modification de la séquence d'ADN, enterrant définitivement le dogme central et ré-ouvrant une porte aux travaux de Lamarck.

Les avancées techniques ont sans cesse remis en question et réorienté les idées scientifiques. Plus récemment, la vision des ET et de l'organisation des génomes a changé et notamment par l'étude des génomes à grande échelle permise par le développement du séquençage de l'ADN à haut débit. Le séquençage de l'ADN a ouvert la voie à une nouvelle discipline : la génomique (et l'épigénomique). L'accès aux séquences d'ADN et la production massive de données nécessitent le développement continu d'outils bioinformatiques afin de trouver dans ce flot de données la ou les réponses aux questions biologiques soulevées.

Aujourd'hui, si leurs mécanismes de régulation et d'activation ne sont pas encore clairement tous caractérisés, quelques exemples illustrent le rôle des ET dans l'évolution des organismes

eucaryotes (Feschotte et Pritham 2007) et des études ont notamment mis en évidence l'implication d'ET dans l'apparition de nouveaux traits agronomiques (Lisch 2013; Song et Cao 2017). Depuis les débuts de la génomique, la détection des mouvements des ET s'est avérée difficile et souvent limitée. Si parfois, l'identification d'ET actifs semble s'apparenter à chercher une aiguille dans une botte de foin, l'étude des ET dans une plante d'intérêt agronomique représente un véritable enjeu dans notre compréhension de l'adaptation des plantes à leur environnement. C'est dans ce contexte général que s'insère mes travaux de thèse.

Mes premiers travaux de thèse ont été effectués sur le riz asiatique, plante modèle des céréales, *Oryza sativa ssp japonica*. Le riz est la première céréale cultivée pour l'alimentation humaine et fait l'objet de nombreux projets de recherche partout dans le monde. Cette plante possède une riche diversité génétique et offre de multiples avantages pour la génomique : un cycle de vie court, un petit génome (430 mégabases - Mb), un séquençage et une annotation de haute qualité (*International Rice Genome Sequencing Project, IRGSP 2005*). De plus, nous disposons également d'une base de données expertisée d'ET (Copetti et al. 2015). Ces nombreux avantages décrit ci-dessus font d'*Oryza sativa* un modèle intéressant pour l'étude des ET chez les plantes.

Mes travaux de thèses se sont alors organisés autour de différents axes :

1 - Développement d'un outil bioinformatique dédié à l'identification des éléments mobiles d'un génome : mon premier objectif fut la mise en place de différentes stratégies bioinformatiques pour analyser, identifier et dresser une liste exhaustive des ET les plus actifs dans un tissu donné.

2 – Étude de l'activité des ET au cours du développement du grain de riz : l'étude des ET suscite de nombreuses questions qui restent sans réponse. Quelle est la part mobile d'un génome ? Quels sont les facteurs (environnementaux, moléculaires) qui activent les ET ? Quels sont les mécanismes d'activation des ET ? Quels impacts ont l'activité des ET au cours du développement d'un organisme ? J'ai consacré une partie de mes travaux de thèse à tenter de répondre à ces questions en caractérisant l'activité des ET dans différents tissus et à différents stades de développement du grain de riz *Oryza sativa*.

3 - Étude de l'activité des ET chez différents organismes : Nathan Springer en conférence scientifique (Phenome 2017, source : Twitter) a souligné que « même s'il était difficile de travailler avec, il ne faut pas ignorer les transposons ». Au cours de mes travaux et au fil des

collaborations, nous avons étendu et développé notre stratégie à un ensemble de génomes eucaryotes complexes (maïs, blé, pin, peuplier, etc...) dans le but de rendre plus facile l'étude des ET.

Le premier chapitre de ce manuscrit est consacré à une synthèse bibliographique des connaissances actuelles sur l'activité des ET dans les génomes eucaryotes et les techniques aujourd'hui utilisées pour les identifier. Le second chapitre définit et caractérise les ADN extrachromosomiques circulaires, témoins de l'activité des ET et de la stabilité d'un génome. Le troisième chapitre est divisé en trois parties et regroupe les résultats que j'ai obtenus lors de mes travaux de thèse. Pour finir, la discussion générale résume mes travaux et les perspectives qui en découlent. Les publications sont incluses dans le manuscrit.

CHAPITRE 1

- L'ACTIVITÉ DES ÉLÉMENTS TRANSPOSABLES AU SEIN DES
GÉNOMES -

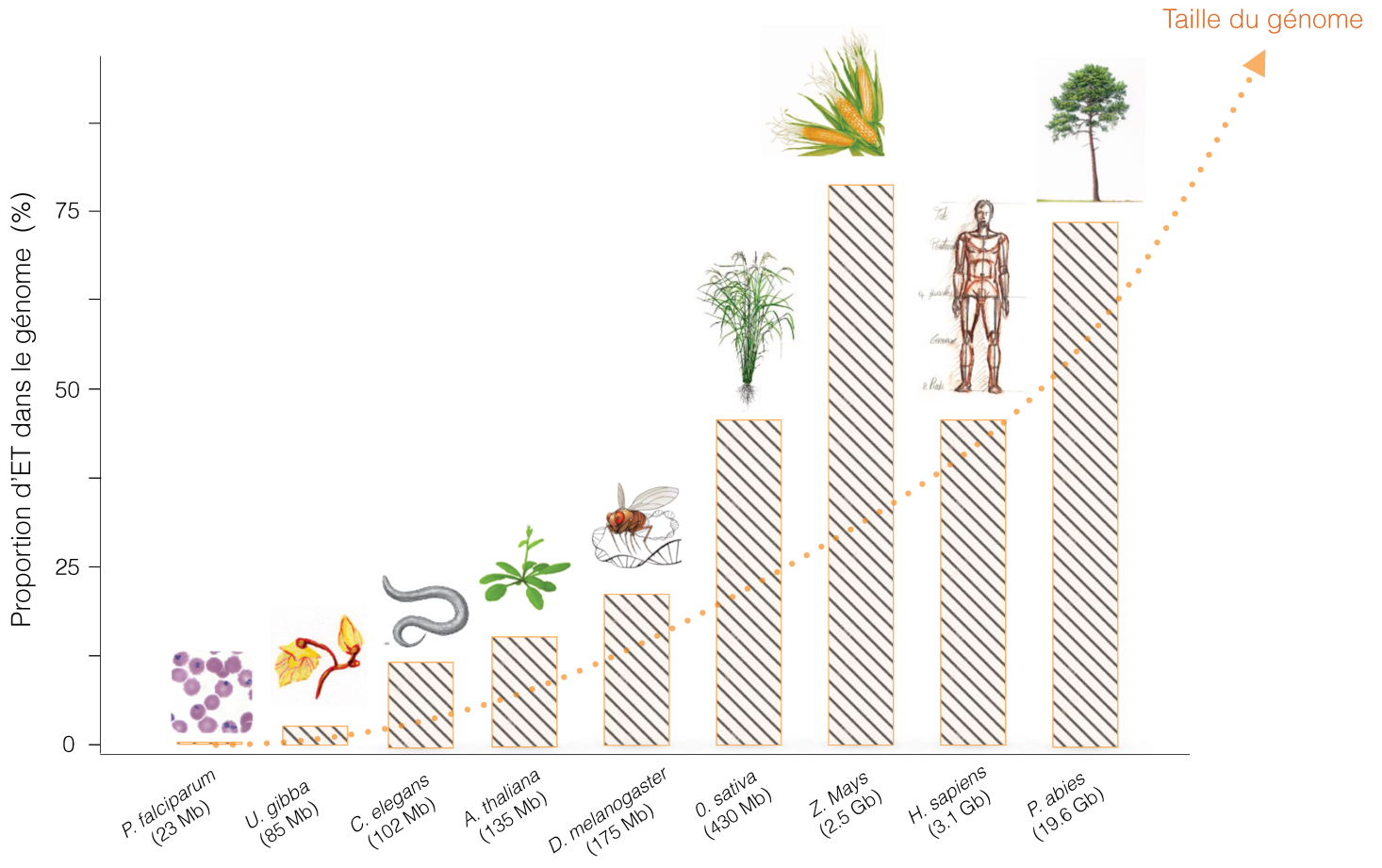


Figure 1. Distribution des ET (en %) en fonction de la taille des génomes des organismes eucaryotes.

1. Caractérisation des éléments transposables

1.1 Classification des éléments transposables

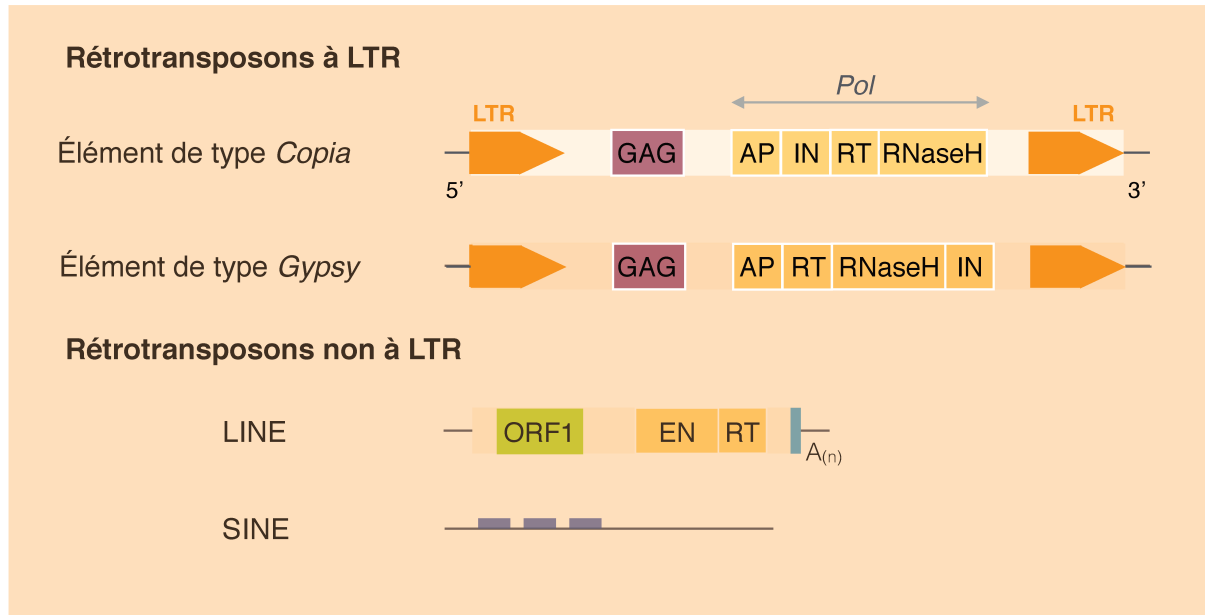
Les ET sont des constituants majeurs des génomes eucaryotes pouvant représenter jusqu'à plus de 80% d'un génome hôte, chez le maïs par exemple (Schnable et *al.* 2009). La taille d'un génome eucaryote est étroitement liée à son contenu en ET. Plus un génome est grand, plus le nombre d'ET est conséquent (Figure 1). À ce jour, tous les génomes eucaryotes séquencés possèdent des ET à deux exceptions près celui du parasite humain *Plasmodium falciparum* (23 Mb, aucun ET détecté ; Gardner et *al.*, 2002) et celui de l'algue verte *Micromonas pusilla* (12-13 Mb, aucun ET détecté ; Worden et *al.* 2009). L'abondance et la riche diversité des ET a conduit à une classification rigoureuse de ces éléments.

À partir de la proposition de Finnegan en 1989, les ET sont divisés en deux grandes classes définies par leur mécanisme de transposition : les rétrotransposons (classe I) et les transposons à ADN (classe II). Depuis 1989, les connaissances sur les mécanismes, sur la diversité et sur la complexité des ET n'ont cessé de s'accroître et la classification des ET est sujette à de nombreux débats dans la communauté. De nouvelles améliorations et différentes classifications ont été apportées et proposées (Piégu et *al.* 2015) et ne cesseront d'évoluer dans le futur. Finalement, ici est présentée une classification simplifiée des principales classes d'ET basée sur la classification proposée par Wicker et *al.* (2007) dans le but d'évaluer l'étendue de la diversité de ces éléments dans les génomes hôtes eucaryotes. Les rétrotransposons transposent *via* un mécanisme dit de « copier-coller » à l'aide d'un ARN intermédiaire alors que les transposons à ADN transposent *via* un mécanisme dit de « couper-coller » sans ARN intermédiaire (Wicker et *al.* 2007) (Figure 2). Chaque classe d'ET dispose d'éléments autonomes et non autonomes. Les éléments autonomes codent les enzymes nécessaires à leur transposition contrairement à la mobilité des éléments non-autonomes qui dépend des enzymes produites par les éléments autonomes de la même famille ou de familles proches.

1.1.1 Classe I : les rétrotransposons

Selon la classification de Wicker et *al.* (2007), l'organisation des rétrotransposons est basée sur des caractéristiques mécanistiques et sur l'organisation et la phylogénie de la transcriptase réverse (RT). Les rétrotransposons sont divisés en deux ordres : les rétrotransposons à LTR (*Long Terminal Repeat*) et les rétrotransposons non LTR.

Classe I : rétrotransposons (éléments à ARN)



Classe II : transposons à ADN

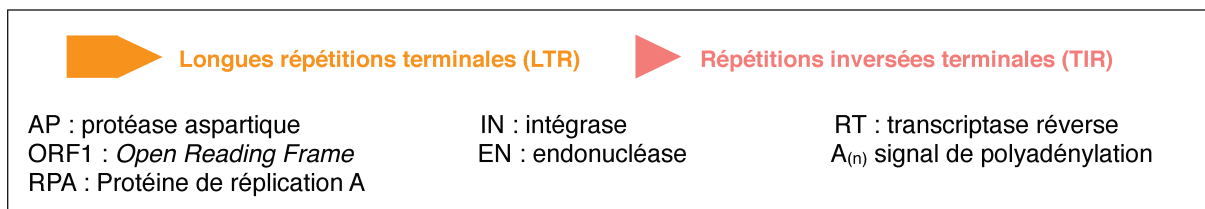
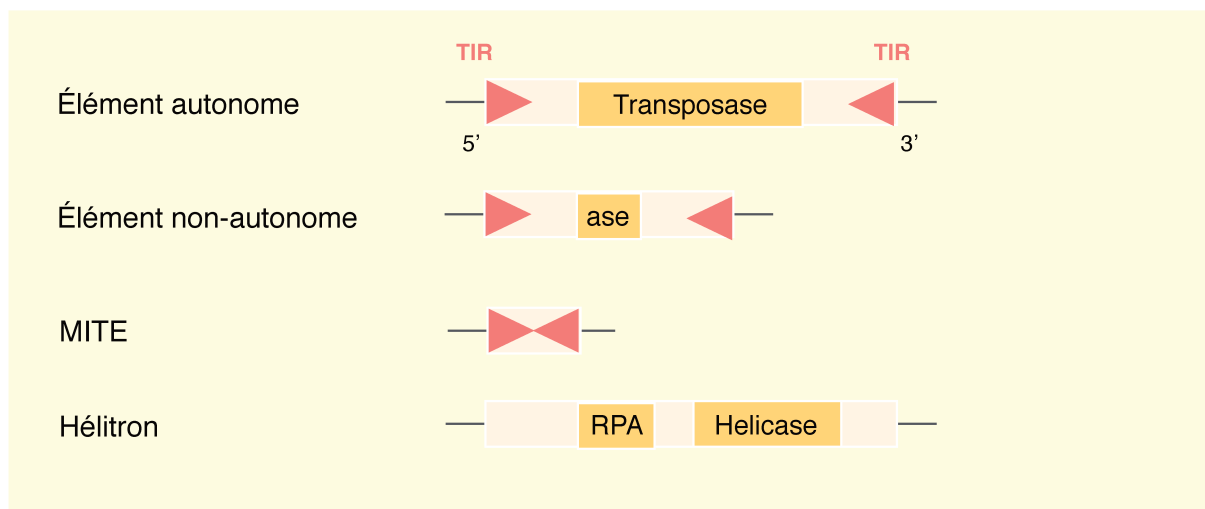


Figure 2. Classification simplifiée des éléments transposables (Wicker et al., 2007)

Les rétrotransposons à LTR comprennent deux séquences répétées à leur extrémité, commençant en général par TG en 5' et se terminant par CA en 3'. Le LTR contient la séquence promotrice de l'ARN polymérase II (Pol II) et marque le début de la transcription mais il indique également la fin de la transcription et le signal de la polyadénylation. Les rétrotransposons contiennent généralement une ou deux ORF (*Open Reading Frame*) qui codent la GAG, une protéine qui participe à la formation d'une particule de type viral (VLP) et la POL, une polyprotéine clivée en 4 protéines : une RT, une RNase H, une protéase aspartique (AP) et une intégrase (IN) (Figure 2). Les rétrotransposons à LTR sont divisés en deux superfamilles, *Copia* et *Gypsy* qui diffèrent par l'ordre de la RT et de l'IN dans la région codante *Pol*. Les rétrotransposons à LTR sont prédominants dans les génomes des plantes contrairement aux génomes animaux où les rétrotransposons non LTR sont majoritaires.

Les rétrotransposons non LTR sont divisés en deux classes : les LINE et les SINE (*Long and Short Interspread Nuclear Elements*). Les LINE codent pour une RT et une endonucléase et leurs transcrits sont polyadénylés en 3'. Les LINE représentent jusqu'à 20% du génome humain (Lander et al. 2001) alors que chez les plantes les LINE semblent rares. Les SINE sont des éléments non autonomes qui dépendent des LINE pour transposer. La région interne des SINE est très variable et dépend de la famille de l'élément. L'élément le plus étudié appartenant à la classe des SINE est l'élément *Alu* qui représente à lui seul 11% du génome humain (Lander et al. 2001).

1.1.2 Classe II : les transposons à ADN

Les transposons à ADN (classe II) sont caractérisés par la présence de séquences répétées inversées (TIR) de part et d'autre de leurs extrémités. Les éléments autonomes qui codent pour une transposase sont divisés en différentes superfamilles telles que *Ac/Ds*, *CACTA*, *MULE*. Certains éléments de ces superfamilles peuvent évoluer en éléments non autonomes soit en accumulant des mutations dans leur région codante soit par la délétion complète de la région codante conduisant à la formation de MITE (*Miniature Inverted-repeat Transposable Element*) (Figure 2). À titre d'exemple, dans le génome du riz, les MITE sont les éléments avec le plus grand nombre de copies (environ 90000 copies dans certaines variétés (Jiang et al. 2004)). Il existe aussi un autre type de transposon à ADN, les héliçons qui transposent à l'aide d'un mécanisme de cercle roulant (Yang et Bennetzen 2009).

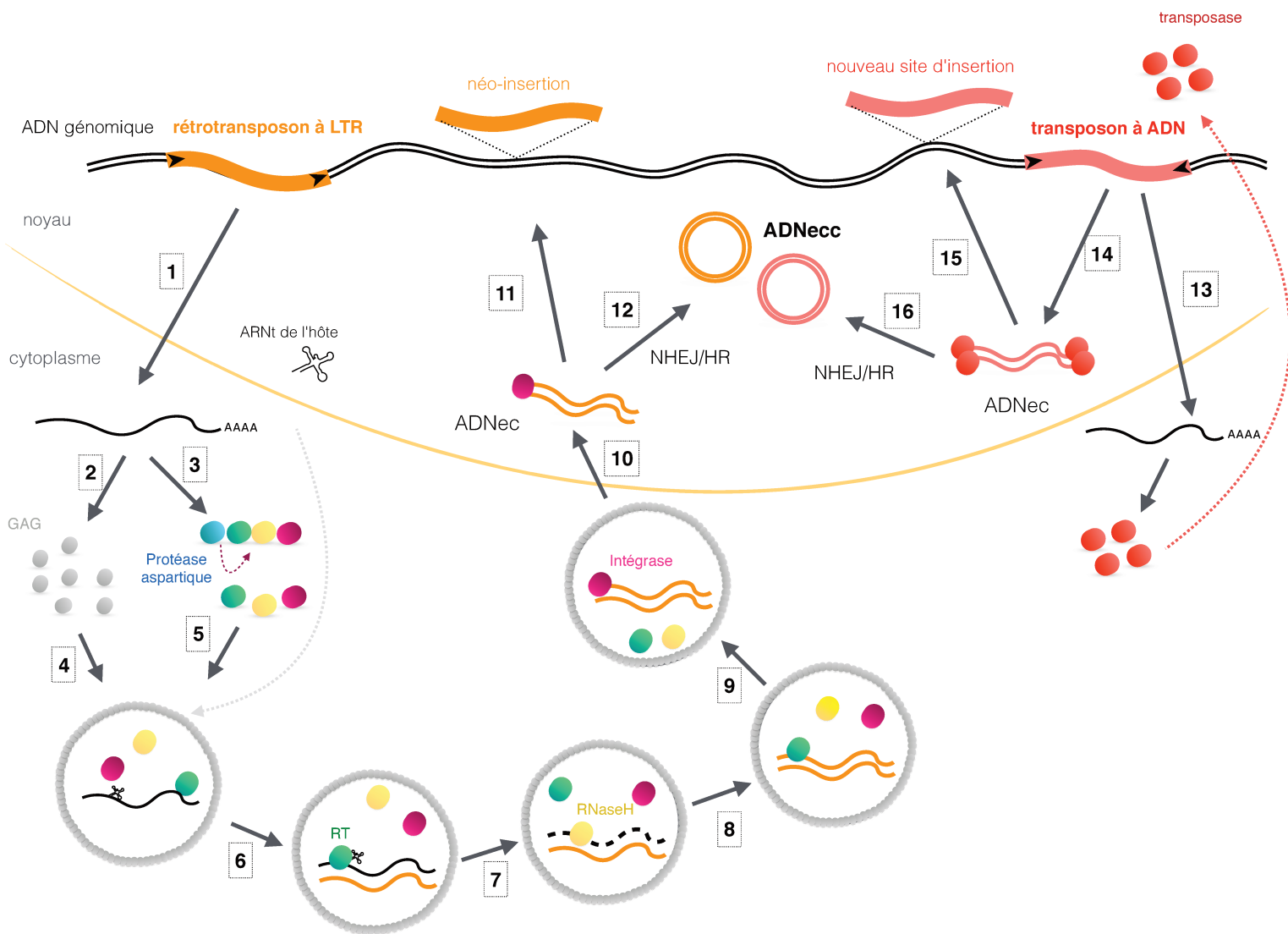


Figure 3. Cycle de vie des éléments transposables. Le cycle de vie d'un RT à LTR (classe 1) est composé de différentes étapes (Schulman 2013). Le LTR en 5' contient la séquence promotrice de l'ARN polymérase II et marque le début de la transcription (1) et à l'inverse, le LTR en 3' indique la fin de la transcription et le signal de la polyadénylation. Les transcrits du RT sont utilisés à la fois comme matrice pour la traduction (2,3) et pour la réverse transcription en double brin d'ADN (6-8). Dans le cytoplasme, la polyprotéine est autoclivée en 4 protéines (3) : une réverse transcriptase (RT, cercles verts), une RNase H (cercles jaunes), une protéase aspartique (AP, cercles bleus) et une intégrase (IN, cercles roses). L'interaction entre plusieurs protéines GAG (cercles gris) permet la protection des transcrits et des 4 protéines à l'intérieur d'une particule de type viral (*Virus-like particle*; VLP) (4,5). La liaison entre l'ARNt du génome hôte et le PBS (*Primer Binding Site*) inséré à la fin 3' du LTR 5' initie la réverse transcription des transcrits en ADN *via* la RT (6). Durant la réverse transcription, la RNaseH dégrade l'ARN matrice (7) et le brin complémentaire est réverse-transcrit (8). La copie d'ADN extrachromosomique (ADNec) synthétisée est associée à l'IN (9) et migre dans le noyau via des mécanismes non identifiés (10) afin de ré-intégrer le génome hôte et ainsi créer une nouvelle copie de l'élément (11). L'ADNec peut également être reconnu par les mécanismes de réparation de l'ADN et est alors capturé par le système *non-homologous end-joining* (NHEJ) induisant la formation d'ADN extrachromosomique circulaires (ADNecc) (12). L'ADNecc peut également être formé par recombinaison homologue (HR). Les transposons à ADN (classe 2) autonomes codent une protéine : la transposase (13). Ces protéines se lient à chacune des extrémités de l'élément et excisent l'élément de son locus (14) et l'insèrent à un nouveau locus (15). Avant sa ré-insertion, l'ADNec peut être reconnu par les mécanismes de réparation de l'ADN et être capturé par le système NHEJ induisant la formation d'ADNecc (16). L'ADNecc peut également être formé par HR.

1.2 Les mécanismes de transposition

1.2.1 Mécanisme de « copier-coller »

Le cycle de vie d'un rétrotransposon à LTR est composé de différentes étapes (Figure 3). La première étape de rétrotransposition est initiée par la transcription de l'élément et les transcrits migrent dans le cytoplasme où ils sont utilisés comme matrice pour la traduction et également comme matrice pour la reverse transcription en double brin d'ADN (Schulman 2013). Dans le cytoplasme, la polyprotéine est autoclivée et l'interaction entre plusieurs protéines GAG forme la VLP. À l'intérieur de la VLP, l'ARN messager (ARNm) est reverse-transcrit en ADN extrachromosomique (ADNec). Cet ADNec nouvellement synthétisé est associé à l'IN et migre dans le noyau *via* des mécanismes encore méconnus afin de réintégrer le génome hôte et de créer une nouvelle copie de l'élément. Ainsi et contrairement aux transposons à ADN, à chaque cycle de rétrotransposition, une nouvelle copie du rétrotransposon est produite et explique pourquoi les rétrotransposons sont souvent les éléments majoritaires dans les génomes. Néanmoins, l'ADNec peut également être reconnu par les mécanismes de réparation de l'ADN avant sa réinsertion dans le génome et être capturé par la voie du *Non-Homologous End-Joining* (NHEJ) induisant la formation d'ADN extrachromosomique circulaire (ADNec). La formation de ces ADNec a été démontrée à partir de travaux effectués sur les rétrovirus (Li et *al.* 2001; Kilzer et *al.* 2003). Des ADNec issus d'ET peuvent également être formés par recombinaison homologue (HR) (voir Chapitre 2 page 29).

Sans pression de sélection, les ET accumulent des mutations au cours du temps et peuvent perdre leur capacité à transposer. Il est en revanche possible de déterminer l'âge de l'insertion d'un rétrotransposon à LTR en calculant l'identité entre ses deux LTR. En effet, les LTR jouent un rôle majeur dans la reverse-transcription de l'ARNm en ADNec (Figure 4) (Schulman 2013). Les LTR d'un rétrotransposon sont constitués de trois domaines : U5 (unique en 5' de l'ARNm), R (répété aux deux extrémités de l'ARNm), U3 (unique en 3' de l'ARNm). La transcription d'un rétrotransposon à LTR démarre dans le domaine R du LTR en 5' et se termine dans le domaine R du LTR 3', ainsi à chaque extrémité du transcrit, une partie de la séquence du LTR n'est pas transcrite. La reverse-transcription est initiée par la liaison d'un ARN de transfert (ARNt) du génome hôte au PBS (*Primer Binding Site*) présent en aval du LTR 5' de l'élément. La synthèse de l'ADN est réalisée de 5' vers 3'. Ce fragment d'ADN synthétisé qui comporte la séquence U5 et R du LTR, est transféré du côté 3' de la matrice d'ARN et s'hybride au niveau de la région R du LTR. Le brin est alors synthétisé du LTR 3' complet jusqu'au PBS,

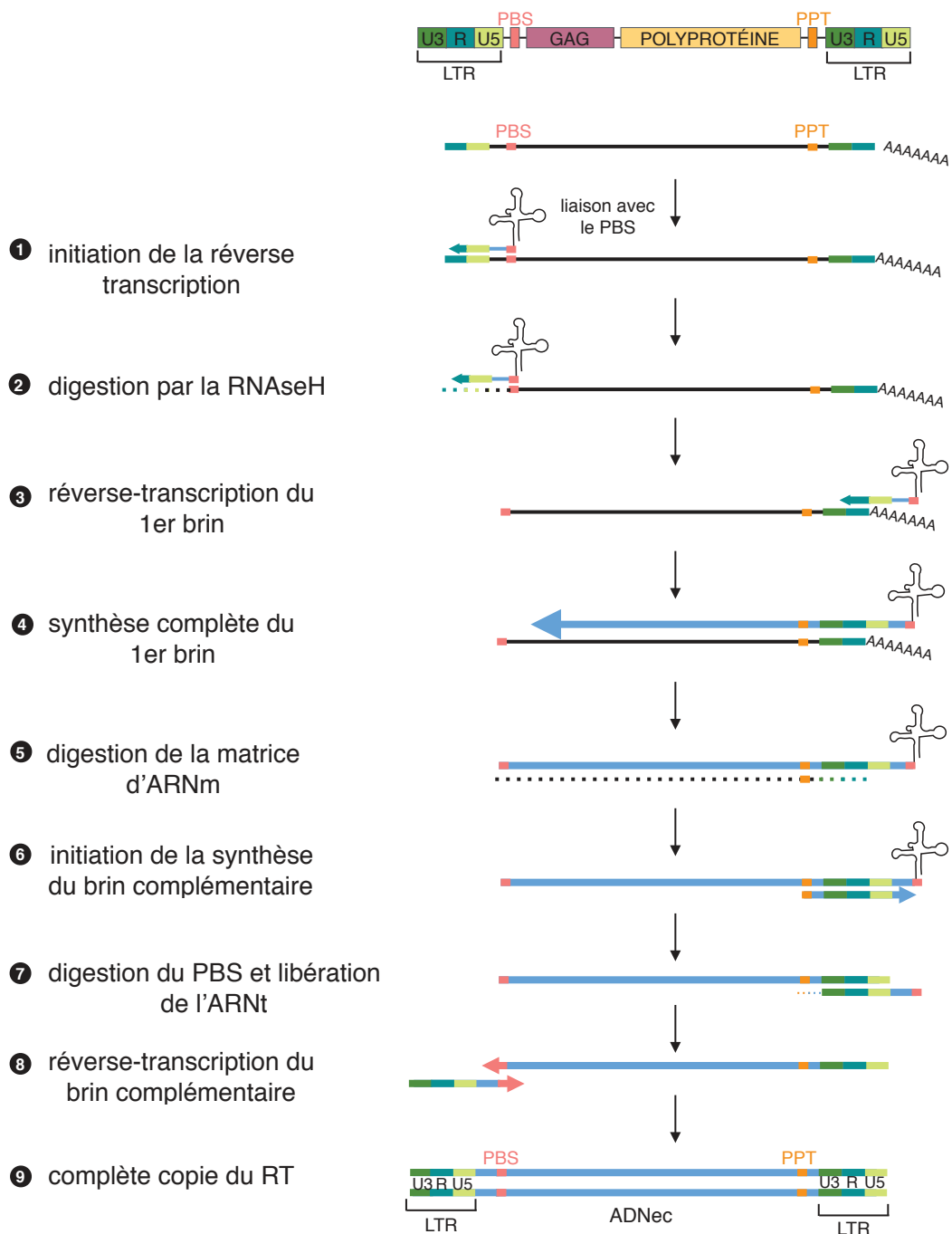


Figure 4. Étapes de la réverse-transcription de l'ARNm en ADN linéaire double brin (ADNec). L'ARNt du génome hôte se lie au PBS qui sert d'amorce pour initier la réverse-transcription (1). La protéine RNaseH dégrade l'ARN hybridé à l'ADN nouvellement synthétisé (2). L'ADN synthétisé est transféré du côté 3' de la matrice d'ARN et s'hybride au niveau de la région R du LTR (3). Le brin entier est synthétisé par la RT (4). La protéine RNaseH dégrade l'ARN excepté le PPT qui résiste à la dégradation (5). La synthèse de l'ADN du brin complémentaire démarre au PPT, utilisé alors comme amorce d'initiation (6). La RNaseH dégrade le PPT et l'ARNt du génome hôte est libéré de l'ADN (7). Les deux régions du PBS s'hybrident (8) et la synthèse des deux brins s'étend afin de synthétiser entièrement la copie du RT (9). Les flèches indiquent le sens de la réverse-transcription, les pointillés représentent l'ARN dégradé. Le LTR est divisé en trois régions : U5 (Unique en 5' de l'ARNm), R (Répété aux deux extrémités de l'ARNm), U3 (Unique en 3' de l'ARNm) ; PBS : *Primer Binding Site* ; PPT : *Poly-Purine Tract*. Figure adaptée des travaux de Sarafianos et al. (2001).

le LTR 5' n'est donc pas synthétisé en entier à ce stade du processus. La synthèse du LTR 5' du brin complémentaire démarre au PPT (*Poly-Purine Tract*) présent du côté 3' de l'élément. La synthèse est réalisée de 5' vers 3'. Le fragment d'ADN complémentaire nouvellement synthétisé comporte la séquence complète du LTR et s'hybride au brin codant au niveau du PBS. La dernière étape de la reverse-transcription consiste à compléter la synthèse des deux brins afin d'obtenir la séquence complète de l'élément. Le processus de reverse-transcription a été caractérisé notamment grâce aux travaux réalisés sur le virus du HIV chez l'homme (Sarafianos et al. 2001; Menéndez-Arias et al. 2017). En définitive, du fait que le LTR d'une extrémité joue le rôle de matrice pour synthétiser le deuxième LTR, au moment de l'insertion d'un rétrotransposon dans le génome, les deux LTR sont identiques. Au fil du temps, des mutations s'accumulent et les séquences des deux LTR divergent et à l'aide du taux de substitution moyen, l'âge de l'insertion peut être déterminé à partir de la divergence moyenne entre les deux LTR (SanMiguel et al. 1998). Il est donc possible de déterminer si une famille de rétrotransposons à LTR semble récemment active ou non.

1.2.2 Mécanisme de « couper-coller »

Les transposons à ADN ont un cycle de transposition qui semble relativement court comparé aux éléments de classe I et se déroule uniquement dans le noyau. Les enzymes transposases reconnaissent et se lient aux TIR aux deux extrémités de l'élément. L'élément est excisé de son locus et est inséré à un nouveau locus (Figure 3) (Muñoz-López et García-Pérez 2010). Cependant, avant sa réinsertion, l'ADN peut être reconnu par les mécanismes de réparation de l'ADN (NHEJ ou HR) induisant la formation d'ADNec (Li et al. 2001; Sundaresan et Freeling 2007). Paradoxalement à leur mécanisme de transposition, il s'avère que certaines familles de transposons à ADN ont largement proliféré dans les génomes hôtes (exemple de *mPing* chez le riz (Jiang et al. 2003) et de *Heartbreaker* chez le maïs (Zhang et al. 2000)). Deux hypothèses ont été avancées pour expliquer la création de nouvelles copies par les transposons à ADN (Feschotte et Pritham 2007). La première est que lorsque le transposon s'excise de son locus, si l'élément est présent sur la chromatide sœur, le site excisé peut être réparé par recombinaison et l'élément est ainsi réintroduit au locus excisé. La deuxième hypothèse implique que durant la phase de réplication de l'ADN, le transposon s'excise de son locus nouvellement répliqué et se réinsère à un locus non encore répliqué. La deuxième hypothèse a été étayée par une étude sur les transposons IS de bactérie dont la transposition serait couplée à la réplication de l'hôte (Ton-Hoang et al. 2010).

1.3 Distribution des éléments transposables au sein des génomes hôtes

La distribution des ET diffère selon les familles d'ET et selon les génomes. Le site d'insertion d'un ET dans le génome influence fortement le devenir de cet élément. En effet, les ET qui s'insèrent préférentiellement dans des régions géniques auront tendance à se multiplier plus rapidement car c'est un environnement favorable à la transcription. En revanche, les possibilités d'être détectés et silencés par l'hôte seront plus élevées. Par exemple, les transposons à ADN *Mutator* chez le maïs et *Pack-MULE* chez le riz et les MITE *mPing* chez le riz, ciblent spécifiquement les régions riches en gènes (Liu et al. 2009; Naito et al. 2009; Jiang et al. 2011). Au contraire, certaines familles s'insèrent dans des régions « *safe havens* » pauvres en gènes, telles que les régions centromériques ou péri-centromériques devenant alors des « cimetières » à ET où les éléments restent inactifs et perdent rapidement leur capacité à transposer (Chuong et al. 2016a).

Les sites d'intégration des ET sont directement liés aux propriétés des enzymes qui interagissent avec l'ADN pour permettre l'insertion d'une nouvelle copie dans le génome hôte. L'interaction de l'enzyme avec le génome hôte diffère selon l'enzyme et induit indirectement une variabilité et des spécificités des sites ciblés par l'élément. À titre d'exemple, l'élément LINE L1 très actif chez les mammifères (Richardson et al. 2014), est spécifiquement inséré dans des régions riches en T. La queue polyadénylée (polyA) de l'ARNm de L1 interagit avec des sites riches en T et la reverse-transcription a lieu au site d'intégration ce qui explique la spécificité d'insertion de L1 (Sultana et al. 2017). En revanche, les transposases des transposons à ADN et les IN des rétrotransposons à LTR ciblent de courtes séquences d'ADN réparties sur l'ensemble du génome hôte suggérant que les transposons à ADN et les rétrotransposons à LTR ont peu de sites spécifiques d'insertions. Il existe cependant des exceptions comme par exemple le rétrotransposon ZAM chez *Drosophila melanogaster* qui reconnaît et s'insère spécifiquement dans des sites CGCGCG (Faye et al. 2008; Sultana et al. 2017).

Néanmoins, au-delà des spécificités enzymatiques de chacun des mécanismes de transposition, des facteurs chromatiniques et nucléaires jouent également un rôle dans l'intégration des ET. Par exemple chez *Arabidopsis*, Tsukahara et al. (2012) ont étudié la spécificité d'insertion de certains rétrotransposons à LTR actifs dans deux espèces d'*Arabidopsis*. Les auteurs ont remarqué que les néo-insertions du rétrotransposon *Évadé* (*ATCOPIA93*) étaient localisées dans

des régions géniques chez *Arabidopsis thaliana* contrairement à l'élément *Évadé-like* présent dans le génome d'*A. lyrata* qui s'insère préférentiellement dans les régions répétées centromériques. Afin de comprendre les spécificités d'insertions des ET, les auteurs ont introduit l'élément *Évadé-like* dans le génome d'*A. thaliana* et ont montré que les néo-insertions d'*Évadé-like* étaient spécifiquement localisées dans une séquence satellite répétée en tandem dans les centromères d'*A. thaliana* (Tsukahara et al. 2012). Or, les séquences satellites centromériques évoluent rapidement et même si les deux espèces d'*Arabidopsis* sont des espèces proches, les séquences satellites sont divergentes (environ 30%) au sein des deux génomes (Kawabe et Nasuda 2005). *Évadé* et *Évadé-like* ont des sites d'insertions spécifiques différents et les auteurs suggèrent qu'*Évadé-like* pourrait reconnaître des marques caractéristiques des centromères et qui expliquerait sa spécificité d'insertion (Tsukahara et al. 2012). De manière plus générale, les régions hétérochromatiques sont très riches en ET et de nombreuses études ont montré des biais d'insertions de certaines familles d'ET (Bushman 2003; Pereira 2004; Baucom et al. 2009). L'abondance d'ET dans l'hétérochromatine *versus* l'euchromatine semble être induite par des interactions spécifiques entre des ET et des marques épigénétiques associées avec l'hétérochromatine (Gao et al. 2008). Certains chercheurs ont également suggéré que le mécanisme de réparation de l'ADN étant moins efficace dans l'hétérochromatine, les cassures d'ADN pourraient alors favoriser l'insertion d'ET (Dimitri 1997). La présence de ces ET dans l'hétérochromatine pourrait être un mécanisme sélectionné par les génomes hôtes pour maintenir les marques épigénétiques répressives afin de conserver la compaction de l'ADN caractéristique de l'hétérochromatine (Lippman et al. 2004; Slotkin et Martienssen 2007).

Enfin, la distribution des ET à l'échelle d'un génome peut également traduire le niveau d'activité des ET au sein du génome. Chez *Arabidopsis thaliana*, malgré des ET spécifiques des régions géniques (cf. *Évadé*), les ET sont très majoritairement présents dans les régions centromériques et péri-centromériques contrairement aux autres plantes qui ont un génome plus grand et une proportion d'ET plus importante telles que le riz *Oryza sativa* où les ET sont distribués tout le long des chromosomes (Mirouze et Vitte 2014). Les copies d'ET accumulées dans les régions centromériques et péri-centromériques s'éliminent plus lentement que les copies insérées dans les régions géniques. La différence de temps d'élimination entre les ET est expliquée par (1) l'absence de pression de sélection, (2) par l'accumulation de mutations spontanées et (3) par un faible taux de recombinaison dans les régions centromériques et

péricentromériques (Wright et *al.* 2003). La prédominance de copies d'ET dans ces régions semble témoigner de la faible activité des ET chez *Arabidopsis* contrairement au riz.

Outre le devenir propre de l'ET, l'environnement dans lequel l'élément s'insère peut fortement moduler le devenir de la cellule ou de l'organisme hôte. Afin de contrôler leur prolifération et de réprimer leur pouvoir mutagène les ET sont strictement régulés par l'hôte.

2. Mécanismes de contrôle des éléments transposables

Les organismes hôtes ont développé de rigoureux mécanismes épigénétiques tels que la méthylation de l'ADN assurant une répression stable de la transcription des ET et un contrôle de leur prolifération. Au cours des divisions cellulaires, la méthylation de l'ADN est maintenue par des méthyltransférases de maintenance alors que les jeunes copies d'ET sont ciblées par de la méthylation *de novo* initiée par un mécanisme spécifique des plantes la voie appelée RdDM (*RNA-dependent DNA Methylation*). De plus en plus d'études épigénétiques sont réalisées chez le riz et les principaux régulateurs de la méthylation de l'ADN sont aujourd'hui identifiés. Nous avons regroupé les connaissances actuelles sur les régulations épigénétiques et plus précisément sur les acteurs de la méthylation de l'ADN dans une revue publiée dans le journal *Epigenomes*. Je renvoie le lecteur à cette revue pour la partie descriptive des mécanismes de méthylation notamment chez le riz. Dans cette revue, nous avons également discuté de l'importance de la méthylation de l'ADN dans le contrôle des caractères agronomiques et des futures perspectives à considérer dans ce domaine. Je terminerai cette partie en détaillant plus précisément certaines conditions particulières dans lesquelles les mécanismes épigénétiques peuvent être relâchés et conduire à la réactivation des ET.

Review

DNA Methylation in Rice and Relevance for Breeding

Sophie Lanciano * and Marie Mirouze * 

IRD, DIADE, University of Montpellier, Laboratory of Plant Genome and Development, University of Perpignan, 66860 Perpignan, France

* Correspondence: sophie.lanciano@ird.fr (S.L.); marie.mirouze@ird.fr (M.M.);
Tel.: +33-468-662-119 (S.L. & M.M.)

Academic Editor: Etienne Bucher

Received: 3 June 2017; Accepted: 30 June 2017; Published: 4 July 2017

Abstract: The challenge of sustaining food security in the context of global changes is at the heart of plant research. Environmental stresses, in particular, are known to impact genome stability and epigenetic mechanisms. Epigenetic pathways are well characterized in plants, particularly in the dicotyledon model plant *Arabidopsis thaliana*, but an increasing number of epigenetic and epigenomic studies are also performed on rice (*Oryza sativa*). Rice represents a major food crop of worldwide importance and is also a good model for monocotyledons owing to its relatively small genome size and fully sequenced well-annotated genome. Today, the main regulators of DNA methylation are identified in rice. Moreover, compared to *Arabidopsis*, rice has an important evolutionary history due to human selection since its domestication. DNA methylation may be involved in both adaptation and agronomic performances and thus, a better understanding of epigenetic regulations in rice should contribute to improving the adaptation of crops to a changing environment. In this review, we expose the current knowledge on DNA methylation in rice and future perspectives to be considered.

Keywords: epigenetics; epigenomics; *Oryza sativa*; *Arabidopsis thaliana*; DNA methylation; transposable elements; siRNAs

1. Introduction

Epigenetics is defined as the study of chromatin marks including DNA methylation and histones post-translational modifications. Chromatin accessibility modulates DNA replication and repair, expression of genes and transposable elements (TEs) activity in both plants and animals. In plants, cytosine methylation occurs at three contexts (CG, CHG, and CHH; where H is A, T, or C) whereas in mammals mainly CG dinucleotides are methylated [1]. In addition, genome-wide DNA methylation profiles are different between plants and mammals. Indeed, mammalian genomes are strongly methylated: for example in human embryonic stem cells 72–85% of CGs are methylated [2] compared to *Arabidopsis thaliana* where the methylation levels are 24% for CGs, 6.7% for CHGs, and 1.7% for CHHs, respectively [3]. DNA methylation is predominantly found at TEs and repeats, ensuring the maintenance of TEs silencing. Furthermore, DNA methylation is involved in important developmental processes and stress responses in both plants and animals [4,5]. Epigenetic regulatory mechanisms affect the reproductive development, flowering regulation, and stress responses and thus could potentially play a role in crop improvement [6].

Asian rice (*Oryza sativa*) is one of the most important food crops worldwide and is the best model for cereal genomics. Interest for rice research also resides in the rich source of genetic diversity of the species. Whether epigenetic mechanisms and epigenomic variations have accumulated during the long history of selection and domestication in rice and could contribute to adaptation and agronomic traits is a major question in rice research. Epigenetic regulations have been dissected in great detail in *A. thaliana*, but are still poorly characterized in rice, although recent studies have shed some light on epigenetic

regulation in this crop and we refer the reader to excellent recent reviews [7,8]. Here, we first discuss the relevance of epigenetic regulations for breeding through recent examples of epigenetic control of agricultural traits in rice, and we then focus our review on recent work deciphering the epigenetic regulators involved in the maintenance and establishment of DNA methylation in this species.

2. Epigenetic Regulations Are Involved in Agricultural/Adaptive Traits

In both *Arabidopsis* and rice, the crucial role of DNA methylation is well established in various developmental processes such as seed/embryo development, gametophyte development, and flowering time control [9–14]. Furthermore, epimutations (DNA methylation variations) can be heritable or reversible and thus, might allow phenotypic variation and quick response to environmental changes [15,16]. In *A. thaliana*, a massive analysis of 1107 methylomes from the 1001 Genomes collection demonstrates that the intraspecies epigenomic diversity extent can be correlated with both climate and geographical origin [17]. These natural and spontaneous DNA methylation variations induce alterations of gene transcription and can induce the emergence of new adaptive traits.

Indeed, in plants, many analyses have demonstrated the role of DNA methylation in stress responses [18] and in rice the number of epigenetic studies in stress conditions is increasing. For example, Secco et al. [19] have shown the impact of inorganic phosphate (Pi) deficiency on gene transcription and DNA methylation patterns in rice and in *Arabidopsis*. They identified more differentially methylated regions (DMRs) in response to Pi starvation in rice compared to *Arabidopsis*. These DNA methylation changes occur mainly at TEs located near to Pi starvation-induced (*PSI*) genes. They observed that these specific TEs are hypermethylated in response to Pi starvation, suggesting a mechanism repressing their activity in order to limit their deleterious effects and to facilitate *PSI* gene transcription. These differences could be explained by significant differences of TE density (15% of the genome in *Arabidopsis* vs. 40% in *O. sativa*).

Rice productivity is also affected by two major stresses, salinity and water deficit. Garg et al. [20] identified DMRs associated with differential expression of genes involved in abiotic stress responses in three cultivars—one stress sensitive and two drought and salinity tolerant, respectively—in normal conditions. These DMRs could explain the resistance phenotype. In addition, Wang et al. [21] demonstrated that, under drought conditions, a drought-resistant genotype has a more stable methylome than a drought-sensitive genotype, suggesting again the influence of DNA methylation in abiotic stress response. Interestingly, drought-induced epimutations seem to be non-random and are inherited from generation to generation [22]. Finally, new questions emerge on agricultural practices and on the impact of exposition to pesticides or heavy metal in the soil. Recent studies [23,24] demonstrated that DNA methylation patterns are affected in rice exposed to these substances.

Epigenomic diversity is strongly influenced by TE content and activity [25]. TEs contribute to plant evolution [26] and could influence agricultural traits [27,28]. A picture is emerging where TE polymorphisms and their associated epigenetic marks could contribute to the evolution of gene networks [29] and play a key role in adaptation [30]. Beneficial roles of TEs in rice are reviewed by Song and Cao [31]. One example recently discovered in rice is the natural epiallele *Epi-rav6* [22]. The hypomethylation of a MITE (Miniature Inverted-Repeat Transposable Element) inserted in the *RAV6* promoter induces an over-expression of the gene, resulting in an increase in leaf angle. Undoubtedly, TEs represent a new source of adaptive traits for crop breeding.

Overall, these observations highlight the importance of epigenetic mechanisms in stress responses and raise the question of the molecular actors involved. Furthermore, the seminal work on *Arabidopsis* epigenetic recombinant inbred lines [32,33] has shown that basic knowledge on epigenetic mutation is instrumental if one wants to introduce epigenetic diversity using crosses between mutant and wild type plants. Growing literature on rice epigenetic mutants starts to build a picture of epigenetic regulations in this species. Here, we focus on mutants affecting DNA methylation, as the best-studied epigenetic mark so far.

3. Main Regulators of DNA Methylation in Rice

The methylome (genome-wide DNA methylation) is monitored by DNA methyltransferases (DNA METs), and for some part maintained during replication and thus transmitted across cell divisions. A more dynamic part of the methylome is controlled by 24-nucleotide, small interfering RNAs (siRNAs) via an RNA-directed DNA methylation (RdDM) pathway involving DNA MET activity. In plants, at least three classes (or three major classes) of DNA MET genes have been identified: DNA methyltransferases (*METs*), plant specific chromomethyltransferases (*CMTs*), and domain rearranged methyltransferases (*DRMs*). Genes directly or indirectly involved in DNA methylation are listed in Table 1 and studies on the corresponding rice mutants are detailed. Of note, most of these mutants have been produced by callus culture (*Tos17* insertion, T-DNA, or RNA interference (RNAi) techniques) also known to affect DNA methylation [34] and TE activity. Results should therefore be interpreted with caution.

In *A. thaliana*, MET1, the ortholog of the mammalian Dnmt1 [35], is the major CG methylase and ensures the maintenance of CG methylation [11]. In rice, two closely related putative *MET1* genes are present: *OsMET1-1* and *OsMET1-2* [36]. The loss-of-function mutant *Osmet1-2* presents strong developmental defects in seed development and vegetative growth [37]. The global CG methylation level is reduced by 76% in the homozygous *Osmet1-2* mutant compared to the wild type (WT). CHG and CHH methylation are also affected (6.6 and 43%, respectively). However, the genome-wide CG methylation level in the *Arabidopsis met1* mutant is decreased by 98%, suggesting a redundant function between *OsMET1-1* and *OsMET1-2* in rice. Nevertheless, *Osmet1-1* mutants do not show discernible developmental phenotype [38] and thus, *OsMET1-1* seems to have a minimal and/or redundant function in the maintenance of CG methylation. Consistently, *OsMET1-2* is expressed at higher levels than *OsMET1-1* [37]. Gene expression is largely altered in *Osmet1-2* (13% misregulated genes) while it is only slightly affected in *A. thaliana met1* mutant (2%), suggesting that a large proportion of genes are regulated directly or indirectly by DNA methylation in rice. Lastly, both transcriptional activity of TEs and 24 nucleotide (nt) siRNA production are disturbed in *Osmet1-2* mutant, indicating that *OsMET1-2* could be involved in transcriptional silencing of TEs [37].

CMTs are plant-specific DNA METs, characterized by the presence of a chromo (chromatin organization modifier) domain and a bromo-adjacent homology (BAH) domain in the N-terminal region. In *Arabidopsis*, three *CMT* genes have been identified: *CMT2* is known to establish CHH methylation [39] while *CMT3* is a major CHG MET [1,40,41]. *CMT1* is only weakly expressed and its function is still unknown [42]. In rice, there are three *CMT* genes: *OsCMT2*, *OsCMT3a*, and *OsCMT3b*. *OsCMT3a* is the only functional *CMT3* ortholog and is involved in the maintenance of DNA hypermethylation at CHG sites during DNA replication [43]. The loss-of-function *Oscmt3a* mutation affects the expression of genes and TEs [43]. TEs are predominantly transcriptionally activated and transgenerational TE mobility is observed in mutants confirming the role of *OsCMT3a* in their control. In contrast to *Arabidopsis cmt3*, *Oscmt3a* mutant displays pleiotropic developmental phenotypes: early flowering, short stature, and low fertility. *Oscmt3b* mutant does not present any morphological abnormality and *OsCMT3b* is expressed only in panicles, suggesting that *OsCMT3b* could play a minor role in CHG methylation. Finally *OsCMT2* is closely related to *CMT2* [43], suggesting that *OsCMT2* may play a role in CHH methylation, although no *Oscmt2* mutant is described yet.

Table 1. Genes involved in DNA methylation in rice.

	Proteins	Locus ID	Mutation	Expression	Description/Phenotype	Functions	References
Maintenance of DNA methylation	OsMET1-2 (DNA METHYLTRANSFERASE 1)	LOC_Os07g08500	T-DNA insertion (<i>Tos17</i>)	KO	All germinated seedlings undergo quick necrotic death	Maintain DNA methylation at CG sites during DNA replication. Two copies <i>MET1-1</i> and <i>MET1-2</i>	[37,44]
	OsDRM1a	LOC_Os11g01810	/	/	Downregulated by jasmonic acid	Not expressed, lack of methyltransferase motifs	[45]
	OsDRM1b	LOC_Os12g01800	/	/	Downregulated by jasmonic acid		
	OsCMT3 (CHROMOMETHYLTRANSFERASE)	LOC_Os10g01570	T-DNA insertion (<i>Tos17</i>)	KO	No difference in the vegetative phase. Early reproductive stage, 15% shorter stature and decreased fertility	Maintain DNA methylation at CHG sites during DNA replication	[43,44]
Chromatin remodeler	OsDDM1a (DECREASE in DNA METHYLATION)	LOC_Os09g27060	RNAi mutants	KD	93% identity between both DDM1 homologs; dwarf phenotype; hypomethylation in later generations of selfed progenies	Remodeling histones ATPases. Maintenance of cytosine methylation; Required for maintenance of TE silencing	[46,47]
	OsDDM1b (DECREASED in DNA METHYLATION)	LOC_Os03g51230	RNAi mutants	KD	/	Maintenance of cytosine methylation	[46,47]
RdDM	OsDRM2 (DOMAINS REARRANGED METHYLTRANSFERASE)	LOC_Os03g02010	Gene targeting through homologous recombination	KO	Reduction of vegetative growth and semi-dwarf phenotype. Reduction in the de novo methylation at transposons and 5S repeat sequences	De novo DNA methylation at CHH sites directed by siRNAs. Major DRM1/2-type methyltransferase gene in rice	[44,45,48,49]
	OsDCL3a (DICER LIKE PROTEIN 3)	LOC_Os01g68120	RNAi mutant	KD	Pleiotropic phenotypes affecting agricultural traits: plant height, angle of flag leaf, smaller panicles. Similar phenotypes as RNAi mutants of AGO4ab-1 and RDR2-2	Biogenesis of 24-nt long miRNAs (lmiRNAs) which can direct DNA methylation (<i>cis</i> and <i>trans</i>); 24 nt siRNA biogenesis	[50–52]
	OsDCL3b (DICER LIKE PROTEIN 3)	LOC_Os10g34430	RNAi mutant	KD	/	Panicle and early seed-specific and require for 24 nt phased small RNAs. <i>DCL3a</i> is expressed at a much higher level than <i>DCL3b</i>	[50,53]
	OsDCL4 (DICER LIKE PROTEIN 4)	LOC_Os04g43050	/	KO	Severe spikelet defects including thread-like lemma and male sterility	Biogenesis of 21 nt siRNA in panicles and seedlings	[53,54]
	OsRDR1 (RNA DEPENDENT RNA POLYMERASE 1)	LOC_Os02g50330	T-DNA insertion (<i>Tos17</i>)	KO	Ephemeral phenotypic fluctuations occurred only under some abiotic stress conditions	Role in the production and amplification of exogenous, virus-derived siRNAs (vsiRNAs) in infected plants and in some abiotic stress responses. Role in maintaining the intrinsic locus-specific CHH methylation patterns	[55]

Table 1. Cont.

	Proteins	Locus ID	Mutation	Expression	Description/Phenotype	Functions	References
RdDM	OsRDR2 (RNA DEPENDENT RNA POLYMERASE 2)	LOC_Os04g39160	RNAi mutant	KD	Similar phenotypes as RNAi mutants of AGO4ab-1 and OsDCL3a.	Role not studied yet but could be similar to ATRDR2 (according to its expression pattern)	[50]
	OsRDR6 (RNA DEPENDENT RNA POLYMERASE 6)	LOC_Os01g34350	SNP (G -> T)	Temperature dependent	Spikelet defects	Biogenesis of 21 nt and 24 nt siRNAs (different from <i>Arabidopsis</i>) and resistance against virus	[53,56]
	AGO4a/AGO4b	LOC_Os01g16870/LOC_Os04g06770	RNAi mutants	KD	Similar phenotypes as RNAi mutants of OsDCL3a and RDR2-2	High similarity with <i>Arabidopsis</i> AGO4	[50]
	OsAGO1s (4 OsAGO1 homologs OsAGO1a, OsAGO1b, OsAGO1c, OsAGO1d)	LOC_Os02g45070/LOC_Os04g47870/LOC_Os02g58490/LOC_Os06g51310	RNAi mutants	KD	Various developmental defects	miRNA mediated gene regulation	[50]
	WAF1 (WAVY LEAF1)	LOC_Os07g06970	NMU mutagenesis	KO	Seedling lethality due to defects of SAM maintenance or pleiotropic phenotypes in leaf morphology and floral development. Phenotypes similar to <i>sho1</i> and <i>sho2</i> mutants deficient in DCL4 and AGO7, respectively	Methylates 3' terminal nucleotide of siRNAs; HEN1 (HUA ENHANCER 1) homolog	[57]
5-meC DNA glycosylase/lyases	OsROS1a	LOC_Os01g11900	knock-in targeting	KO	Severe underdeveloped endosperm phenotype	There are 4 ROS1 orthologs (ROS1a-d). ROS1a is the most expressed gene compared to ROS1b-d. ROS1a and <i>Arabidopsis</i> DME gene could have analogous functions in the endosperm	[58]
	DNG701 (OsROS1c)	LOC_Os05g37350	T-DNA insertion, RNAi	KO; KD; OE	The progeny of <i>ros1c</i> mutant present two seed phenotypes, normal seeds and wrinkled seeds	ROS1a and ROS1c could play different roles in seed development. Could be involved in the control of transposition.	[58,59]
	DML3a (DEMETER LIKE 3) and DML3b	LOC_Os04g28860/LOC_Os02g29380	/	/	/	/	[10,58]

DDM1: decrease in DNA methylation; KD: knock-down; KO: knock-out; NMU: N-nitroso-N-methylurethane; OE: over-expressed; SAM: shoot apical meristem; t-DNA: transfer DNA.

DRM is required for the maintenance of non-CG methylation and for de novo methylation in all the three contexts CG, CHG, and CHH [60]. In *Arabidopsis*, de novo methylation is established by the RdDM pathways where siRNAs guide DRM2 to the region to be methylated [61]. DRM2, which is homologous to mammalian Dnmt3, is also functionally redundant with CMT3 in the maintenance of non-CG methylation at many loci [62]. DRM2 is expressed at much higher levels than DRM1 and suggests that DRM2 is the predominant de novo DNA MET in *Arabidopsis* [63]. In rice, homozygous *Osdrm2* mutants show severe developmental defects such as a semi-dwarfed phenotype, reductions in tiller number, abnormal panicle architecture, and complete sterility [45]. The genome-wide DNA methylation level is reduced to 17% in *Osdrm2* compared to the WT and the RdDM pathway is deficient. In contrast to *Arabidopsis*, OsDRM2 is the major CHH MET in rice [46]. Heterologous expression of *OsDRM2* in yeast confirmed that OsDRM2 could methylate DNA de novo [48]. In rice, two other DRM genes were identified respectively, *DRM1a* and *DRM1b*. These two genes are not expressed and might not encode functional DNA METs [45].

In addition to DNA METs, DNA methylation is regulated by chromatin factors. DDM1 (DECREASE IN DNA METHYLATION 1), a SWItch/Sucrose Non-Fermentable (SWI2/SNF2)-like chromatin remodeling protein, is necessary to maintain DNA methylation in *Arabidopsis* [64,65]. Two genes are orthologous of *DDM1* in rice designated as *OsDDM1a* and *OsDDM1b*. These two genes share 93% similarity and their expression patterns are also similar. Interestingly, only the double mutants *Osddm1a* and *Osddm1b* show severe developmental defaults and a complete sterility, suggesting functional redundancy [46,47]. The global DNA methylation level is reduced by 54% in the double mutant *Osddm1a Osddm1b* compared to the WT. Genome-wide analyses of DNA methylation have shown that OsDDM1 is involved in both CG and CHG methylation at euchromatic and heterochromatic regions but is also involved in CHH methylation of small TEs such as MITEs mainly located in euchromatic regions [46]. Transcriptomic analyses in *Arabidopsis ddm1* mutants have shown an important deregulation of gene transcription [39,66] suggesting that OsDDM1 could also play a role in gene regulation and ensure normal plant development.

In plants, DNA demethylation is ensured by 5-mC DNA glycosylase enzymes that are encoded by four genes in *Arabidopsis*: REPRESSOR OF SILENCING (*ROS1*) [67], DEMETER (*DME*) [68], DEMETER-LIKE2 (*DML2*) AND DEMETER-LIKE3 (*DML3*) [69]. In rice, phylogenetic analyses identified six DNA glycosylases: 4 *ROS1* homologs (*ROS1a-d*) and two *DML3* (*DML3a* and *DML3b*) homologs and no *DME* homolog [10]. Expression analysis showed that *ROS1a* is the most expressed gene compared to *ROS1b-d* and *DML3a-b* in the five tissues tested (seedling leaf, seedling root, anther, pistil, and immature seed) [58]. Maternal null *ros1a* mutants present severe underdeveloped endosperm phenotype, reminiscent of the *dme* mutant phenotype in *Arabidopsis*, suggesting that *ROS1a* and *DME* could have analogous functions in the endosperm. Even if the DNA glycosylase function of *ROS1a* is not demonstrated yet, maternal and paternal null *ros1a* alleles are not transmitted to the next generation, suggesting that DNA methylation level play important roles in gametophyte development. Characterization of the *ROS1c/DNG701* gene [59] showed that *ROS1c* is required for the demethylation of the *Tos17* retrotransposon in rice calli, therefore suggesting it could be involved in the control of transposition. In addition, the progeny of the knockout mutant *ros1c* present a proportion of 10% of wrinkled seeds that could be due to the impact of the mutation on the endosperm hypomethylation. The cause of this phenotype and its low penetrance is yet unknown. It is less pronounced in the *ros1a* mutant, suggesting that *ROS1a* and *ROS1c* could play different roles in seed development.

Altogether these studies have shown that rice mutants affected in the DNA methylation machinery present severe developmental phenotypes in both vegetative and reproductive stages in contrast to *Arabidopsis*. In addition, a clear difference of DNA methylation patterns at the chromosome-wide level is observed between both species [70]. Here again, these differences could be explained by significant differences of TE density. Plant species with a high TE content (*O. sativa*, *Zea mays*, etc.) seem to require more robust DNA methylation mechanisms than species with a low TE content such as *Arabidopsis*.

In maize for instance, genetic perturbation of the methylome lead to developmental phenotypes and some mutant combination cannot be recovered due to lethality [71].

4. Establishment of De Novo DNA Methylation

TEs are mobile genetic elements able to proliferate in their host genomes. They represent a main source of genomic diversity and an evolutionary force in both plants and animals. Host genomes have established strong regulations and young TE copies are silenced by epigenetic marks such as cytosine methylation, ensuring a stable repression of TE expression and preventing their proliferation. As mentioned above, de novo DNA methylation can be initiated via the RdDM mechanism, a plant-specific pathway through which siRNAs target homologous DNA regions to methylate it. The RdDM pathway is well characterized in *Arabidopsis* [72,73] but largely unstudied in rice. In this part, we will detail RdDM pathways based on the model plant *A. thaliana* (Figure 4a) to highlight a comprehensive overview of what have been identified in rice (Figure 4b).

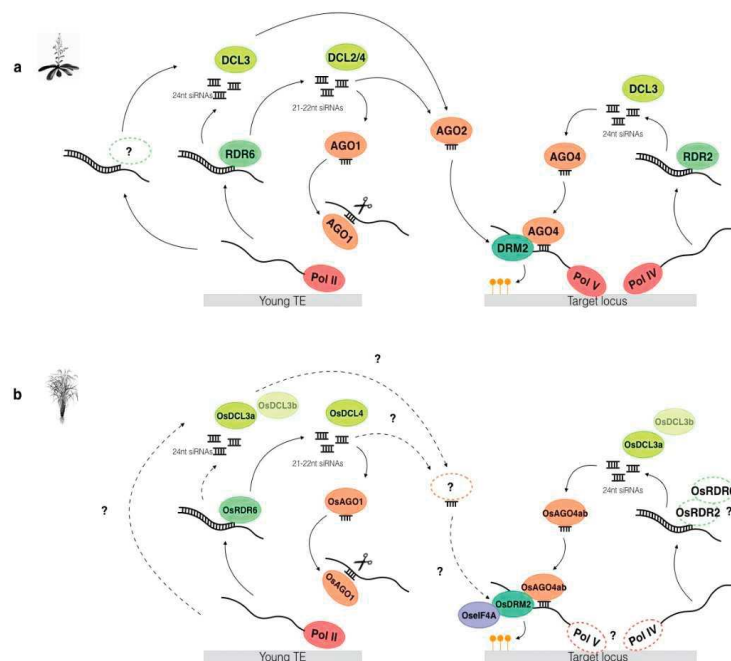


Figure 1. RNA-directed DNA methylation (RdDM) pathways in *Arabidopsis* and in rice. (a) In *Arabidopsis*, the “canonical” RdDM pathway (right panel) is initiated by the RNA polymerase IV (Pol IV) that generates a single strand RNA (ssRNA) of the target locus which is a template for the RNA-dependent RNA polymerase 2 (RDR2) to generate a double strand RNA (dsRNA). DICER-like 3 (DCL3) cleaves dsRNAs to 24 nucleotide (nt) siRNAs then loaded into ARGONAUTE 4 (AGO4). AGO4-bound siRNAs can base-pair with the nascent RNA polymerase V (Pol V) transcript or with DNA leading to the recruitment of DRM2 to establish de novo methylation of the target locus and to mediate transcriptional gene silencing (TGS) of transposable elements (TEs). The “non-canonical” pathway (left panel) incorporate components that are typically associated with post-transcriptional gene silencing (PTGS). A young TE copy, future RdDM target, is transcribed by RNA polymerase II (Pol II) to produce mRNAs. Some of these Pol II transcripts can be converted into dsRNAs by RDR6 and processed by DCL2 and DCL4 into 21–22 nt siRNAs, resulting in AGO1-mediated PTGS by the cleavage of TE mRNAs. These dsRNAs can also initiate de novo DNA methylation by three independent pathways: RDR6–DCL3–RdDM, RDR6–RdDM, and DCL3–RdDM involving AGO2, Pol V, and DRM2. These non-canonical pathways enhance methylation and reinforce TGS. A transition from

PTGS to TGS occurs when high levels of Pol II and RDR6-dependent dsRNAs saturate DCL2 and DCL4 enzymes and become available for processing by DCL3, which produces 24 nt siRNAs that trigger canonical RdDM and TGS of TEs. (b) In rice, only some genes involved in RdDM pathways are functionally characterized. DCLs, AGOs, and RDRs are encoded by one of the largest multigene families and some genes, like *OsAGO4a* and *OsAGO4b* or *OsRDR2* and *OsRDR6* seem to have functional redundancy. However, *OsDCL3a* and *OsDCL3b* seem to have some similar roles, *OsDCL3b* being mostly involved in panicle and early seed development. Dotted lines represent hypothetical pathways or hypothetical proteins, not yet characterized.

Main classes of regulators involved in this complex mechanism include: RNAi machinery ensured by Dicer-like (DCL), Argonaute (AGO), and RNA-dependent RNA polymerases (RDRs) genes families, de novo MET DRMs and two plant-specific RNA polymerases Pol IV and Pol V. The “canonical” RdDM pathway is initiated by Pol IV which generates a single strand RNA (ssRNA) of the target locus which is a template for RDR2 to generate a double strand RNA (dsRNA). DCL3 cleaves dsRNAs to 24 nucleotides siRNAs (24 nt siRNAs) which are stabilized by methylation at their 3'-OH groups by HUA ENHANCER 1 (HEN1) and loaded into AGO4. AGO4-bound siRNAs can base-pair with the nascent Pol V transcript or with DNA [74], leading to the recruitment of DRM2 to establish de novo methylation of the target locus and to mediate transcriptional gene silencing (TGS) of TEs.

In addition to the canonical pathway, various “non-canonical” forms of RdDM have recently been identified in *Arabidopsis* [73]. These mechanisms partly incorporate components that are typically associated with post-transcriptional gene silencing (PTGS). Initially, future RdDM targets—for instance, a young TE copy—are transcribed by Pol II to produce mRNAs. Some of these Pol II transcripts can be converted to dsRNAs by RDR6 and processed by DCL2 and DCL4 into 21–22 nt siRNAs, resulting in AGO1-mediated PTGS by the cleavage of TE mRNAs. However, these dsRNAs can also initiate de novo DNA methylation by three independent pathways: RDR6–DCL3–RdDM, RDR6–RdDM, and DCL3–RdDM involving AGO2, Pol V, and DRM2. These non-canonical pathways enhance methylation and reinforce TGS. A transition from PTGS to TGS occurs when high levels of Pol II- and RDR6-dependent dsRNAs saturate DCL2 and DCL4 enzymes and become available for processing by DCL3, which produces 24 nt siRNAs that trigger canonical RdDM and TGS of TEs. Finally, RdDM mechanisms seem to be able to compensate for each other [75].

So far, only some of the rice RdDM machinery components have been functionally characterized by using RNAi mutants and expression analyses [50]. In plants, DCLs, AGOs, and RDRs are encoded by small multigenic families and curiously in rice these families gather 32 genes, one of the largest numbers of these genes among the plant species analyzed so far. These results suggest that there is a complex and robust network of siRNAs in rice [76]. Nevertheless, only a few genes have been fully identified as partners in the RdDM mechanism. For example, *OsDCL3a* is involved in 24 nt siRNA processing [51]. RdDM mutants display several developmental alterations as lower plant height and smaller panicles compared to WT. RNAi mutants of *OsAGO4a* and *OsAGO4b*, homologs of *Arabidopsis* AGO4 [52], and RNAi mutant of *OsRDR2* have similar phenotypes suggesting that *OsRDR2* and *OsAGO4ab* are involved in the same pathway. The function of *OsRDR2* is not clearly established yet its expression pattern is similar to *RDR2* in *Arabidopsis* at earlier stages of flower development, suggesting a similar role [50]. *OsRDR6* is involved in the biogenesis of 21 and 24 nt siRNAs [53,56] suggesting a possible redundancy between *OsRDR2* and *OsRDR6*. In addition, the loss-of-function of *OsRDR1* causes alteration of CHH methylation indicating a possible role in the RdDM pathway [55]. Finally, *WAVY LEAF1* (*WAF1*) has been identified as an ortholog of *Arabidopsis* *HEN1*. Rice *waf1* mutants show strong pleiotropic phenotypes and do not survive 10 days after germination. Abe et al. [57] have shown that the siRNAs abundance was decreased in this mutant compared to WT. Finally, as in *Arabidopsis*, *WAF1* is required for the stabilization of siRNAs.

RNA polymerases are composed of at least 12 subunits, forming a large holoenzyme [77]. Nuclear RNA Polymerases D and E (NRPD and NRPE) are specific subunits of Pol IV and Pol V, respectively, and are derived from the duplication of Pol II subunit, NRPB [78]. Orthologs of *NRPD* and *NRPE* have

been identified in rice [79] and present the same domain structure as in *Arabidopsis*. No study has been published yet on Pol IV and Pol V in rice nevertheless similarities of structure suggest a similar role.

Interestingly, no rice knockout mutant has been described for any of the early actors of the RdDM pathway, so far [76]. The difficulty of obtaining such mutants suggests that they could be sterile and weak alleles or RNAi lines could be of interest. This might also underline the more drastic effect of affecting the RdDM pathway in this species than in *A. thaliana*. Recently, Bousios and Gaut [80] discussed the importance of studying epigenetic mechanisms in a wide range of species. Indeed, due to the singularities of the *A. thaliana* genome (small genome with weak TE activity), the generalization of the current epigenetic model to plants at large may not be fully relevant. There is therefore a need for a comprehensive study of these mechanisms in another model species, such as rice.

5. Conclusions and Perspectives

Food security and climate changes are serious global concerns and it is now well established that epigenetic regulations are strongly influenced by the environment and could provide a reversible yet heritable source of variation for rapid adaptation. Rice is one of the most important food crops worldwide and its evolution is explained by a long history of selection and domestication. Natural epimutations have accumulated in this species and we propose that they could contribute to adaptation and to agronomic traits in this species. Over recent years, the number of rice epigenetic studies has significantly increased. In response to different stresses, rice epialleles have been identified and are correlated with the appearance of new adaptive traits. Nevertheless, the underlying mechanisms are still unclear and lots of questions remain unanswered. Massive comparative studies of rice varieties will provide an opportunity to improve our knowledge of epigenetic regulations and also to identify new agronomically interesting epialleles. Finally, the fact that epigenetic mechanisms are closely linked to TE content and are thus species-specific, stresses the importance of expanding these epigenetic studies to a large number of crop species.

Acknowledgments: We thank Olivier Panaud (University of Perpignan, France) and two anonymous reviewers for helpful comments and our colleagues at the Institute of Research for Development (IRD) and Laboratory of Plant Genome and Development (LGDP) for stimulating discussions. S.L. is supported by an ANR fellowship (French National Agency for Research). This work was supported by the IRD, the FAiD (Fédération d'aide pour le développement, <http://faid.univ-perp.fr/>), and the French National Agency for Research (ANR-13-JSV6-0002).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Du, J.; Johnson, L.M.; Jacobsen, S.E.; Patel, D.J. DNA methylation pathways and their crosstalk with histone methylation. *Nat. Rev. Mol. Cell Biol.* **2015**, *16*, 519–532. [[CrossRef](#)] [[PubMed](#)]
2. Chen, P.-Y.; Feng, S.; Joo, J.W.J.; Jacobsen, S.E.; Pellegrini, M. A comparative analysis of DNA methylation across human embryonic stem cell lines. *Genome Biol.* **2011**, *12*, R62. [[CrossRef](#)] [[PubMed](#)]
3. Cokus, S.J.; Feng, S.; Zhang, X.; Chen, Z.; Merriman, B.; Haudenschild, C.D.; Pradhan, S.; Nelson, S.F.; Pellegrini, M.; Jacobsen, S.E. Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* **2008**, *452*, 215–219. [[CrossRef](#)] [[PubMed](#)]
4. Law, J.A.; Jacobsen, S.E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat. Rev. Genet.* **2010**, *11*, 204–220. [[CrossRef](#)] [[PubMed](#)]
5. Allis, C.D.; Jenuwein, T. The molecular hallmarks of epigenetic control. *Nat. Rev. Genet.* **2016**, *17*, 487–500. [[CrossRef](#)] [[PubMed](#)]
6. Springer, N.M. Epigenetics and crop improvement. *Trends Genet.* **2013**, *29*, 241–247. [[CrossRef](#)] [[PubMed](#)]
7. Chen, X.; Zhou, D.-X. Rice epigenomics and epigenetics: Challenges and opportunities. *Curr. Opin. Plant Biol.* **2013**, *16*, 164–169. [[CrossRef](#)] [[PubMed](#)]
8. Deng, X.; Song, X.; Wei, L.; Liu, C. Epigenetic regulation and epigenomic landscape in rice. *Nat. Sci. Rev.* **2016**, *3*, 309–327. [[CrossRef](#)]
9. Gehring, M.; Bubb, K.L.; Henikoff, S. Extensive demethylation of repetitive elements during seed development underlies gene imprinting. *Science* **2009**, *324*, 1447–1451. [[CrossRef](#)] [[PubMed](#)]

10. Zemach, A.; Kim, M.Y.; Silva, P.; Rodrigues, J.A.; Dotson, B.; Brooks, M.D.; Zilberman, D. Local DNA hypomethylation activates genes in rice endosperm. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 18729–18734. [[CrossRef](#)] [[PubMed](#)]
11. Saze, H.; Mittelsten Scheid, O.; Paszkowski, J. Maintenance of CpG methylation is essential for epigenetic inheritance during plant gametogenesis. *Nat. Genet.* **2003**, *34*, 65–69. [[CrossRef](#)] [[PubMed](#)]
12. Gazzani, S.; Gendall, A.R.; Lister, C.; Dean, C. Analysis of the molecular basis of flowering time variation in *Arabidopsis* accessions. *Plant Physiol.* **2003**, *132*, 1107–1114. [[CrossRef](#)] [[PubMed](#)]
13. Shi, J.; Dong, A.; Shen, W.-H. Epigenetic regulation of rice flowering and reproduction. *Front. Plant Sci.* **2014**, *5*, 803. [[CrossRef](#)] [[PubMed](#)]
14. Xing, M.-Q.; Zhang, Y.-J.; Zhou, S.-R.; Hu, W.-Y.; Wu, X.-T.; Ye, Y.-J.; Wu, X.-X.; Xiao, Y.-P.; Li, X.; Xue, H.-W. Global analysis reveals the crucial roles of DNA methylation during rice seed development. *Plant Physiol.* **2015**, *168*, 1417–1432. [[CrossRef](#)] [[PubMed](#)]
15. Weigel, D.; Colot, V. Epialleles in plant evolution. *Genome Biol.* **2012**, *13*, 249. [[CrossRef](#)] [[PubMed](#)]
16. Quadrana, L.; Colot, V. Plant transgenerational epigenetics. *Annu. Rev. Genet.* **2016**, *50*, 467–491. [[CrossRef](#)] [[PubMed](#)]
17. Kawakatsu, T.; Huang, S.-S.C.; Jupe, F.; Sasaki, E.; Schmitz, R.J.; Urich, M.A.; Castanon, R.; Nery, J.R.; Barragan, C.; He, Y.; et al. Epigenomic diversity in a global collection of *Arabidopsis thaliana* accessions. *Cell* **2016**, *166*, 492–505. [[CrossRef](#)] [[PubMed](#)]
18. Mirouze, M.; Paszkowski, J. Epigenetic contribution to stress adaptation in plants. *Curr. Opin. Plant Biol.* **2011**, *14*, 267–274. [[CrossRef](#)] [[PubMed](#)]
19. Secco, D.; Wang, C.; Shou, H.; Schultz, M.D.; Chiarenza, S. Stress induced gene expression drives transient DNA methylation changes at adjacent repetitive elements. *eLife* **2015**, *4*, e09343. [[CrossRef](#)] [[PubMed](#)]
20. Garg, R.; Narayana Chevala, V.; Shankar, R.; Jain, M. Divergent DNA methylation patterns associated with gene expression in rice cultivars with contrasting drought and salinity stress response. *Sci. Rep.* **2015**, *5*, 14922. [[CrossRef](#)] [[PubMed](#)]
21. Wang, W.; Qin, Q.; Sun, F.; Wang, Y.; Xu, D.; Li, Z. Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription. *Front. Plant Sci.* **2016**, *39*, 61–69. [[CrossRef](#)]
22. Zheng, X.; Chen, L.; Xia, H.; Wei, H.; Lou, Q.; Li, M.; Li, T.; Luo, L. Transgenerational epimutations induced by multi-generation drought imposition mediate rice plant's adaptation to drought condition. *Sci. Rep.* **2017**, *7*, 39843. [[CrossRef](#)] [[PubMed](#)]
23. Feng, S.J.; Liu, X.S.; Tao, H.; Tan, S.K.; Chu, S.S.; Oono, Y.; Zhang, X.D.; Chen, J.; Yang, Z.M. Variation of DNA methylation patterns associated with gene expression in rice (*Oryza sativa*) exposed to cadmium. *Plant Cell Environ.* **2016**, *39*, 2629–2649. [[CrossRef](#)] [[PubMed](#)]
24. Lu, Y.C.; Feng, S.J.; Zhang, J.J.; Luo, F.; Zhang, S. Genome-wide identification of DNA methylation provides insights into the association of gene expression in rice exposed to pesticide atrazine. *Sci. Rep.* **2016**, *6*, 18985. [[CrossRef](#)] [[PubMed](#)]
25. Quadrana, L.; Bortolini Silveira, A.; Mayhew, G.F.; LeBlanc, C.; Martienssen, R.A.; Jeddeloh, J.A.; Colot, V. The *Arabidopsis thaliana* mobilome and its impact at the species level. *eLife* **2016**, *5*, e15176. [[CrossRef](#)] [[PubMed](#)]
26. Lisch, D. How important are transposons for plant evolution? *Nat. Rev. Genet.* **2013**, *14*, 49–61. [[CrossRef](#)] [[PubMed](#)]
27. Martin, A.; Troadec, C.; Boualem, A.; Rajab, M.; Fernandez, R.; Morin, H.; Pitrat, M.; Dogimont, C.; Bendahmane, A. A transposon-induced epigenetic change leads to sex determination in melon. *Nature* **2009**, *461*, 1135–1138. [[CrossRef](#)] [[PubMed](#)]
28. Ong-Abdullah, M.; Ordway, J.M.; Jiang, N.; Ooi, S.-E.; Kok, S.-Y.; Sarpan, N.; Azimi, N.; Hashim, A.T.; Ishak, Z.; Rosli, S.K.; et al. Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm. *Nature* **2015**, *525*, 533–537. [[CrossRef](#)] [[PubMed](#)]
29. Chuong, E.B.; Elde, N.C.; Feschotte, C. Regulatory activities of transposable elements: From conflicts to benefits. *Nat. Rev. Genet.* **2016**, *18*, 71–86. [[CrossRef](#)] [[PubMed](#)]
30. Rey, O.; Danchin, E.; Mirouze, M.; Loot, C.; Blanchet, S. Adaptation to global change: A transposable element-epigenetics perspective. *Trends Ecol. Evol.* **2016**, *31*, 514–526. [[CrossRef](#)] [[PubMed](#)]

31. Song, X.; Cao, X. Transposon-mediated epigenetic regulation contributes to phenotypic diversity and environmental adaptation in rice. *Curr. Opin. Plant Biol.* **2017**, *36*, 111–118. [[CrossRef](#)] [[PubMed](#)]
32. Reinders, J.; Wulff, B.B.H.; Mirouze, M.; Mari-Ordóñez, A.; Dapp, M.; Rozhon, W.; Bucher, E.; Theiler, G.; Paszkowski, J. Compromised stability of DNA methylation and transposon immobilization in mosaic *Arabidopsis* epigenomes. *Genes Dev.* **2009**, *23*, 939–950. [[CrossRef](#)] [[PubMed](#)]
33. Johannes, F.; Porcher, E.; Teixeira, F.K.; Saliba-Colombani, V.; Simon, M.; Agier, N.; Bulski, A.; Albuissou, J.; Heredia, F.; Audigier, P.; et al. Assessing the impact of transgenerational epigenetic variation on complex traits. *PLoS Genet.* **2009**, *5*, e1000530. [[CrossRef](#)] [[PubMed](#)]
34. Stroud, H.; Ding, B.; Simon, S.A.; Feng, S.; Bellizzi, M.; Pellegrini, M.; Wang, G.-L.; Meyers, B.C.; Jacobsen, S.E. Plants regenerated from tissue culture contain stable epigenome changes in rice. *eLife* **2013**, *2*, e00354. [[CrossRef](#)] [[PubMed](#)]
35. Finnegan, E.J.; Dennis, E.S. Isolation and identification by sequence homology of a putative cytosine methyltransferase from *Arabidopsis thaliana*. *Nucleic Acids Res.* **1993**, *21*, 2383–2388. [[CrossRef](#)] [[PubMed](#)]
36. Pavlopoulou, A.; Kossida, S. Plant cytosine-5 DNA methyltransferases: Structure, function, and molecular evolution. *Genomics* **2007**, *90*, 530–541. [[CrossRef](#)] [[PubMed](#)]
37. Hu, L.; Li, N.; Xu, C.; Zhong, S.; Lin, X.; Yang, J.; Zhou, T.; Yuliang, A.; Wu, Y.; Chen, Y.-R.; et al. Mutation of a major CG methylase in rice causes genome-wide hypomethylation, dysregulated genome expression, and seedling lethality. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 10642–10647. [[CrossRef](#)] [[PubMed](#)]
38. Yamauchi, T.; Moritoh, S.; Johzuka-Hisatomi, Y.; Ono, A.; Terada, R.; Nakamura, I.; Iida, S. Alternative splicing of the rice *OsMET1* genes encoding maintenance DNA methyltransferase. *Plant Physiol.* **2008**, *165*, 1774–1782. [[CrossRef](#)] [[PubMed](#)]
39. Zemach, A.; Kim, M.Y.; Hsieh, P.-H.; Coleman-Derr, D.; Eshed-Williams, L.; Thao, K.; Harmer, S.L.; Zilberman, D. The *Arabidopsis* nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin. *Cell* **2013**, *153*, 193–205. [[CrossRef](#)] [[PubMed](#)]
40. Lindroth, A.M.; Cao, X.; Jackson, J.P.; Zilberman, D.; McCallum, C.M.; Henikoff, S.; Jacobsen, S.E. Requirement of CHROMOMETHYLASE3 for maintenance of CpXpG methylation. *Science* **2001**, *292*, 2077–2080. [[CrossRef](#)] [[PubMed](#)]
41. Stroud, H.; Do, T.; Du, J.; Zhong, X.; Feng, S.; Johnson, L.; Patel, D.J.; Jacobsen, S.E. Non-CG methylation patterns shape the epigenetic landscape in *Arabidopsis*. *Nat. Struct. Mol. Biol.* **2014**, *21*, 64–72. [[CrossRef](#)] [[PubMed](#)]
42. Ashapkin, V.V.; Kutueva, L.I.; Vanyushin, B.F. Plant DNA methyltransferase genes: Multiplicity, expression, methylation patterns. *Biochemistry* **2016**, *81*, 141–151. [[CrossRef](#)] [[PubMed](#)]
43. Cheng, C.; Tarutani, Y.; Miyao, A.; Ito, T.; Yamazaki, M.; Sakai, H.; Fukai, E.; Hirochika, H. Loss of function mutations in the rice chromomethylase *OsCMT3a* cause a burst of transposition. *Plant J.* **2015**, *83*, 1069–1081. [[CrossRef](#)] [[PubMed](#)]
44. Sharma, R.; Singh, R.M.; Malik, G. Rice cytosine DNA methyltransferases-gene expression profiling during reproductive development and abiotic stress. *FEBS J.* **2009**, *276*, 6301–6311. [[CrossRef](#)] [[PubMed](#)]
45. Moritoh, S.; Eun, C.-H.; Ono, A.; Asao, H.; Okano, Y.; Yamaguchi, K.; Shimatani, Z.; Koizumi, A.; Terada, R. Targeted disruption of an orthologue of DOMAINS REARRANGED METHYLASE 2, *OsDRM2*, impairs the growth of rice plants by abnormal DNA methylation. *Plant J.* **2012**, *71*, 85–98. [[CrossRef](#)] [[PubMed](#)]
46. Tan, F.; Zhou, C.; Zhou, Q.; Zhou, S.; Yang, W.; Zhao, Y.; Li, G.; Zhou, D.-X. Analysis of chromatin regulators reveals specific features of rice DNA methylation pathways. *Plant Physiol.* **2016**, *171*, 2041–2054. [[CrossRef](#)] [[PubMed](#)]
47. Higo, H.; Tahir, M.; Takashima, K.; Miura, A.; Watanabe, K.; Tagiri, A.; Ugaki, M.; Ishikawa, R.; Eiguchi, M.; Kurata, N.; et al. DDM1 (decrease in DNA methylation) genes in rice (*Oryza sativa*). *Mol. Genet. Genom.* **2012**, *287*, 785–792. [[CrossRef](#)] [[PubMed](#)]
48. Pang, J.; Dong, M.; Li, N.; Zhao, Y.; Liu, B. Functional characterization of a rice de novo DNA methyltransferase, *OsDRM2*, expressed in *Escherichia coli* and yeast. *Biochem. Biophys. Res. Commun.* **2013**, *432*, 157–162. [[CrossRef](#)] [[PubMed](#)]
49. Dangwal, M.; Malik, G.; Kapoor, S.; Kapoor, M. De novo methyltransferase, *OsDRM2*, interacts with the ATP-dependent RNA helicase, *OseIF4A*, in rice. *J. Mol. Biol.* **2013**, *425*, 2853–2866. [[CrossRef](#)] [[PubMed](#)]

50. Kapoor, M.; Arora, R.; Lama, T.; Nijhawan, A.; Khurana, J.P.; Tyagi, A.K.; Kapoor, S. Genome-wide identification, organization and phylogenetic analysis of Dicer-like, Argonaute and RNA-dependent RNA Polymerase gene families and their expression analysis during reproductive development and stress in rice. *BMC Genom.* **2008**, *9*, 451. [[CrossRef](#)] [[PubMed](#)]
51. Wei, L.; Gu, L.; Song, X.; Cui, X.; Lu, Z.; Zhou, M.; Wang, L.; Hu, F.; Zhai, J.; Meyers, B.C.; et al. Dicer-like 3 produces transposable element-associated 24-nt siRNAs that control agricultural traits in rice. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 3877–3882. [[CrossRef](#)] [[PubMed](#)]
52. Wu, L.; Zhou, H.; Zhang, Q.; Zhang, J.; Ni, F.; Liu, C.; Qi, Y. DNA methylation mediated by a microRNA pathway. *Mol. Cell* **2010**, *38*, 465–475. [[CrossRef](#)] [[PubMed](#)]
53. Song, X.; Wang, D.; Ma, L.; Chen, Z.; Li, P.; Cui, X.; Liu, C.; Cao, S.; Chu, C.; Tao, Y.; et al. Rice RNA-dependent RNA polymerase 6 acts in small RNA biogenesis and spikelet development. *Plant J.* **2012**, *71*, 378–389. [[CrossRef](#)] [[PubMed](#)]
54. Liu, B.; Chen, Z.; Song, X.; Liu, C.; Cui, X.; Zhao, X.; Fang, J.; Xu, W.; Zhang, H.; Wang, X.; et al. *Oryza sativa* dicer-like4 reveals a key role for small interfering RNA silencing in plant development. *Plant Cell* **2007**, *19*, 2705–2718. [[CrossRef](#)] [[PubMed](#)]
55. Wang, N.; Zhang, D.; Wang, Z.; Xun, H.; Ma, J.; Wang, H.; Huang, W.; Liu, Y.; Lin, X.; Li, N.; et al. Mutation of the *RDR1* gene caused genome-wide changes in gene expression, regional variation in small RNA clusters and localized alteration in DNA methylation in rice. *BMC Plant Biol.* **2014**, *14*, 177. [[CrossRef](#)] [[PubMed](#)]
56. Hong, W.; Qian, D.; Sun, R.; Jiang, L.; Wang, Y.; Wei, C.; Zhang, Z.; Li, Y. OsRDR6 plays role in host defense against double-stranded RNA virus, *Rice Dwarf Phytoreovirus*. *Sci. Rep.* **2015**, *5*, 11324. [[CrossRef](#)] [[PubMed](#)]
57. Abe, M.; Yoshikawa, T.; Nosaka, M.; Sakakibara, H.; Sato, Y.; Nagato, Y.; Itoh, J.-I. *WAVY LEAF1*, an ortholog of Arabidopsis *HEN1*, regulates shoot development by maintaining microRNA and trans-acting small interfering RNA accumulation in rice. *Plant Physiol.* **2010**, *154*, 1335–1346. [[CrossRef](#)] [[PubMed](#)]
58. Ono, A.; Yamaguchi, K.; Fukada-Tanaka, S.; Terada, R.; Mitsui, T.; Iida, S. A null mutation of *ROS1a* for DNA demethylation in rice is not transmittable to progeny. *Plant J.* **2012**, *71*, 564–574. [[CrossRef](#)] [[PubMed](#)]
59. La, H.; Ding, B.; Mishra, G.P.; Zhou, B.; Yang, H.; Bellizzi, M.D.R.; Chen, S.; Meyers, B.C.; Peng, Z.; Zhu, J.-K.; Wang, G.-L. A 5-methylcytosine DNA glycosylase/lyase demethylates the retrotransposon *Tos17* and promotes its transposition in rice. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 15498–15503. [[CrossRef](#)] [[PubMed](#)]
60. Cao, X.; Jacobsen, S.E. Role of the Arabidopsis DRM methyltransferases in de novo DNA methylation and gene silencing. *Curr. Biol.* **2002**, *12*, 1138–1144. [[CrossRef](#)]
61. Matzke, M.A.; Mosher, R.A. RNA-directed DNA methylation: An epigenetic pathway of increasing complexity. *Nat. Rev. Genet.* **2014**, *15*, 394–408. [[CrossRef](#)] [[PubMed](#)]
62. Cao, X.; Jacobsen, S.E. Locus-specific control of asymmetric and CpNpG methylation by the *DRM* and *CMT3* methyltransferase genes. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 16491–16498. [[CrossRef](#)] [[PubMed](#)]
63. Cao, X.; Springer, N.M.; Muszynski, M.G.; Phillips, R.L.; Kaeppler, S.; Jacobsen, S.E. Conserved plant genes with similarity to mammalian de novo DNA methyltransferases. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 4979–4984. [[CrossRef](#)] [[PubMed](#)]
64. Jeddelloh, J.A.; Stokes, T.L.; Richards, E.J. Maintenance of genomic methylation requires a SWI2/SNF2-like protein. *Nat. Genet.* **1999**, *22*, 94–97. [[PubMed](#)]
65. Brzeski, J.; Jerzmanowski, A. Deficient in DNA methylation 1 (DDM1) defines a novel family of chromatin-remodeling factors. *J. Biol. Chem.* **2003**, *278*, 823–828. [[CrossRef](#)] [[PubMed](#)]
66. Lippman, Z.; Gendrel, A.-V.; Black, M.; Vaughn, M.W.; Dedhia, N.; McCombie, W.R.; Lavigne, K.; Mittal, V.; May, B.; Kasschau, K.D.; et al. Role of transposable elements in heterochromatin and epigenetic control. *Nature* **2004**, *430*, 471–476. [[CrossRef](#)] [[PubMed](#)]
67. Agius, F.; Kapoor, A.; Zhu, J.-K. Role of the Arabidopsis DNA glycosylase/lyase ROS1 in active DNA demethylation. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 11796–11801. [[CrossRef](#)] [[PubMed](#)]
68. Choi, Y.; Gehring, M.; Johnson, L.; Hannon, M.; Harada, J.J.; Goldberg, R.B.; Jacobsen, S.E.; Fischer, R.L. DEMETER, a DNA glycosylase domain protein, is required for endosperm gene imprinting and seed viability in arabidopsis. *Cell* **2002**, *110*, 33–42. [[CrossRef](#)]
69. Penterman, J.; Zilberman, D.; Huh, J.H.; Ballinger, T.; Henikoff, S.; Fischer, R.L. DNA demethylation in the Arabidopsis genome. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 6752–6757. [[CrossRef](#)] [[PubMed](#)]

70. Mirouze, M.; Vitte, C. Transposable elements, a treasure trove to decipher epigenetic variation: Insights from *Arabidopsis* and crop epigenomes. *J. Exp. Bot.* **2014**, *65*, 2801–2812. [[CrossRef](#)] [[PubMed](#)]
71. Li, Q.; Eichten, S.R.; Hermanson, P.J.; Zaunbrecher, V.M.; Song, J.; Wendt, J.; Rosenbaum, H.; Madzima, T.F.; Sloan, A.E.; Huang, J.; et al. Genetic perturbation of the maize methylome. *Plant Cell* **2014**, *26*, 4602–4616. [[CrossRef](#)] [[PubMed](#)]
72. Bucher, E.; Reinders, J.; Mirouze, M. Epigenetic control of transposon transcription and mobility in *Arabidopsis*. *Curr. Opin. Plant Biol.* **2012**, *15*, 503–510. [[CrossRef](#)] [[PubMed](#)]
73. Cuerda-Gil, D.; Slotkin, R.K. Non-canonical RNA-directed DNA methylation. *Nat. Plants* **2016**, *2*, 16163. [[CrossRef](#)] [[PubMed](#)]
74. Lahmy, S.; Pontier, D.; Bies-Etheve, N.; Laudíe, M.; Feng, S.; Jobet, E.; Hale, C.J.; Cooke, R.; Hakimi, M.-A.; Angelov, D.; et al. Evidence for ARGONAUTE4–DNA interactions in RNA-directed DNA methylation in plants. *Genes Dev.* **2016**, *30*, 2565–2570. [[CrossRef](#)] [[PubMed](#)]
75. Panda, K.; Ji, L.; Neumann, D.A.; Daron, J.; Schmitz, R.J.; Slotkin, R.K. Full-length autonomous transposable elements are preferentially targeted by expression-dependent forms of RNA-directed DNA methylation. *Genome Biol.* **2016**, *17*, 170. [[CrossRef](#)] [[PubMed](#)]
76. Arikiti, S.; Zhai, J.; Meyers, B.C. Biogenesis and function of rice small RNAs from non-coding RNA precursors. *Curr. Opin. Plant Biol.* **2013**, *16*, 170–179. [[CrossRef](#)] [[PubMed](#)]
77. Ream, T.S.; Haag, J.R.; Wierzbicki, A.T.; Nicora, C.D.; Norbeck, A.D.; Zhu, J.-K.; Hagen, G.; Guilfoyle, T.J.; Pasa-Tolić, L.; Pikaard, C.S. Subunit compositions of the RNA-silencing enzymes Pol IV and Pol V reveal their origins as specialized forms of RNA polymerase II. *Mol. Cell* **2009**, *33*, 192–203. [[CrossRef](#)] [[PubMed](#)]
78. Tucker, S.L.; Reece, J.; Ream, T.S.; Pikaard, C.S. Evolutionary history of plant multisubunit RNA polymerases IV and V: Subunit origins via genome-wide and segmental gene duplications, retrotransposition, and lineage-specific subfunctionalization. *Cold Spring Harb. Symp. Quant. Biol.* **2010**, *75*, 285–297. [[CrossRef](#)] [[PubMed](#)]
79. Huang, Y.; Kendall, T.; Forsythe, E.S.; Dorantes-Acosta, A.; Li, S.; Caballero-Pérez, J.; Chen, X.; Arteaga-Vázquez, M.; Beilstein, M.A.; Mosher, R.A. Ancient origin and recent innovations of RNA polymerase IV and V. *Mol. Biol. Evol.* **2015**, *32*, 1788–1799. [[CrossRef](#)] [[PubMed](#)]
80. Bousios, A.; Gaut, B.S. Mechanistic and evolutionary questions about epigenetic conflicts between transposable elements and their plant hosts. *Curr. Opin. Plant Biol.* **2016**, *30*, 123–133. [[CrossRef](#)] [[PubMed](#)]



© 2017 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

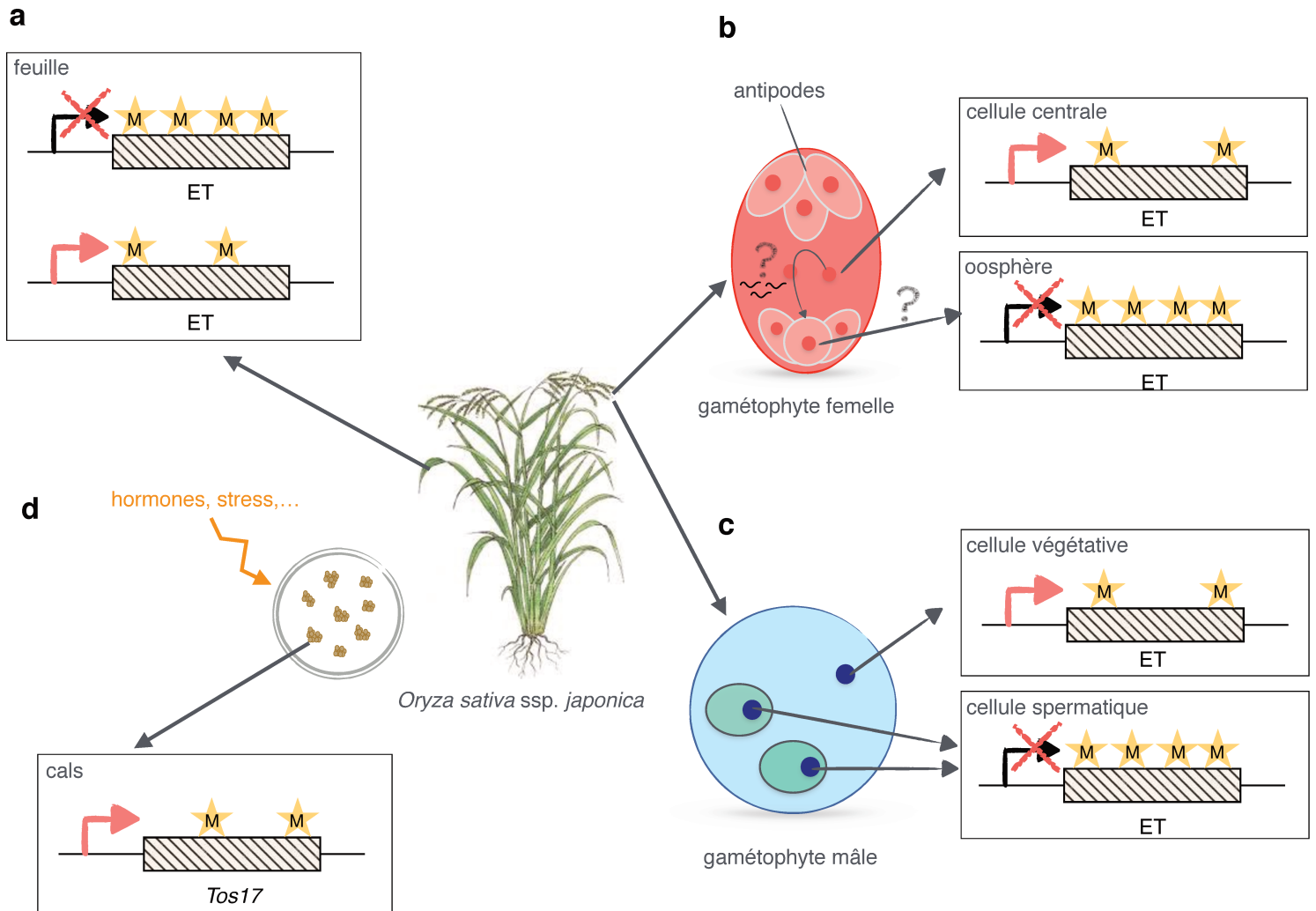


Figure 5. Modèle pour le contrôle de l'activité des ET au cours du développement et en conditions de stress chez le riz. (a) Durant le développement végétatif, la plupart des ET sont épigénétiquement contrôlés (étoiles: méthylation de l'ADN) et non ou faiblement transcrits. (b) De plus, le modèle actuel sur le contrôle épigénétique des ET propose que la transmission des marques épigénétiques à la génération suivante implique une ré-activation transcriptionnelle transitoire des ET dans les cellules centrales du gamétophyte femelle et (c) dans la cellule végétative du gamétophyte mâle (grain de pollen) (Martinez et Kohler, 2017). Le modèle suggère que la réactivation transcriptionnelle des ET conduit à la production de petits ARN (siRNA) de 21nt et de 24nt qui migrent respectivement vers l'oosphère et la cellule spermatique afin de cibler les ET à inactiver dans les lignées germinales. Les points d'interrogations soulignent les incertitudes du modèle (par exemple le mouvement des petits ARNs dans le gamétophyte femelle). (d) Les conditions de stress telles que la culture *in vitro* peuvent également affecter l'épigénome conduisant à l'activation d'ET tel que *Tos17* (Stroud et al., 2013).

2.2 Réactivation des éléments transposables au cours du développement

De par la forte répression épigénétique exercée sur les ET, peu de familles d'ET semblent capables d'être actives au cours du développement d'une plante. (Figure 5). Aujourd'hui, nous connaissons peu d'exemples d'ET actifs dans une plante sauvage et en conditions normales. Néanmoins, certains facteurs développementaux sont capables d'activer la transcription de certains éléments (Figure 5). Par exemple, chez le maïs, le transposon à ADN MuDR est transitoirement exprimé durant le développement végétatif de la plante (Li et al. 2010). Cette réactivation est expliquée par une perte transitoire de la méthylation de l'ADN au loci de MuDR durant la transition de la plante juvénile vers la plante adulte (Li et al. 2010). Aucune transposition cependant n'a été observée dans ce cas, la transposase n'étant pas détectée.

Durant la phase de reproduction sexuelle des plantes à fleurs, la transmission des marques épigénétiques d'une génération à une autre nécessite une reprogrammation épigénétique autrement dit un changement global de la méthylation de l'ADN et des modifications des histones (Kawashima et al. 2015). Le gamétophyte mâle est constitué d'une cellule végétative haploïde qui formera le tube pollinique et de deux cellules spermatiques haploïdes qui sont les gamètes mâles (Figure 5). Le gamétophyte femelle est constitué de trois cellules antipodes, de deux synergides qui entourent l'oosphère haploïde, le gamète femelle, et d'une cellule centrale qui contient deux noyaux polaires (Figure 5). Durant la fécondation, une des deux cellules spermatiques féconde l'oosphère et la deuxième cellule spermatique fusionne avec les deux noyaux polaires de la cellule centrale. Cette double fécondation conduit à la formation d'un albumen triploïde et d'un embryon diploïde (Raghavan 2003).

Chez *Arabidopsis thaliana* l'analyse du méthylome des cellules végétatives du gamétophyte mâle (grain de pollen) a montré que la méthylation en contexte CG était réduite alors que la méthylation en contexte CHH augmentait (Figure 5). Cette hypométhylation est induite par une sous-régulation de *DDMI* (*Decrease in DNA Methylation 1*) et par la surexpression de *DME* (*DEMETTER*), qui conduit à la réactivation transcriptionnelle des ET (Slotkin et al. 2009; Schoft et al. 2011). Park et al. (2016) ont montré par l'analyse du méthylome de la cellule centrale que de façon similaire à la cellule végétative, la cellule centrale subit une hypométhylation en CG mais que cette hypométhylation n'est pas globale suggérant une déméthylation locus-spécifique. En revanche, le niveau de méthylation en CHH semble similaire à celui de la cellule végétative. Dernièrement, Ingouff et al. (2017) ont développé un senseur fluorescent de méthylation (DYNAMETs) qui permet de suivre la méthylation en CG et CHH en temps réel à

l'échelle d'une cellule chez *Arabidopsis*. Leur étude a notamment révélé que la méthylation de l'oosphère subissait quelques fluctuations au niveau de la méthylation en CG aux premiers stades du développement tandis que le niveau de méthylation en CHH est relativement stable. Pour finir, après la double fécondation, le niveau de méthylation dans les contextes est relativement réduit dans l'albumen par rapport à l'embryon (Hsieh et al. 2009; Ibarra et al. 2012). Park et al. (2016) suggèrent que l'initiation de cette hypométhylation est induite par l'hypométhylation locus-spécifique observée dans la cellule centrale. Ces études soulignent la dynamique de l'épigénome au cours du développement.

Le modèle aujourd'hui proposé suggère que la réactivation des ET dans ces deux tissus non transmis à la génération suivante permettrait l'induction de petits ARN de 21 nucléotides (nt) qui migreraient dans l'oosphère et dans les cellules spermatiques afin de cibler la méthylation d'ADN *de novo* dans les tissus transmis à la génération suivante (Slotkin et al. 2009; Martínez et Köhler 2017). Ce modèle reste toutefois controversé et a été à plusieurs reprises réfuté. Par exemple, l'expression de microARN (miARN) artificiels spécifiquement exprimés dans les cellules végétatives du gamétophyte mâle a montré que ces miARN n'étaient pas capables d'induire du *silencing* dans la cellule spermatique remettant en cause le modèle précédemment établi (Grant-Downton et al. 2013). De plus, aucune connexion cytoplasmique directe n'a été observée entre la cellule végétative et la cellule spermatique. Néanmoins, très récemment Martínez et al. (2016) ont montré que les petits ARN interférents (siARN) de 21nt chargés par des protéines AGO étaient capables de migrer de la cellule végétative vers la cellule spermatique dans le grain de pollen contrairement au miARN, validant ainsi le modèle de Slotkin et al. (2009). Outre la méthylation de l'ADN, les variants d'histones transmises par les parents subissent également une reprogrammation dans l'embryon. Effectivement, Ingouff et al. (2010) ont montré que les variants d'histones parentaux étaient remplacés par des variants d'histones H3 synthétisés *de novo* suggérant une reprogrammation des marques chromatiniennes quelques heures après la fécondation et qui pourrait initier le développement de l'embryon.

Le processus de reprogrammation épigénétique semble être conservé chez les plantes car la transcription des ET (Anderson et al. 2013) et l'hypométhylation ont également été observées dans les tissus des gamétophytes mâles et femelles ainsi que dans l'albumen (après la fécondation) chez différentes espèces de riz (Zemach et al. 2010; Xing et al. 2015; Park et al.

2016). Toutefois, les mouvements de ces ARN dans le gamétophyte femelle doivent être confirmés et de nombreux travaux doivent être réalisés dans le but de préciser le rôle et le devenir de ces petits ARN après la fécondation (Martínez et Köhler 2017). Finalement, l'activité transcriptionnelle des ET lors du développement ne conduit pas nécessairement à la transposition d'éléments et à ce jour, seule une activité transpositionnelle d'un élément MULE, *AtMula*, dans les grains de pollen d'*Arabidopsis thaliana* a pu être observée (Slotkin et al. 2009).

2.3 Réactivation des éléments transposables en conditions de stress

Comme précédemment mentionné, la méthylation de l'ADN peut être modifiée en réponse à des conditions environnementales. Ces changements de méthylation peuvent être à l'origine de l'activité de certains ET. L'exemple le plus documenté est la réactivation de *Tos17* chez le riz lors de la culture prolongée de cals (prolifération cellulaire obtenue par culture *in vitro*) (Hirochika et al. 1996). Une perte de méthylation au locus de *Tos17* expliquerait sa mobilité (Stroud et al. 2013). La réactivation de cet élément dans les cals a notamment permis la construction de banques de mutants chez le riz (Jeon et al. 2000; Sallaud et al. 2003; Wei et al. 2013). De façon surprenante, cette perte de méthylation est spécifique à certaines familles et le méthylome est affecté durablement à certains loci, même après régénération des plantules, bien que le mécanisme impliqué ne soit pas encore connu (Stroud et al. 2013).

D'autres exemples montrent que les ET ont acquis la capacité à se réactiver en conditions de stress. Par exemple, *ONSEN*, un des rétrotransposons à LTR les plus étudiés chez *Arabidopsis thaliana* est activé en réponse à un stress thermique (Ito et al. 2011). Un élément de réponse à la chaleur est présent dans le LTR d'*ONSEN* (Cavrak et al. 2014) et lors d'un stress thermique, le facteur de transcription (FT) HRE (*heat shock factor A2*) se lierait au LTR d'*ONSEN* induisant la transcription de l'élément. Néanmoins, les néo-insertions d'*ONSEN* sont détectées uniquement après un stress thermique dans des plantes mutées pour la voie du RdDM (par exemple mutées pour pol IV) et non chez les plantes sauvages (Ito et al. 2011). D'autre part, des études ont montré que certains éléments ont également acquis des séquences *cis* régulatrices leur permettant de contourner les mécanismes de contrôle. Chez le tabac par exemple, les rétrotransposons à LTR *Tnt1* et *Tto1* sont activés spécifiquement en réponse à une attaque de pathogènes. Ils possèdent dans leurs promoteurs le motif répété CCAACC(N)₇CT, homologue

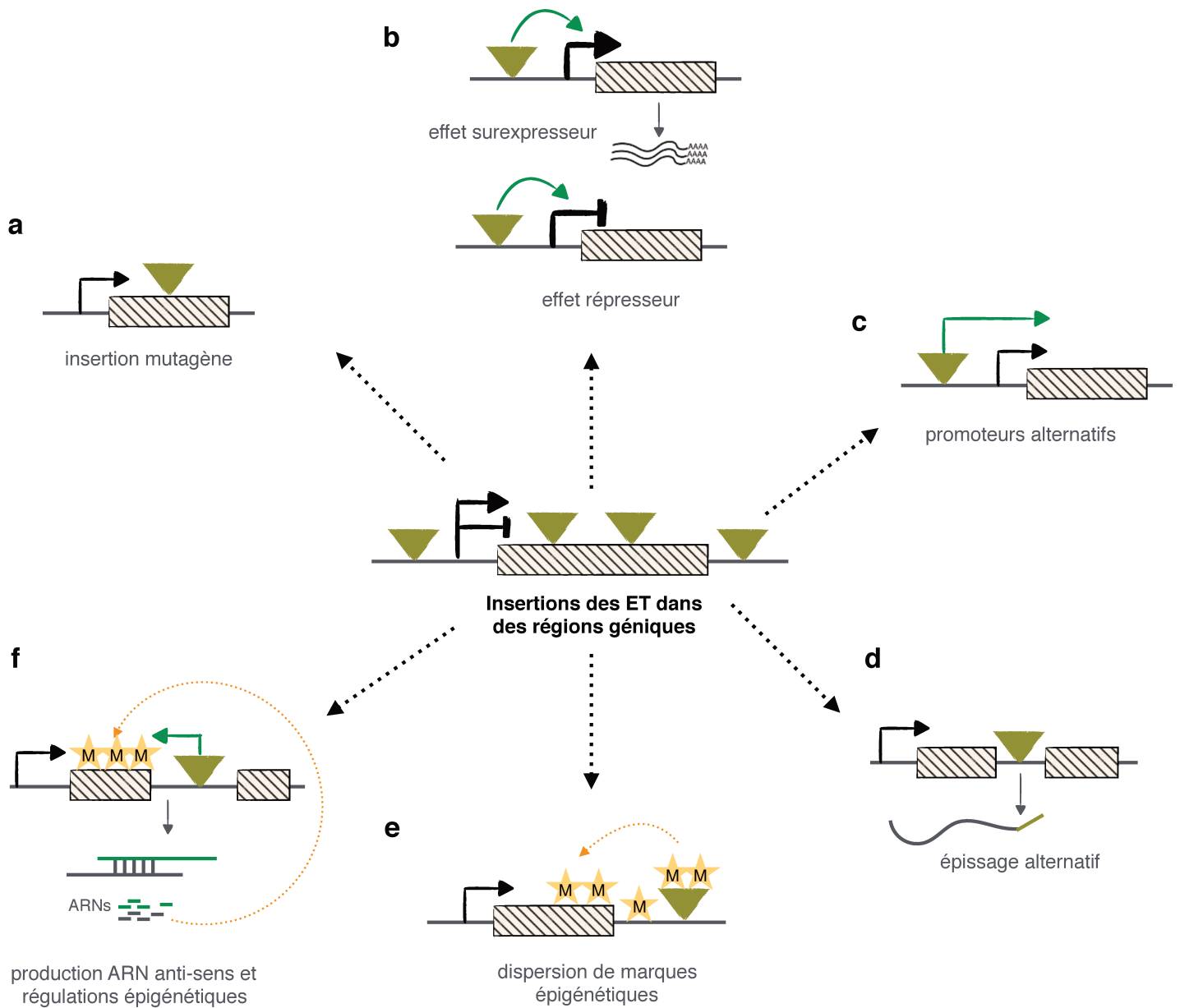


Figure 6. Impact des ET sur les régulations transcriptionnelles des gènes. Par leur mobilité et leur activité promotrice, les ET peuvent être mutagènes (a) et induire des changements de régulations transcriptionnelles (b,c,d). Les ET attirent également des marques épigénétiques répressives et peuvent donc avoir un rôle comme régulateur épigénétique (e, f). Un triangle vert symbolise l'insertion d'un ET et une étoile orange symbolise la méthylation de l'ADN. Figure adaptée de Gogvadze et Buzdin (2009) et Wei et al. (2016).

de la séquence dite *H-box* qui est présente dans de nombreux gènes impliqués dans les mécanismes de défenses des plantes (Mhiri et al. 1997; Hirochika 1997). Il semblerait que ces éléments *cis*-régulateurs puissent expliquer la spécificité d'activité de ces deux rétrotransposons.

En définitive, différents exemples cités ci-dessus montrent que même si l'ET est transcrit, la transposition de l'élément n'est pas automatique. C'est souvent la combinaison entre deux facteurs : une levée de *silencing* et un stimulus activateur (par exemple une perte de méthylation et un facteur environnemental) qui permettront la transposition et donc la présence de néo-insertions de l'élément dans le génome hôte (Grandbastien 2015) comme décrit ci-dessus pour *ONSEN* chez *Arabidopsis* (Nozawa et al. 2017).

3. Rôle des éléments transposables au sein des génomes eucaryotes

Les ET sont-ils bénéfiques ou néfastes pour le génome hôte ? Depuis la découverte des ET dans les génomes, la question de leurs rôles et de leurs effets sur les génomes hôtes fait débat dans la communauté. Très tôt certains scientifiques ont suspecté que les ET pouvaient être impliqués dans la régulation transcriptionnelle des gènes (Davidson et Britten 1979; McClintock 1984). Pourtant, cette idée a longtemps été effacée par la pensée dominante de l'époque qui réduisait les ET au rang de simples parasites, sans fonction biologique (Orgel et Crick 1980). Cette hypothèse était soutenue par leur répllication dite « égoïste » et par leurs liens phylogénétiques étroits avec les rétrovirus (Lerat et Capy 1999). Aujourd'hui, les études sur les rôles des ET se multiplient et nul ne peut ignorer l'implication majeure des ET dans l'évolution des organismes (Feschotte et Pritham 2007; Lisch 2013; Chuong et al. 2016a). Les insertions d'ET dans des régions codantes ont un fort potentiel mutagène et peuvent conduire à des changements de régulation transcriptionnelle et post-transcriptionnelle tels que des épissages alternatifs, la création de nouveaux promoteurs et de nouvelles marques épigénétiques conduisant à des effets activateurs ou répresseurs sur le niveau d'expression des gènes affectés (Figure 6) (Gogvadze et Buzdin 2009; Wei et Cao 2016). Dans cette partie, nous verrons quelques exemples choisis pour illustrer les impacts des ET sur les génomes hôtes et tenter de répondre à cette brûlante question.

3.1 Les éléments transposables : une source significative d'éléments régulateurs

Comme mentionné précédemment, les ET sont distribués tout le long des chromosomes et possèdent une riche diversité de séquences promotrices. Aujourd'hui de nombreux indices suggèrent que les ET offrent une nouvelle source significative de promoteurs et d'éléments *cis*-régulateurs pour les gènes eucaryotes (Feschotte 2008; Chuong et *al.* 2016a; Hirsch et Springer 2017). À titre d'exemple, l'analyse des séquences promotrices chez l'homme a montré qu'environ 25% des promoteurs des gènes contenaient des ET dans leur séquence (Jordan et *al.* 2003) alors que chez le riz 58% des gènes étaient associés à des MITE (Lu et *al.* 2012). Récemment, différentes approches ont été développées afin d'identifier les sites d'initiation de la transcription (TSS) basée sur le séquençage des terminaisons 5' des ARN (Morton et *al.* 2014) ou sur des modèles de prédiction développés à partir de larges bases de données de promoteurs (Shahmuradov et *al.* 2017). Ces approches permettent ainsi d'étudier l'origine et l'évolution des réseaux de régulation géniques et ont notamment montré que chez le maïs 1% des TSS identifiés sont des séquences provenant d'ET (Hirsch et Springer 2017). Les ET peuvent également affecter les réseaux de régulations de gènes en créant ou en éliminant des sites de fixation des facteurs de transcription (FT) par le biais de nouvelles insertions (Feschotte 2008). Quelques exemples ont été identifiés chez les animaux et environ 20% des sites de fixation de FT sont associés à des ET dans les cellules embryonnaires humaines (Kunarso et *al.* 2010). Aujourd'hui aucun exemple n'a encore été décrit chez les plantes.

Les ET sont également régulés par des mécanismes épigénétiques qui peuvent jouer un rôle central dans la régulation tissu-spécifique de gènes. Xie et *al.* (2013) ont montré que l'expression tissu-spécifique de certains gènes chez l'homme était corrélée à la présence à proximité d'ET hypométhylés spécifiquement dans le même tissu. En effet, l'analyse de méthylomes issues de 11 types cellulaires humains a mis en évidence l'hypométhylation spécifique de certaines familles d'ET dans un tissu et non dans les 10 autres testés. Ces mêmes ET se trouvaient à proximité de gènes fortement transcrits dans le tissu. Les auteurs ont montré par la construction de rapporteurs GUS que 26 ET hypométhylés présentaient une activité surexpresser et suggèrent que l'expression tissu-spécifique de ces gènes peut être influencée par le statut épigénétique des ET qui les entourent (Xie et *al.* 2013).

Dernièrement, un nouveau mécanisme de régulation médié par un ET a été découvert chez le riz. Cho et Paszkowski (2017) ont identifié un ET, appelé *MIKKI*, fortement transcrit dans les racines et qui présente un site de fixation tronqué d'un miARN miR171. Chez *Arabidopsis*,

miR171 est connu pour cibler les transcrits du gène *SCL21* (*SCARECROW-like 21*) qui code pour un facteur de transcription impliqué dans le développement racinaire. Leur étude a montré que les transcrits de *MIKKI* jouaient le rôle d'éponge à micro ARN. En effet, osa-miR171 (miR171 de *Oryza sativa*) se fixe aux transcrits de *MIKKI* ce qui a pour effet l'accumulation des transcrits du gène *SCL21* dans les racines. Dans les panicules *MIKKI* est réprimé et osa-miR171 cible *SCL21* induisant la dégradation des transcrits. *MIKKI* participe donc à la régulation tissulaire du gène *SCL21*. Cette régulation post-transcriptionnelle semble conservée chez différentes variétés de riz.

L'acquisition de nouvelles séquences régulatrices induite par la dispersion des ET dans les génomes participe à l'évolution des gènes et à l'adaptation des organismes à des changements environnementaux ou développementaux (Song et Cao 2017). Chez le riz, un des exemples les mieux documentés concerne le MITE *mPing*. Une amplification massive de cet élément a été observée par Susan Wessler dans une variété EG4 cultivée dans le nord du Japon (Naito et al. 2006). Comme précédemment évoqué, *mPing* s'insère préférentiellement dans les régions 5' des gènes (Naito et al. 2009). Les auteurs de l'étude ont montré que les gènes comportant *mPing* dans leur séquence étaient surexprimés en réponse à des signaux environnementaux tels que le sel, le froid et la sécheresse (Naito et al. 2009). La multiplication de *mPing* pourrait ainsi avoir contribué à générer de nouveaux réseaux de régulation en réponses à différents stress environnementaux. Ito et al. (2011) ont également montré que chez *Arabidopsis* les insertions d'*ONSEN* (voir Introduction page 16) dans des régions géniques pouvaient induire la surexpression de ces gènes en réponse à un stress thermique. De même chez le maïs, des études transcriptionnelles ont montré que différentes familles d'ET étaient associées à la surexpression de certains gènes en réponses à différents stress abiotiques (Makarevitch et al. 2015).

Il existe aussi des exemples d'ET également impliqués dans la réponse à des stress biotiques. Par exemple, chez la variété indonésienne de riz Tjahaja, le LTR du rétrotransposon *Renovator* modifie l'activité du gène voisin *Pit* en devenant le promoteur du gène et induit sa surexpression (Hayashi et Yoshida 2009; Wei et Cao 2016). Ce gène de la famille NBS-LRR est impliqué dans la résistance au champignon pathogène *Magnaporthe grisea* et sa surexpression conduit à augmenter le spectre de résistance de la variété Tjahaja.

Les ET constituent donc un véritable réservoir de séquences régulatrices dans les génomes eucaryotes. Inévitablement, en dérégulant l'activité transcriptionnelle des gènes, les ET peuvent également avoir de forts impacts sur le phénotype des organismes hôtes et être ainsi à l'origine

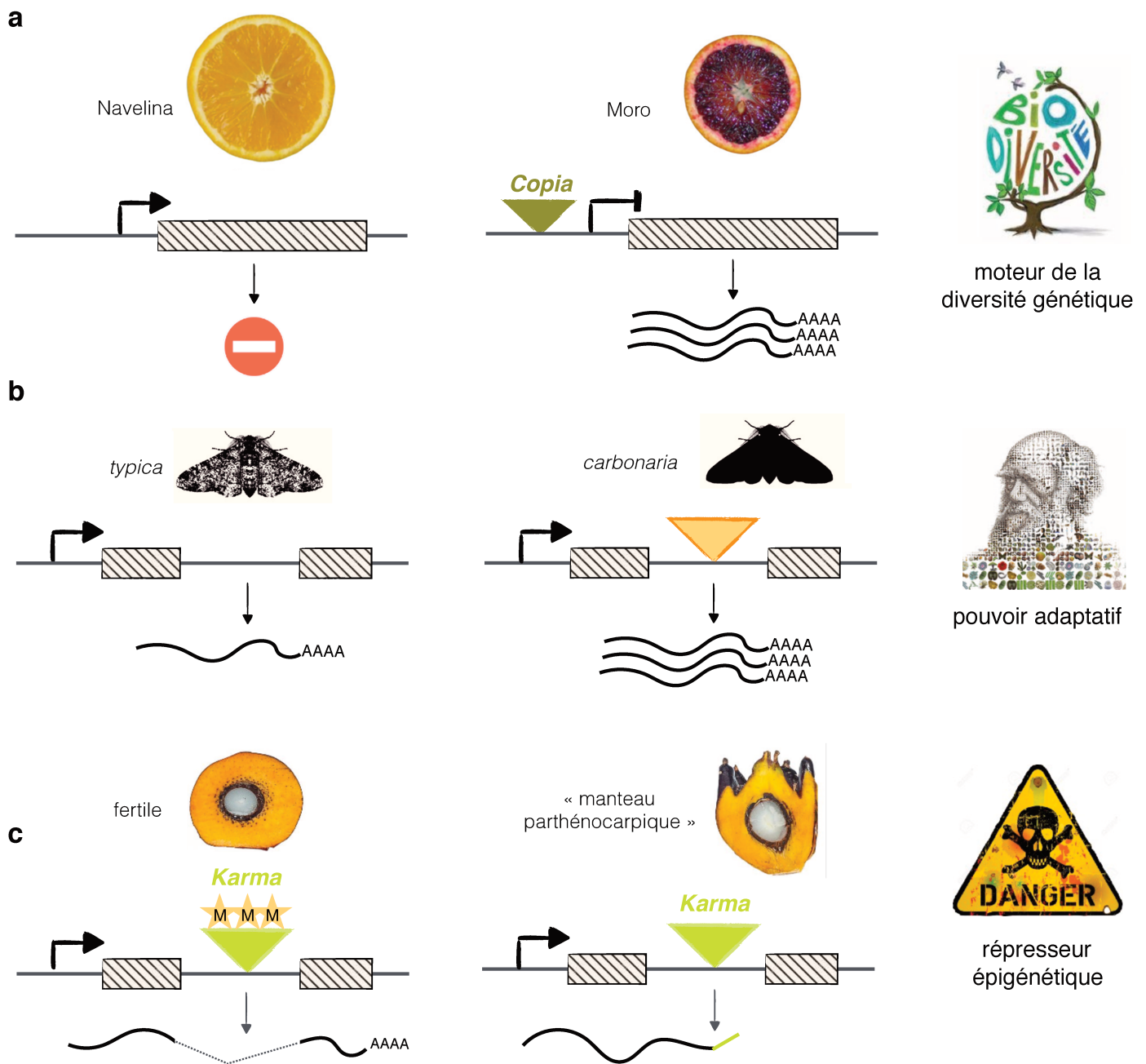


Figure 7. Impacts phénotypiques des ET. (a) Les ET peuvent être mutagènes et ainsi contribuer à la diversité génétique. Par exemple, la couleur sanguine des oranges de la variété Moro est expliquée par la présence d'un rétrotransposon de type *Copia* dans la région promotrice du gène *Ruby*, régulateur de la biosynthèse d'anthocyanes. L'insertion du rétrotransposon induit la surexpression du gène et conduit à la suraccumulation de pigmentation (Butelli et al. 2012). (b) Le dimorphisme du phalène du bouleau, *Biston betularia*, représente un bel exemple de la théorie de la sélection naturelle et de la réponse évolutive face aux changements environnementaux. Durant la révolution industrielle, la population de la forme « *carbonaria* » a fortement augmenté au détriment de la forme *typica*, population plus importante auparavant. Suite à l'industrialisation, la couleur noire du papillon conférait un avantage sélectif car plus difficilement repérable par les prédateurs sur les bouleaux noircies par la pollution industrielle. La forme *carbonaria* est expliquée par la présence d'un ET (*carbonaria-TE*) dans le premier intron du gène *cortex* induisant une augmentation des transcrits du gène. Ce gène impliqué dans la régulation du cycle cellulaire semble réguler les motifs de pigmentation au cours du développement cellulaire (van't Hof et al., 2016 ; Nadeau et al., 2016). (c) Les ET attirent les marques épigénétiques répressives et jouent ainsi un rôle de régulateur épigénétique. Par exemple, chez le palmier à huile, la culture *in vitro* induit deux phénotypes: fruit fertile et «manteau parthénocarpique». Le rétrotransposon *Karma* est inséré dans l'intron du gène *deficiens* impliqué dans le développement floral. Durant la culture *in vitro*, l'hypométhylation de *Karma* induit l'épissage alternatif du gène responsable du phénotype « manteau parthénocarpique » (Ong-Abdullah et al., 2015).

de traits adaptatifs.

3.2 L'impact des éléments transposables sur le phénotype des organismes hôtes

Après avoir discuté des effets moléculaires et génétiques causés par les insertions d'ET, dans cette partie, nous décrirons l'impact phénotypique d'une insertion d'ET.

Dans la revue présentée dans la partie 2, nous avons discuté du rôle de la méthylation de l'ADN sur les caractères agronomiques. Les ET peuvent également être à l'origine de certains caractères utilisés lors de la domestication. Par exemple la pigmentation des oranges sanguines *Citrus sinensis* (Figure 7a) est due à un gène de type *Myb* R2R3, aussi appelé *Ruby* qui participe à la régulation de la biosynthèse des anthocyanes (Butelli et al. 2012). La couleur rouge de la variété Moro par exemple est due à la présence d'un rétrotransposon de type *Copia* dans la séquence promotrice du gène *Ruby* induisant une surexpression du gène et donc une production importante d'anthocyanes contrairement à une variété orange telle que Navelina par exemple (Figure 7a) (Butelli et al. 2012). Dans cet exemple, l'activité de l'ET a contribué à la création de diversité génétique.

Un des exemples les plus frappants pour illustrer le rôle adaptatif des ET est celui du phalène du bouleau, *Biston betularia*. *B. betularia* est une espèce dimorphique et deux types de pigmentation existent : la forme *typica* (pigmentation blanche et noire) et la forme *carbonaria* (pigmentation noire) (Figure 7b). Au cours de la révolution industrielle en Angleterre, la population *carbonaria* a pris le dessus sur la population *typica* auparavant majoritaire. Durant l'industrialisation, la couleur noire permettait aux phalènes d'être plus difficilement repérables par les prédateurs sur les bouleaux noircis par l'industrialisation, leur conférant ainsi un avantage sélectif. Il est intéressant de constater que le dimorphisme de pigmentation est expliqué par l'insertion d'un ET (*carbonaria*-TE) dans le premier intron du gène *cortex* qui conduit à la surexpression du gène (Van't Hof et al. 2016). Le gène *cortex* impliqué dans le cycle cellulaire semble réguler les motifs de pigmentation au cours du cycle cellulaire (Nadeau et al. 2016). Ici, le polymorphisme lié à l'ET a conféré un avantage sélectif.

Même si de nombreux exemples montrent l'effet bénéfique des ET sur les génomes hôtes, il existe également de nombreux exemples qui témoignent de leurs effets délétères. Chez l'humain par exemple, certains cancers et autres maladies sont très fréquemment associés avec l'activité du LINE L1 (Burns 2017; Scott et Devine 2017). Récemment, chez les plantes, le LINE *Karma* a été impliqué dans la stérilité du palmier à huile (Figure 7c). La culture *in vitro*

induit deux types de phénotypes chez le palmier à huile : le type sauvage fertile et le type « manteau parthénocarpique ». *Karma* est inséré dans un intron du gène *deficiens*, impliqué dans le développement floral. Durant la culture *in vitro*, *Karma* peut être déméthylé et cette hypométhylation induit l'épissage alternatif du gène conduisant à la stérilité du fruit.

Ces trois exemples présentés ci-dessus montrent les impacts positifs ou négatifs très marquants des ET sur la régulation de l'expression des gènes et par conséquent sur le phénotype de l'organisme et soulignent les liens très complexes qui unissent ces ET avec leur génome hôte.

3.3 Bénéfiques ou néfastes ?

Les ET une « arme à double tranchant » ? C'est ce que semble penser Chuong et al (2016a) en concluant leur excellente revue sur le rôle des ET dans l'évolution des réseaux de régulation. Longtemps, les scientifiques se sont demandés pourquoi les ET étaient conservés dans les génomes hôtes et pourquoi ils n'avaient pas été éliminés si leurs effets n'étaient que néfastes. Les différents exemples présentés dans cette partie ont montré que, s'il est vrai que les ET peuvent être néfastes pour l'hôte, il est vrai aussi que les ET constituent un véritable réservoir de séquences régulatrices et participe à l'adaptation des organismes à leur environnement. En réponse à de fortes régulations épigénétiques, les ET évoluent rapidement et ont acquis une grande diversité d'éléments régulateurs qui pourrait expliquer leur capacité à survivre dans les génomes hôtes. La présence de certains ET dans les génomes peut également être expliquée par la « domestication » par les génomes hôtes de certains éléments impliqués dans des innovations biologiques comme c'est le cas pour des rétrovirus endogènes (ERV) impliqués dans le développement du placenta ou la mise en place du système immunitaire chez les mammifères. Chez les primates (dont l'homme), les ERV et notamment la famille de rétrovirus HERV jouent un rôle majeur dans le développement du placenta (Denner 2016). La synthèse de protéines syncytines, protéines de l'enveloppe virale et donc codées par le rétrovirus, induit la formation du tissu appelé syncytiotrophoblaste, nécessaire au bon développement du fœtus. Le processus semble relativement conservé chez les mammifères (Denner 2016). Récemment, l'équipe de Cédric Feschotte a également montré à l'aide de la méthode d'édition CRISPR l'implication des ERV dans la mise en place du système immunitaire chez les mammifères (Chuong et al. 2016b). Les auteurs ont montré par une analyse ChIP-seq que les gènes codant des facteurs de transcription impliqués dans la réponse immunitaire étaient fortement associés à des séquences

d'ERV. L'élimination spécifique de ces ERV a révélé l'implication directe de ces séquences dans la transcription des gènes interférons. La dispersion de ces séquences dans le génome des mammifères contribue à l'évolution de leur système immunitaire.

Les progrès techniques réorientent sans cesse les idées scientifiques et malgré des débuts difficiles dans la communauté scientifique, les ET sont désormais aujourd'hui reconnus pour leur fort potentiel adaptatif et pour leur participation dans l'évolution des organismes hôtes. Au même titre qu'une mutation stochastique apparue dans le génome, les insertions d'ET peuvent être à la fois bénéfiques, néfastes ou simplement neutres pour l'organisme hôte. Au regard des rapides avancées techniques, les futurs travaux ne cesseront d'affiner et de préciser les étroites relations entre les ET et leurs hôtes.

4. Méthodes de détection des éléments transposables actifs

Comme précédemment mentionné, les ET sont strictement régulés et peu d'éléments sont aujourd'hui connus pour être actifs, et il est difficile d'évaluer le taux de transposition dans un organisme. Finalement y a-t-il peu d'exemples connus car il y a peu d'éléments qui transposent, ou est-ce les techniques aujourd'hui disponibles qui ne nous permettent pas de réellement évaluer l'activité des ET au sein des génomes ? Malgré l'évolution des techniques moléculaires et des outils bioinformatiques, suivre l'activité des ET en temps réel reste un challenge. En effet, de par leur nature répétée et la complexité des génomes, l'identification d'ET actifs est encore très limitée. Différentes stratégies moléculaires et bioinformatiques ont été développées pour étudier la transposition en ciblant différentes étapes du cycle de vie des ET. Ces méthodes sont détaillées ci-dessous.

4.1 Techniques moléculaires de détection des éléments transposables

Les propriétés mutagènes des ET permettent de les identifier parfois de façon inattendue par le biais du phénotype que l'ET induit par son insertion. Par exemple un défaut dans le nombre

d'organes floraux et dans l'architecture des rachis est dû à l'insertion du transposon à ADN *hAT* dans le gène *FONI* (*Floral Organ Number 1*) chez le riz (Moon *et al.* 2006). Une anomalie florale dans un mutant de riz a également permis l'identification d'une néo-insertion du rétrotransposon à LTR *Houba* inséré dans le gène *FRIZZY PANICLE* (Komatsu *et al.* 2003).

Au-delà des détections indirectes par des études génétiques de variants phénotypiques, deux stratégies majeures ont été développées pour étudier les mouvements d'ET par des approches moléculaires : (1) l'étude des polymorphismes d'insertions et (2) les études transcriptomiques.

Par recherche de polymorphisme d'insertion, il est possible de détecter la transposition d'un ET candidat à l'aide de la technique du *transposon display*, une des méthodes les plus utilisées dans le domaine. Cette méthode est dérivée de la technique de détection des marqueurs AFLP (*Amplified Fragment Length Polymorphism*) et permet d'isoler et de détecter simultanément toutes les insertions d'un élément (Van den Broeck *et al.* 1998). L'ADN est digéré par deux enzymes de restriction : une enzyme qui coupe à l'intérieur de l'ET et une enzyme qui ne coupe pas à l'intérieur de l'ET. Deux adaptateurs spécifiques des sites de restriction sont liés aux fragments digérés. Les fragments correspondants aux insertions de l'élément étudié sont ensuite sélectionnés et amplifiés par PCR à l'aide d'une amorce spécifique de l'ET marquée radioactivement et une autre spécifique du site de restriction qui ne coupe pas à l'intérieur de l'élément. Les fragments sont ensuite déposés sur un gel polyacrylamide de haute résolution. Pour chacune des insertions, la distance entre l'ET et un site de restriction est statistiquement différente et donc les fragments de chaque insertion ont une taille différente. En comparant plusieurs individus ou variétés, il est possible de visualiser sur gel les polymorphismes d'insertion de l'ET étudié. C'est notamment par cette approche que l'activité des transposons à ADN *mPing* et *Pong* a été détectée (Jiang *et al.* 2003). La principale limite de cette méthode repose sur la nécessité d'avoir au préalable un élément candidat. Cette technique peut être toutefois intéressante pour valider l'activité d'ET détectée par une technique génomique (par exemple par reséquençage du génome, voir au-dessous).

Une autre stratégie consiste à étudier l'activité transcriptionnelle des ET par des sondes spécifiques de chaque famille d'ET. À titre d'exemple, *Tos17* a été identifié grâce au crible d'ADNc de cals de riz par des amorces « universelles » ciblant des domaines conservés de la RT (Hirochika *et al.* 1996). À plus grande échelle, Picault *et al.* (2009) ont mis en place une puce « transposome » regroupant des sondes spécifiques des familles d'ET de riz afin d'étudier la transcription des ET. L'étude a permis l'identification de *Lullaby*, cousin de l'élément *Tos17*

et qui génère aussi des néo-insertions dans les cals (Picault et *al.* 2009). Ces approches transcriptomiques nécessitent ensuite une validation par *transposon display* ou Southern blot afin de valider la mobilité des familles d'ET candidats.

L'avancée des techniques de séquençage a en outre permis le développement d'analyses à l'échelle du génome.

4.2 Techniques de détection des éléments transposables par séquençage

Depuis la méthode de Sanger, les technologies de séquençage se sont multipliées et différentes générations de séquençage dit « à haut débit » ou NGS (*Next Generation Sequencing*) se sont développées. Aujourd'hui, on distingue le séquençage de 2^{ème} et de 3^{ème} génération. Le séquençage de 2^{ème} génération implique une étape d'amplification de l'ADN alors que le séquençage de 3^{ème} génération permet le séquençage d'une seule molécule d'ADN (Figure 8), comme illustré plus loin dans cette partie.

4.2.1 Analyses transcriptionnelles par RNA-seq

L'analyse des transcrits par séquençage des ARN (RNA-seq) permet d'identifier les ET transcriptionnellement actifs sans la nécessité d'utiliser des sondes. Comme toute analyse transcriptionnelle elle ne prend pas en compte la capacité de ces éléments à produire des protéines d'une part et d'une autre part à produire des protéines fonctionnelles. En effet la transcription des ET est la première étape du cycle de vie des ET. Néanmoins, il existe des mécanismes de *silencing* post-transcriptionnels (voir Introduction page 13) qui ciblent et inactivent les transcrits des ET. De même, les ET accumulent des mutations qui peuvent conduire à la synthèse de protéines non fonctionnelles ou l'absence de traduction. L'étude des ET par RNA-seq ne permet donc pas d'avoir une image précise des éléments qui sont transpositionnellement actifs. De nombreuses stratégies ont ainsi été développées pour avoir accès directement aux néo-insertions par séquençage d'ADN.

4.2.2 Capture d'ADN de rétrotransposon (RC-seq, ATLAS-seq)

Chez l'humain, il existe trois familles de rétrotransposons connues pour être fortement actives : *L1*, *Alu* et *SVA* (Mills et *al.* 2007). Afin d'étudier les insertions somatiques chez l'homme, Baillie et *al.* (2011) ont développé une stratégie appelée *retrotransposon capture sequencing* ou

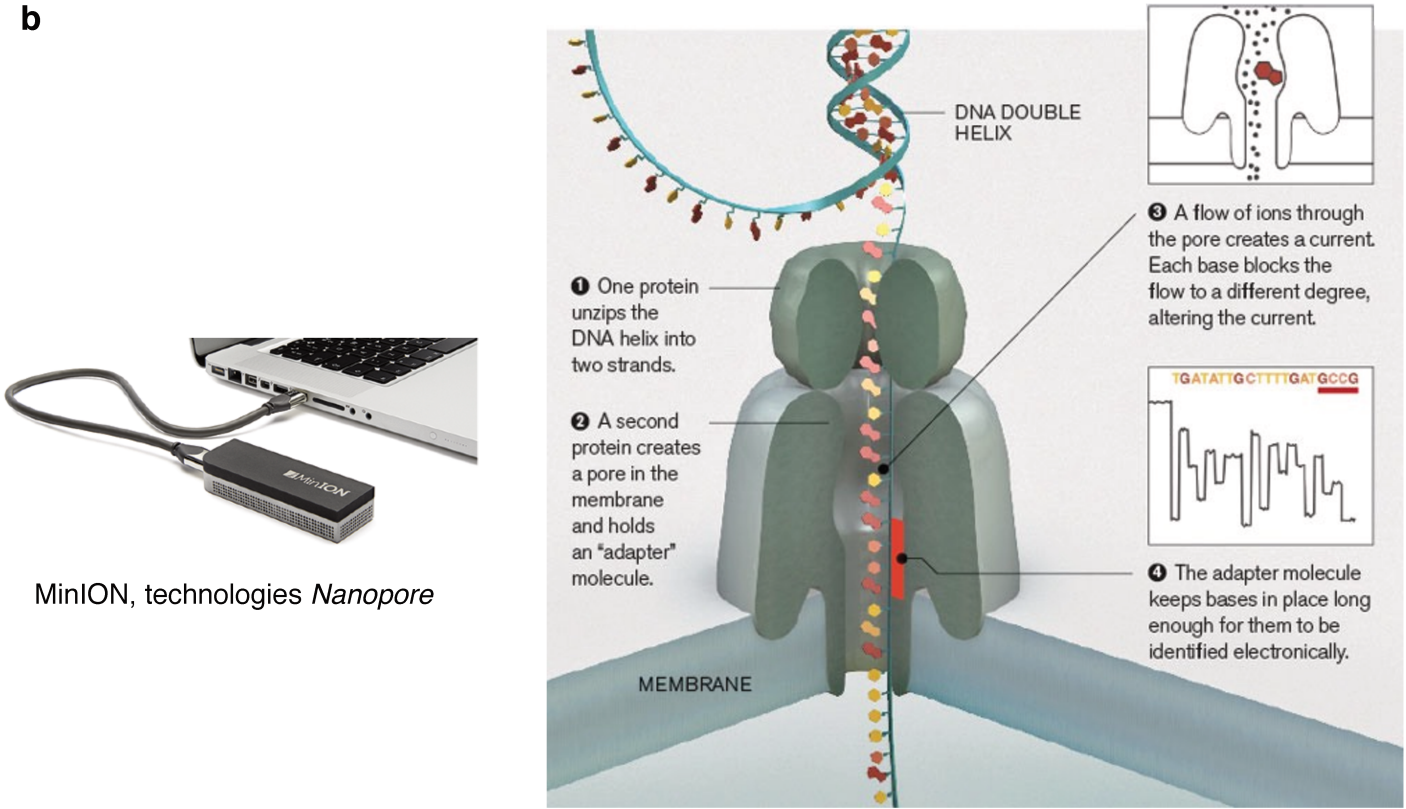
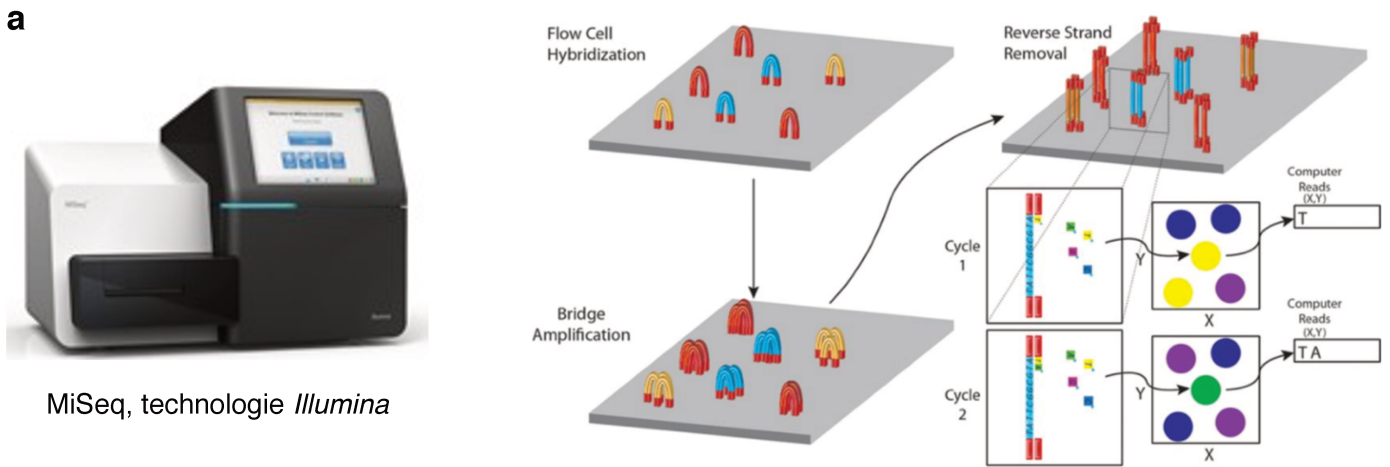


Figure 8. Principe du séquençage de 2ème et 3ème génération (a) Dans le séquençage Illumina la banque s'hybride sur la *flow cell* et est amplifiée par pont et des clusters d'ADN sont formés. Le séquençage se déroule en cycle de 3 étapes: incorporation des nucléotides, détection et photoclivage. Pour chacun des cycles, 4 nucléotides bloqués sont envoyés, l'incorporation d'une base provoque l'arrêt de l'élongation et le nucléotide est débloqué par photoclivage. Chaque nucléotide incorporé est identifié par imagerie. Le cycle est répété un grand nombre de fois afin d'étendre la lecture de séquençage (Image: Churko et al., 2013). (b) Dans le séquençage Nanopore la molécule d'ADN passe par un pore dans une membrane et est bloquée à l'aide d'un adaptateur. Un flux d'ions circule dans le pore et crée un courant. Chaque groupe de bases qui constitue la molécule d'ADN bloque le flux à différentes intensités altérant le courant et est identifié électroniquement (Image Nanopore).

RC-seq qui consiste à capturer tous les fragments d'ADN qui contiennent la séquence d'une de ces trois familles à l'aide d'une puce avec des amorces spécifiques à ces trois éléments (Baillie et *al.* 2011). Les fragments capturés sont ensuite séquencés par NGS. Seules les insertions natives (déjà présentes dans le génome de référence) et les insertions somatiques sont séquencées et il est donc facile d'identifier à partir de ces données, les néo-insertions de ces trois rétrotransposons. Cette stratégie permet d'avoir une couverture de séquençage très importante et limite donc la présence de faux-positifs. Les faux-positifs sont des régions du génome pour lesquelles des néo-insertions sont détectées mais qui se révèlent être des artefacts de séquençage. Plus récemment Philippe et *al.* (2016) ont développé une technique d'amplification PCR ciblée des insertions de rétrotransposons L1 couplée à du séquençage NGS et à une analyse bioinformatique fine. Les régions flanquantes 5' et 3' sont analysées séparément permettant de détecter les insertions tronquées, fréquentes chez les *LINE*. Ces techniques, par leurs grandes couvertures et leurs précisions, permettent de détecter des néo-insertions même somatiques chez l'humain où seules 3 familles sont actives. En revanche ces techniques restent limitée à l'étude d'un ou de quelques ET candidats et ne permettent pas une analyse sans *a priori* de tous les ET d'un génome, en particulier végétal.

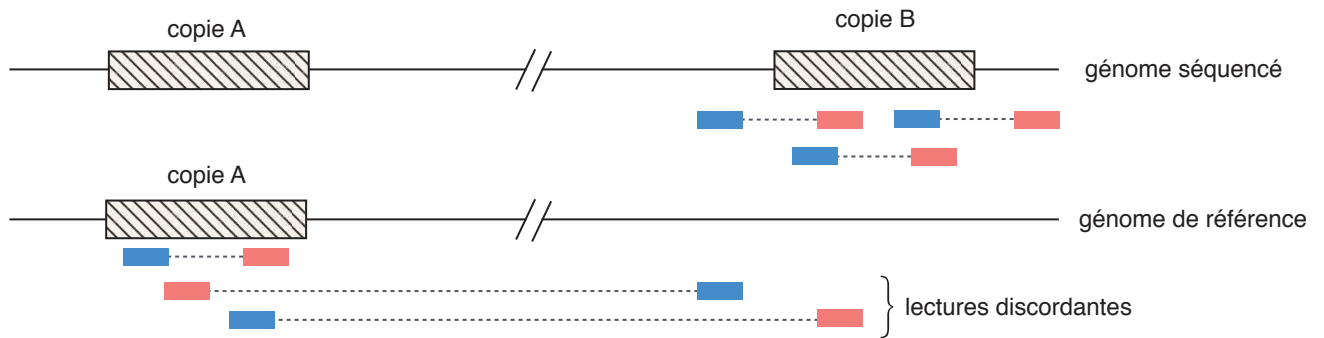
4.2.3 Reséquençage de génome

Le reséquençage des génomes est aujourd'hui la méthode la plus employée pour identifier de nouvelles copies d'ET dans les espèces possédant un génome de référence, et de nombreux outils informatiques ont été développés pour cette détection (El Baidouri et Panaud 2012; Ewing 2015).

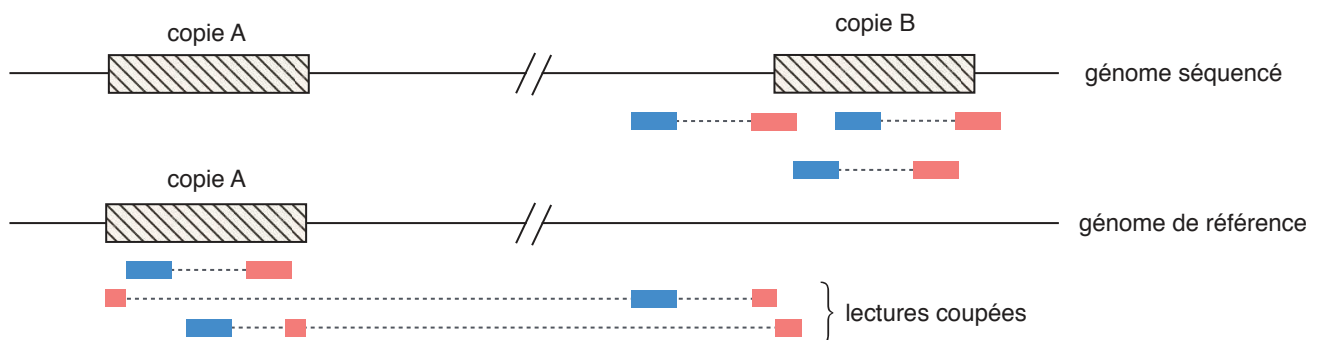
4.2.3.1 Séquençage de 2^{ème} génération et détection des néo-insertions

Le séquençage par la technologie Illumina est actuellement le plus utilisé car il permet un multiplexage (plusieurs banques d'ADN séquencées en même temps) avec des lectures (≤ 300 nt) pairées (séquençage des deux extrémités du fragment d'ADN) et une grande profondeur de séquençage (Figure 8a). L'ADN à séquencer est d'abord fragmenté et des adaptateurs sont liés à chaque extrémité. Les banques d'ADN sont ensuite hybridées sur une puce ou *flow cell* et les fragments sont amplifiés par des ponts à l'aide d'amorces complémentaires aux adaptateurs. Le séquençage est réalisé par cycles de trois étapes : incorporation d'un nucléotide, détection et photoclivage. À chaque cycle, 4 nucléotides (A, T, C, G) bloqués et marqués sont envoyés et l'incorporation d'un des nucléotides, provoque l'arrêt

a Détection des paires de lectures discordantes



b Détection des paires « coupées » (*split reads*)



c Analyse de la profondeur de séquençage

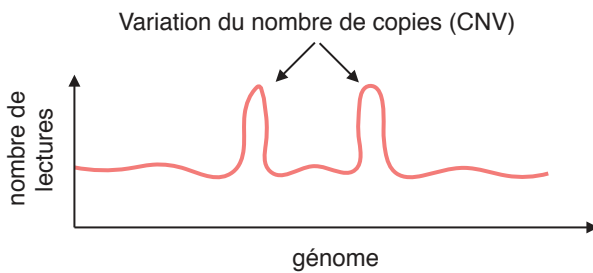


Figure 9. Détection des néo-insertions par reséquençage Illumina. Il existe 3 méthodes pour détecter des néo-insertions à partir de données de séquençage *Illumina*. **(a)** La première méthode consiste à détecter les paires de lectures dites « discordantes ». La copie A d'un ET est la copie native alors que la copie B représente une néo-insertion de l'ET, les paires de lectures sont ici représentées en bleu et rouge. Une paire de lectures est dite discordante lorsque l'écart entre l'alignement des deux lectures d'une même paire est supérieur à l'insert initial. **(b)** La seconde méthode consiste à détecter les lectures « coupées » autrement dit une lecture qui s'aligne à deux endroits différents du génome. **(c)** La troisième méthode consiste à analyser la couverture de séquençage, autrement dit le nombre de lectures alignées pour chaque base du génome. Les ET activés qui se sont donc multipliés par rapport au génome de référence, ont une couverture supérieure au reste du génome et ont donc un nombre de copies qui varie par rapport au génome de référence (*copy number variation*, CNV).

de l'élongation. Le nucléotide incorporé est débloqué par laser et identifié par imagerie. La dernière étape du cycle consiste au clivage du marqueur à l'aide d'un rayon ultraviolet et un nouveau cycle redémarre. Le cycle est répété un nombre de fois qui correspond à la taille de la lecture : pour des lectures de 250nt par exemple, le cycle de séquençage est répété 250 fois.

Différentes analyses bioinformatiques de ces données existent pour détecter des néo-insertions d'ET (Figure 9). La première stratégie consiste à détecter les paires de lectures dites « discordantes » suite à l'alignement des lectures pairées contre le génome de référence (Ewing 2015). Par exemple, la copie A d'un ET est activée dans un génome et une néo-insertion est créée (copie B). Lorsque les paires de lectures qui correspondent à la copie B et aux régions flanquantes de cette néo-insertion sont alignées sur le génome de référence, une première lecture s'aligne sur la copie A de l'ET activé (pas de copie B dans le génome de référence) et la seconde sur la région d'insertion de la copie B. Ainsi, l'écart entre l'alignement des deux lectures d'une même paire est supérieur à l'insert initial et la paire est dite discordante. Cette méthode permet à la fois d'identifier l'élément actif et de détecter la région d'insertion de la néo-copie.

La seconde méthode consiste à détecter les lectures « coupées ». Autrement dit, les lectures qui couvrent une des extrémités de la copie B et une des régions flanquantes de la néo-insertion s'alignent sur le génome de référence en deux parties : une partie sur la copie A et la seconde au point d'ancrage de la copie B. La détection des lectures coupées permet de définir la position précise de la néo-insertion. Cependant ces lectures sont peu nombreuses et leur identification requiert une profondeur de séquençage assez importante.

La troisième méthode consiste à analyser la couverture de séquençage, autrement dit le nombre de lectures alignées pour chaque base du génome. Les ET activés, qui se sont donc multipliés par rapport au génome de référence, ont une couverture supérieure au reste du génome et ainsi, cette analyse permet d'identifier la ou les familles d'ET qui se sont activées et dont les néo-insertions sont difficilement détectables par les deux méthodes précédentes (par exemple les néo-insertions dans les régions répétées). Néanmoins, cette analyse est dédiée uniquement à l'identification de rétrotransposons qui multiplient leurs nombres de copies lorsqu'ils sont actifs contrairement aux transposons à ADN qui ne peuvent donc pas être détectés par cette approche.

Le reséquençage de génome comporte certaines limites. Premièrement, les néo-insertions des ET dans les tissus somatiques ne sont pas héritées à la génération suivante et sont donc

difficilement détectées par reséquençage. De plus, les logiciels d'analyse disponibles ne sont pas capables de détecter des néo-insertions localisées dans des régions répétées du génome ce qui représente une part du génome non négligeable. En effet, la longueur des lectures (maximum 300 nt) ne permet pas d'aligner de façon unique une lecture répétée dans le génome et de nombreuses ambiguïtés demeurent sur la localisation de la néo-insertion. Et pour finir, de nombreux faux-positifs sont détectés par ces approches, surestimant de façon significative le taux de transposition (Ewing 2015). Finalement, ce type d'approche est donc plus adapté à des génomes de petite taille et n'est possible que sur des organismes pour lesquels nous possédons un génome de référence avec une base de données d'ET de très bonne qualité.

4.2.3.2 Séquençage de 3^{ème} génération et détection des néo-insertions

La nouvelle et 3^{ème} génération de séquençage haut débit est en plein essor. Différentes compagnies telles que Pacific Biosciences et Oxford Nanopore ont développé des techniques de séquençage qui permettent d'obtenir des lectures de plusieurs kilobases (kb) (Figure 8). L'équipe dans laquelle j'ai réalisé ma thèse a mis au point l'utilisation de la technique de séquençage de Nanopore au sein du laboratoire dans le but d'identifier les variants structuraux dont les mouvements d'ET.

Le séquençage Nanopore consiste en une puce ou *flow cell* composée de « nanopores » au travers desquels un flux d'ions circule (Figure 8). Le brin de la molécule d'ADN (charge négative) est bloqué dans le pore (charge positive) à l'aide d'un adaptateur et chaque groupe de nucléotide passé dans le pore bloque le flux à différentes intensités. L'intensité du changement permet l'identification des nucléotides.

La présence de longues lectures de séquençage qui couvrent entièrement une néo-insertion (ET et régions flanquantes) révolutionne la détection d'ET actifs et de variants structuraux. Cette technologie permet de détecter des insertions somatiques et des insertions présentes dans des régions répétées contrairement au séquençage Illumina. L'utilisation de la technologie Nanopore a par exemple permis de mettre en évidence des néo-insertions d'un rétrotransposon à LTR, *Évadé*, dans une plante d'*Arabidopsis thaliana* à l'épigénome déstabilisé (Figure 10) (Debladis et al. 2017). Néanmoins, cette technologie en plein essor n'est commercialisée que depuis peu de temps (environ 5 ans) et de nombreuses mises au point sont encore en cours de développement. Aujourd'hui, cette méthode n'est pas encore adaptée à un grand nombre d'organismes ou de tissus et a un coût important. Néanmoins, dans les années qui arrivent, les

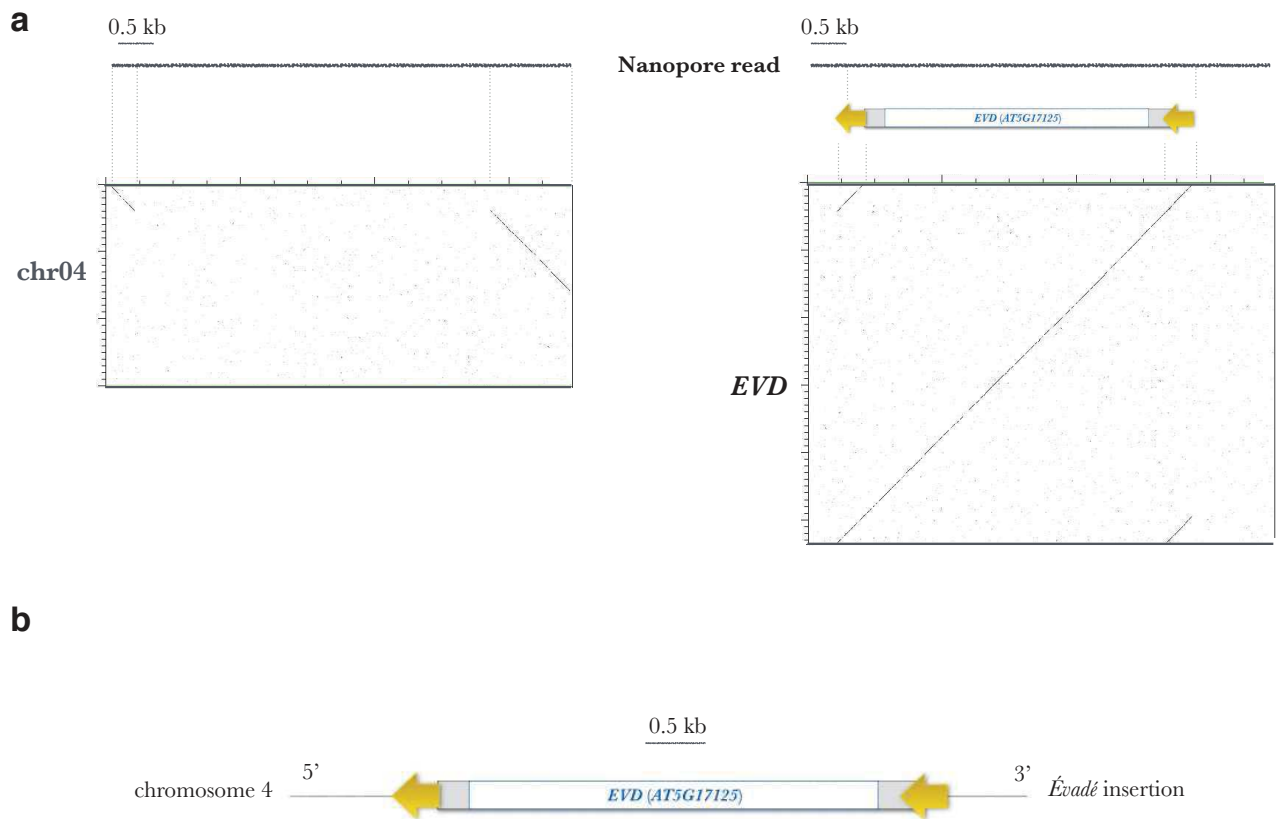


Figure 10. Détection des néo-insertions par reséquençage Nanopore. Le séquençage avec la technologie Nanopore permet d'obtenir des longues lectures, supérieures à 5Kb. Pour la détection des néo-insertions, cela permet d'obtenir une lecture qui couvre entièrement la néo-copie et sa région d'insertion. Par ce type de séquençage, Debladis, Llauro et al. (2017) ont détecté une néo-insertion du RT *EVD* dans l'épiRIL12 d'*Arabidopsis thaliana*. Une lecture Nanopore de 7Kb est alignée contre la région d'insertion sur le chromosome 4 (**a**, gauche) et contre la séquence d'*EVD* (**a**, droite) par dot-plot. (**b**) L'insertion d'*EVD* est représentée avec l'orientation des LTR indiquée par les flèches jaunes. (Figure Debladis, Llauro et al., 2017)

possibilités qu'offriront un séquençage Nanopore risquent de révolutionner la génomique. Par exemple, très récemment, Simpson et *al.* (2017) ont mis au point une analyse pouvant permettre d'évaluer le niveau de méthylation par séquençage Nanopore.

Finale­ment, aucune méthode ne semble aujourd'hui capable d'évaluer le taux de transposition dans un grand nombre de tissus, de conditions et d'organismes. Dans le but de développer une technique alternative, à faible coût et adaptable à un grand nombre d'espèces, nous avons développé durant mes travaux de thèse une analyse par séquençage des formes d'ADNecc, témoins de l'activité des ET. C'est une stratégie de détection certes indirecte mais qui promettait d'offrir l'avantage d'être un meilleur marqueur de transposition que l'activité transcriptionnelle.

Le prochain chapitre consiste à une synthèse bibliographique des connaissances actuelles sur les ADNecc dans les génomes eucaryotes.

CHAPITRE 2

- ÉTUDE DES ADN CIRCULAIRES EXTRACHROMOSOMIQUES -

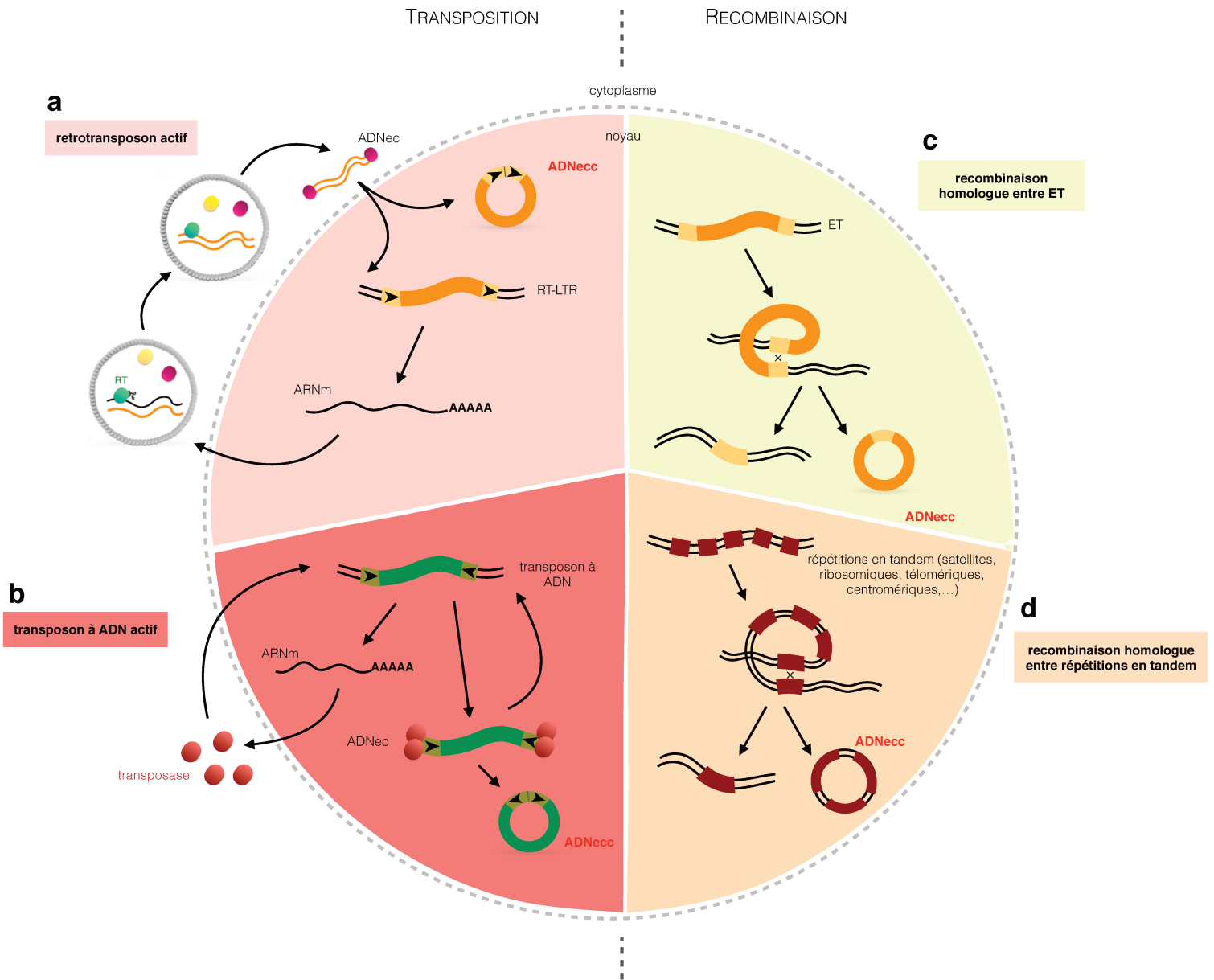


Figure 11. Mécanismes de formation des ADNecc qui constituent le mobilome d'une cellule végétale.

Chez les plantes, il existe deux types d'ADNecc: **(a,b,c)** les cercles formés par à partir d'ET ou de génomes viraux et bactériens (par exemple les plasmides mitochondriaux), **(d)** les cercles dérivés des répétitions en tandem (satellite, télomères, centromères et ADN ribosomiques). La formation de ces cercles peut être expliquée par l'efficacité du mécanisme de réparation de l'ADN qui reconnaît l'ADNec linéaire de la nouvelle copie du rétrotransposon **(a)** ou la copie excisée d'un transposon à ADN **(b)** et qui circularise la molécule par le mécanisme du NHEJ ou par HR pour former un ADNecc. Les cercles de 1 LTR ou 1 TIR peuvent être également formés à partir d'éléments non actifs par HR entre leurs répétitions terminales (LTR ou TIR) aboutissant à la formation de *solo LTR* ou *solo TIR* dans le génome **(c)**. Les répétitions en tandem telles que les séquences ribosomiques, télomériques, centromériques peuvent également générer des ADNecc par HR entre les répétitions **(d)**. Légende voir Figure 3. Les triangles noirs symbolisent les répétitions des rétrotransposons (LTR) et transposons à ADN (TIR).

Dans les cellules eucaryotes, une part de l'ADN n'est pas associée avec les chromosomes nucléaires ou les génomes mitochondriaux et chloroplastiques. Cet ADN, appelé ADN extrachromosomique (ADNec) est très fréquemment présent sous forme d'ADN extrachromosomique circulaire (ADNecc). Les ADNecc ont été détectés il y a plusieurs décennies (Gaubatz 1990) et semblent présents dans tous les organismes eucaryotes aujourd'hui testés (Cohen et Segal 2009). Longtemps considérés comme des produits secondaires sans intérêt biologique majeur, peu d'études ont été réalisées sur le compartiment du génome qu'ils composent. Ce n'est que très récemment que l'intérêt pour les ADNecc a émergé. L'ADNecc semble en effet participer à la plasticité du génome, notamment au niveau de la dynamique des régions répétées et de l'amplification de gènes (Cohen et Segal 2009). Dans ce chapitre, nous définirons les principaux types d'ADNecc et leurs origines puis nous verrons également quelques exemples très récents chez les cellules animales qui illustrent le potentiel et le rôle insoupçonné de ce compartiment génomique. Finalement, nous terminerons en présentant les principales méthodes disponibles aujourd'hui pour étudier et identifier les ADNecc.

1. Définition du mobilome et caractérisation des ADN extrachromosomiques circulaires

1.1 Définition et origine du mobilome

Le mobilome a été défini comme la fraction du génome qui correspond à des plasmides chez les bactéries ou à des ET et des séquences répétées présentes sous forme extrachromosomique dans les cellules eucaryotes (Figure 11) (Siefert 2009). Chez les plantes trois types principaux d'ADNecc sont documentés: (i) les cercles qui proviennent des séquences répétées en tandem (satellites, télomériques, centromériques et ribosomales) et (ii) les cercles qui proviennent des ET et (iii) les cercles formés par les génomes viraux et bactériens (tels que les plasmides mitochondriaux) (Figure 11) (Cohen et Segal 2009).

La formation de ces trois types de cercles est indépendante de la réplication de l'ADN (Cohen et Mechali 2002). De précédentes études suggèrent que les cercles provenant de répétitions en tandem sont générés par recombinaison homologue (Cohen et Mechali 2001) ou par le mécanisme de réparation de l'ADN (Cohen et *al.* 2006). Ces ADNecc semblent assurer une certaine plasticité génomique (Cohen et Segal 2009). Les ADNecc provenant d'ET ont été décrits depuis plusieurs décennies chez les animaux et chez les plantes : un élément de type *Copia* chez la drosophile a été décrit sous forme de cercles (Flavell et Ish-Horowicz 1981) et

chez le tabac des ADNec dérivés du rétrotransposon à LTR *Tto1* ont été observés il y a déjà plus de 20 ans (Hirochika et Otsuki 1995). Ces cercles d'ET peuvent être formés à partir de l'ADN extrachromosomique linéaire par le mécanisme de NHEJ participant au processus de réparation de l'ADN (Figure 11, voir Introduction page 9) (Li et al. 2001; Kilzer et al. 2003) ou par recombinaison homologue entre les répétitions situées à leurs extrémités. La différence entre des ADNec circularisés par NHEJ et ceux circularisés par recombinaison homologue est observée au niveau de la jonction du cercle (Figure 11). Les cercles issus de la voie NHEJ sont joints soit par un ou deux LTR ou soit par un ou deux TIR contrairement aux cercles issus de la recombinaison homologue qui sont strictement composés d'un seul LTR ou d'un seul TIR (Bennetzen et Kellogg 1997).

De nombreuses questions restent en suspens sur les ADNec dans les génomes. Aujourd'hui nous ne savons pas si ces ADNec peuvent tous se réinsérer dans le génome, si ces ADNec peuvent être transcrits ou si ces cercles présentent des marques épigénétiques. Peu d'indices permettent de répondre à ces questions et les avis sont partagés sur le devenir de ces formes circulaires. Certains auteurs suggèrent que les ET sous forme d'ADNec sont des formes transitoires avant leur élimination et participent indirectement au contrôle de leur prolifération (Bennetzen et Kellogg 1997; Cohen et al. 2008). D'autres auteurs en revanche semblent penser que ces formes d'ADN pourraient être des intermédiaires de transposition et se réintégrer dans le génome (Mourier 2016). Il semblerait que de futures études approfondies sur ces ADNec pourraient apporter de nouvelles connaissances sur les ET et sur leurs mécanismes.

1.2 Rôles émergents des ADNec dans les cellules eucaryotes

1.2.1 L'ADNec comme biomarqueur de cancers ?

L'étude du mobilome suscite un intérêt croissant pour la compréhension des mécanismes en jeu dans les cellules cancéreuses. Dans les années 90, Cohen et Lavi (1996) ont montré que le quantité d'ADNec (alors appelé *small polydisperse circular DNA* ou *spcDNA*) par cellule augmentait à la suite d'un traitement carcinogène dans des cellules infectées par le virus SV40 (Cohen et Lavi 1996). Ces cercles contenaient des fragments du virus mais également des répétitions inversées du génome de l'hôte. Les auteurs postulèrent alors que le niveau d'ADNec dans un type cellulaire pourrait refléter son état de stabilité génomique (Cohen et al. 1997). La formation d'ADNec pourrait ainsi être impliquée dans l'amplification génique.

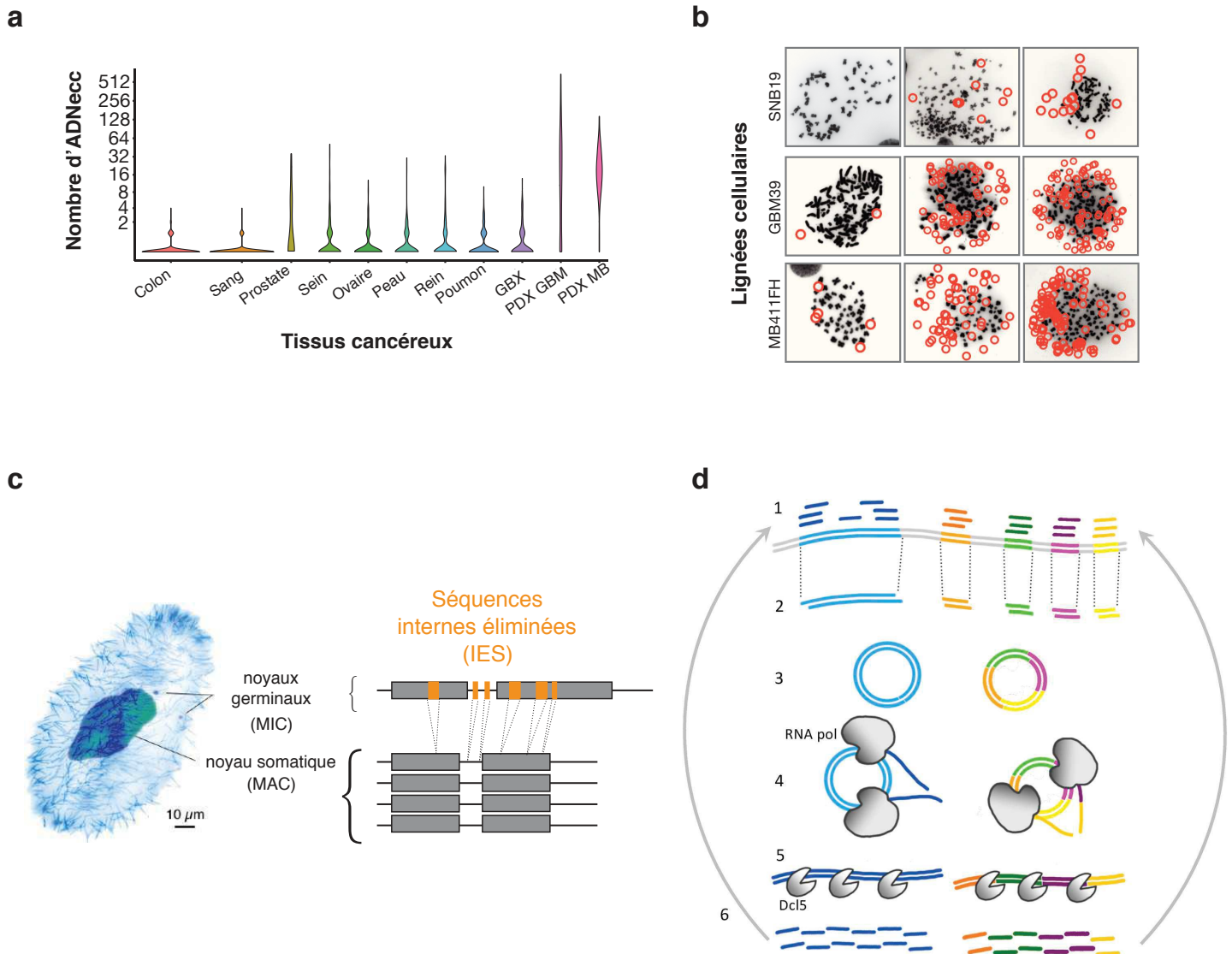


Figure 12. Diversité des ADNecc dans les cellules eucaryotes. (a) Distribution des ADNecc par cellule en métastase en fonction du tissu cancéreux (Turner et al. 2017). GBM : glioblastome ; MB : medulloblastome ; GBX : glioblastome dérivé de xéno greffe greffe d'un organe/tissu/cellule d'une espèce à une espèce différente) ; PDX : patient dérivé de xéno greffe. (b) Détection des ADNecc dans trois lignées cellulaires différentes. Trois cellules par lignées sont montrées. L'ADN est coloré au DAPI et le logiciel *ECdetect* identifie (cercles rouges) et compte le nombre d'ADNecc par cellule (Turner et al. 2017). (c) La paramécie (*Paramecium tetraurelia*) est un organisme unicellulaire avec deux types de noyaux: deux noyaux germinaux (micronucléus MIC) et un noyau somatique (macronucléus). L'ADN est amplifié (environ 800 fois; Duret et al. 2008) et constitue le noyau somatique. Néanmoins, le génome des noyaux germinaux comportent des séquences répétées contrairement au génome du noyau somatique où elles sont éliminées (IES). (d) Modèle de la production d'ARNies à partir des IES (Allen et al. 2017). Les petits ARN produits dans le noyau germinal induisent l'élimination des IES (1). Les IES sont concaténées aléatoirement en ADNecc par la ligase IV (2) et (3). Les ADNecc sont transcrits de façon bidirectionnelle par une ARN polymérase et des ARN double brins sont formés (4). Les ARN double brins sont clivés par Dicer-like 5 (DCL5) et des petits ARN appelés ARNies sont produits (5). Ces ARNies renforcent le contrôle et l'élimination des IES dans le noyau somatique (6).

Par exemple, chez la levure, Moller *et al.* (2015b) ont identifié et séquencé des ADN_{ecc} provenant de gènes dont notamment *HXT6* et *HXT7*, très bien caractérisés dans la littérature et connus pour être amplifiés lorsque les levures sont dans un milieu pauvre en glucose. La forme extrachromosomique de ces deux gènes pourraient alors participer à la stabilité de l'amplification génique. Dans les cellules cancéreuses, les ADN_{ecc} semblent être impliqués dans l'amplification d'oncogènes (Kuttler et Mai 2007). Brièvement, un oncogène est un gène qui à la suite d'une mutation ou d'une modification de son expression induit la formation de cancers. L'amplification d'oncogènes est observée sous deux formes, la première est intrachromosomique et ségrège normalement lors des mitoses alors que la seconde est extrachromosomique et ségrège de façon inégale entre les cellules filles. Deux classes d'ADN_{ecc} ont été détectées dans les tissus cancéreux : (i) des microADN (entre 100 et 400 paires de bases) que l'on retrouve également dans des tissus sains (Shibata *et al.* 2012) et qui proviennent de la voie de réparation de l'ADN (Dillon *et al.* 2015), et (ii) des chromosome dits « double-minute » qui se présentent sous forme de chromosomes circulaires super-enroulés, sans centromère et qui peuvent se répliquer de façon autonome (Hahn 1993).

La présence d'ADN_{ecc} dans les cellules cancéreuses est donc connue depuis des années (Kuttler et Mai 2007) mais c'est seulement très récemment que Turner *et al.* (2017) ont étudié le possible rôle de ces ADN_{ecc} dans l'évolution d'une tumeur. L'équipe de chercheurs a étudié par reséquençage de génome entier 17 types de cancer et a observé des ADN_{ecc} dans 40% des lignées cellulaires étudiées et dans 90% des cancers du cerveau issus de patients. Parmi ces ADN_{ecc} 30% correspondaient à des chromosomes double-minutes. De plus, le nombre d'ADN_{ecc} varie grandement en fonction du type de cancer étudié (Figure 12a). Ces ADN_{ecc} contiennent des oncogènes et certains semblent capables de se réintégrer dans le génome. De façon intéressante les auteurs ont utilisé un modèle mathématique pour prédire l'impact de ces ADN_{ecc} sur l'évolution de la tumeur : la ségrégation inégale des ADN_{ecc} conduit à une distribution très hétérogène (Figure 12b) créant des sous-populations de cellules ce qui favoriserait la résistance aux drogues utilisées pour traiter les cancers et participerait ainsi à l'évolution de la tumeur. Kumar *et al.* (2017) ont également observé la circulation d'ADN_{ecc} et plus précisément de microADN circulaires (Shibata *et al.* 2012) dans les tissus sanguins (plasma et sérum) de patients atteints de cancers. Ces ADN_{ecc} semblent provenir des cellules tumorales et les auteurs suggèrent que si ces ADN_{ecc} ont la capacité d'être transcrits, ces ADN_{ecc} pourraient alors jouer un rôle dans la communication intracellulaire longue distance.

L'ADN_{ecc} : un nouveau biomarqueur pour le diagnostic des cancers ? C'est tout du moins ce

que suggèrent Kumar et *al* (2017) par la facilité d'isoler les ADNecc à partir d'échantillons sanguins. Au-delà des perspectives de biomarqueurs pour diagnostiquer les cancers, étudier les populations d'ADNecc provenant de cellules cancéreuses permet l'accès à une source non négligeable d'information, d'une part au niveau des acteurs de la maladie (oncogènes responsables) et d'autre part au niveau de la ou des mutations possiblement responsables du développement de la tumeur. Ces informations pourront dans l'avenir aider à la mise en place de protocoles thérapeutiques.

1.2.2 Les ADNecc participent à la production des petits ARN chez la paramécie

Dans un tout autre domaine, l'ADNecc peut jouer un rôle central dans la répression des ET/séquences répétées. La paramécie *Paramecium tetraurelia* est un organisme unicellulaire recouvert de cils vibratiles et qui présente un dimorphisme nucléaire : deux noyaux germinaux (micronucléus) et un noyau somatique (macronucléus) (Figure 12c). Le noyau germinifère assure la transmission de l'information génétique lors de la reproduction sexuée et est transcriptionnellement inactif durant le développement et la croissance de l'organisme contrairement au noyau somatique qui assure l'expression de l'information génétique au cours du développement (Jahn et Klobutcher 2002). Durant la reproduction sexuée, le noyau somatique est détruit et un nouveau noyau somatique est formé à partir du noyau germinifère (Jahn et Klobutcher 2002). Curieusement, le génome contenu dans le noyau somatique ne comporte ni séquence répétée, ni ET contrairement au génome contenu dans le noyau germinifère. Le développement du nouveau noyau somatique nécessite l'élimination précise de séquences d'ADN, correspondant à des séquences répétées dont des ET. Ces séquences éliminées sont aussi appelées IES (*Internal Eliminated Sequences*).

L'élimination de ces IES est guidée par des petits ARN générés par le noyau germinifère et l'excision est assurée par *PiggyMac*, une transposase dérivée d'un ET de la famille des *piggyBac* et qui a été domestiquée par le génome hôte (Baudry et *al.* 2009). Une seconde classe de petits ARN spécifiques des IES, les ARNies, est produite dans le noyau somatique après l'excision des IES. Une question restait sans réponse : comment expliquer la production de petits ARN spécifiques de séquences éliminées du génome ? Très récemment, Allen et *al.* (2017) (Allen et *al.* 2017) ont montré que les séquences des IES éliminées du génome somatique étaient concaténées et circularisées par la ligase IV (impliquée dans la réparation de l'ADN) formant ainsi une population hétérogène d'ADNecc (Figure 12d). Ils ont également montré que ces ADNecc étaient transcrits de façon bidirectionnelle dans le noyau somatique

induisant la production de longs ARN double brin clivés par l'enzyme *Dicer-like 5* afin de produire les ARNies responsables de la répression des ET (Allen et *al.* 2017). Ces nouveaux travaux marquent un tournant dans l'étude des ADNec, d'une part parce que c'est une des premières mises en évidence que ces certains ADNec peuvent être transcrits et d'autre part parce qu'ils illustrent le potentiel et les ressources cachées des ADNec dans le contrôle des séquences répétées.

1.3 L'ADNec : un trésor génomique inexploré ?

Récemment, Elizabeth Pennisi (2017) concluait son éditorial consacré aux ADN circulaires dans *Science* par : « les rôles potentiels de ces ADN circulaires font tourner la tête des biologistes », soulignant que l'étude du mobilome suscitait un intérêt grandissant dans la communauté scientifique. Les connaissances sur ces ADN circulaires n'en sont qu'à leur commencement et ouvrent une nouvelle façon de penser sur la plasticité des génomes eucaryotes. Ces derniers mois, les travaux effectués sur les ADNec chez les animaux ont mis en évidence le potentiel de ce « pool d'acides nucléiques » (Kumar et *al.* 2017) trop longtemps négligé. Ces populations d'ADNec sont très hétérogènes selon la cellule, le tissu ou encore l'organisme étudié et des classes très variées d'ADNec ont été identifiées. Ce répertoire génomique encore largement inexploré se révèle être à la fois une source d'information notamment au niveau des mouvements des ET et de la stabilité des séquences répétées en tandem mais aussi un acteur pour l'amplification génique et une ressource génétique dans la défense contre les ET comme c'est le cas chez la paramécie.

L'étude des mobilomes devrait apporter des réponses sur les caractéristiques, les rôles et la compréhension de ces ADNec et pourrait alors jouer un rôle central en génomique et en recherche biomédicale.

2. Méthodes de détection des ADNec

2.1 Méthodes moléculaires pour la détection d'ADNec

Dès les années 80, différentes techniques de laboratoires ont été développées dans le but de d'isoler et d'identifier les ADNec. En 1981, Flavell et Ish-Horowicz ont identifié les premiers

ADNec d'ET en purifiant des cellules de *Drosophila* par gradient de chlorure de césium (CsCl) et en clonant les ADNec afin de les séquencer (Flavell et Ish-Horowitz 1981; 1983). Ils ont détecté par cette méthode des ADNec issus du produit de réverse-transcription d'un élément de la famille *Copia*. Par microscopie électronique, il est également possible d'observer les ADNec comme par exemple des ADNec purifiés par gradient de CsCl provenant d'éléments répétés chez le haricot mungo *Vigna radiata* (Bhattacharyya et Roy 1986) ou encore des microADN circulaires purifiés de cerveaux de souris après purification par exonucléase (Shibata et al. 2012). Une des méthodes les plus fréquemment utilisées est la migration d'ADN sur gel d'électrophorèse à deux dimensions (2D) « neutre-neutre » qui permet de séparer les molécules en fonction de leur taille et de leur structure, suivie d'une hybridation avec une sonde spécifique (Cohen et Lavi 1996). Très récemment, Turner et al. (2017) ont en outre développé une analyse cytogénétique et un logiciel informatique de détection des ADNec *ECdetect* qui permet de localiser automatiquement les ADNec à partir de cultures cellulaires humaines en métaphase marquées au DAPI. La différenciation entre l'ADN génomique et l'ADNec est confirmée à l'aide d'une sonde FISH (*Fluorescence In situ Hybridization*) spécifique des centromères (Figure 12b), les ADNec ne possédant pas de centromère. Cette technique peut s'avérer très puissante pour évaluer le nombre d'ADNec par cellule.

Chacune de ces méthodes permet de valider la présence d'ADNec mais ne permet pas l'identification de toute une population d'ADNec dans un tissu ou échantillon donné. Pour obtenir une caractérisation globale du mobilome d'un tissu, des méthodes de séquençage ont été développées.

2.2 Méthodes génome-entier pour la détection d'ADNec

Dernièrement, Turner et al. (2017) ont étudié les ADNec de cellules cancéreuses par reséquençage de génome entier sans réaliser d'étape de purification des ADNec au préalable. À partir d'une couverture médiane de 1,19X (très faible couverture de séquençage), ils ont détecté les CNV (*Copy Number Variation*) en analysant les variations de couverture de séquençage (voir Introduction page 26). Étant donné la faible couverture de séquençage, la puissance de détection des ADNec est faible.

En revanche, différents groupes ont développé des stratégies pour spécifiquement séquencer des ADNec. Shibata et al. (2012) ont identifié les microADN circulaires provenant de micro-

délétions chromosomiques à des loci géniques chez la souris et l'humain. Une stratégie similaire a été développée chez la levure dans le but d'analyser la présence d'ADNecc (Møller et *al.* 2015b). La méthode employée par ces deux groupes consiste à isoler l'ADNecc à partir d'une extraction classique d'ADN génomique avec digestion de l'ADN linéaire, puis à amplifier les ADNecc à l'aide d'amorces aléatoires et à séquencer les produits obtenus. Néanmoins l'analyse bioinformatique à la suite du séquençage n'a pas été développée, dans ces deux études, dans le but d'obtenir un outil dédié à l'identification des ET actifs, et aucun ET n'a été nouvellement détecté. Ainsi, aujourd'hui l'abondance et l'identification des ADNecc provenant d'ET actifs est inconnue. Pour répondre à cette question, nous avons développé une nouvelle stratégie appelée mobilome-seq, qui consiste à séquencer le mobilome d'un tissu et à analyser l'abondance des ET actifs.

Le prochain chapitre de ce manuscrit est consacré à la description de cette méthode et son application chez plusieurs espèces.

CHAPITRE 3

RÉSULTATS

Partie 1. Le séquençage de l'ADN circulaire extrachromosomique révèle l'activité des rétrotransposons chez les plantes.

Malgré l'évolution des techniques moléculaires et des outils bioinformatiques, suivre l'activité des ET en temps réel reste un défi technique. De par leur nature répétée et la complexité des génomes, l'identification d'ET actifs est donc encore très limitée. Dans le but d'évaluer l'impact des ET sur les génomes et d'améliorer notre compréhension de leurs mécanismes, nous avons développé une stratégie de séquençage à haut débit pour détecter l'ensemble des formes d'ADNec, aussi appelé mobilome (Siefert 2009), témoins de l'activité des ET et par conséquent de l'état épigénétique de l'échantillon analysé. Notre technique du mobilome-seq a été testée et validée sur 2 espèces modèles de plantes : *Arabidopsis thaliana* et *Oryza sativa* pour lesquelles nous disposons d'un génome de référence et d'une base de données d'ET certifiée. Chez le riz asiatique *Oryza sativa ssp japonica* var. *Nipponbare* nous avons séquencé le mobilome à différents stades de développement (culture cellulaire, feuilles, grains, anthères) afin d'identifier de nouveaux ET actifs. Nos premiers travaux ont fait l'objet d'une publication dans la revue *PLoS Genetics*.

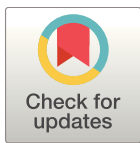
RESEARCH ARTICLE

Sequencing the extrachromosomal circular mobilome reveals retrotransposon activity in plants

Sophie Lanciano^{1,2}, Marie-Christine Carpentier^{2,3}, Christel Llauro^{2,3}, Edouard Jobet^{2,3}, Dagmara Robakowska-Hyzorek¹, Eric Lasserre^{2,3}, Alain Ghesquière¹, Olivier Panaud^{2,3}, Marie Mirouze^{1,2*}

1 Institut de Recherche pour le Développement (IRD), UMR232 DIADE, 911 Avenue Agropolis, Montpellier, France, **2** University of Perpignan, Laboratory of Plant Genome and Development, 58 Avenue Paul Alduy, Perpignan, France, **3** Centre National de la Recherche Scientifique (CNRS), Laboratory of Plant Genome and Development, 58 Avenue Paul Alduy, Perpignan, France

* marie.mirouze@ird.fr



OPEN ACCESS

Citation: Lanciano S, Carpentier M-C, Llauro C, Jobet E, Robakowska-Hyzorek D, Lasserre E, et al. (2017) Sequencing the extrachromosomal circular mobilome reveals retrotransposon activity in plants. *PLoS Genet* 13(2): e1006630. doi:10.1371/journal.pgen.1006630

Editor: Cédric Feschotte, University of Utah School of Medicine, UNITED STATES

Received: October 5, 2016

Accepted: February 10, 2017

Published: February 17, 2017

Copyright: © 2017 Lanciano et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Sequencing data generated in this study have been deposited at the European Nucleotide Archive (<http://www.ebi.ac.uk/ena/data/view/PRJEB13537>).

Funding: SL is supported by a French National Agency for Research PhD fellowship (ANR-13-JSV6-0002). This work was funded by IRD, an Agropolis Fondation grant (Labex AGRO, “RetroCrop” 1202-041) and a young investigator grant from the French National Agency for Research (ANR-13-JSV6-0002 “ExtraChrom”) to

Abstract

Retrotransposons are mobile genetic elements abundant in plant and animal genomes. While efficiently silenced by the epigenetic machinery, they can be reactivated upon stress or during development. Their level of transcription not reflecting their transposition ability, it is thus difficult to evaluate their contribution to the active mobilome. Here we applied a simple methodology based on the high throughput sequencing of extrachromosomal circular DNA (eccDNA) forms of active retrotransposons to characterize the repertoire of mobile retrotransposons in plants. This method successfully identified known active retrotransposons in both *Arabidopsis* and rice material where the epigenome is destabilized. When applying mobilome-seq to developmental stages in wild type rice, we identified *PopRice* as a highly active retrotransposon producing eccDNA forms in the wild type endosperm. The mobilome-seq strategy opens new routes for the characterization of a yet unexplored fraction of plant genomes.

Author summary

Long time considered as « junk DNA », the evolutive force of transposable elements (TEs) is now well established and TEs contribute strongly to eukaryote genome plasticity. However, it is difficult to fully characterize the mobile part of a genome, or active mobilome, and tracking TE activity remains challenging. We therefore propose to use the detection of extrachromosomal circular DNA as a diagnostic for plant TE activity. Our mobilome-seq technique allowed to identify a new active retrotransposon in wild type rice seeds, and will represent a powerful strategy in characterizing the somatic activity of TEs to evaluate their impact on genome stability and to better understand their adaptive capacity in multi-cellular eukaryotes.

MM. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Transposable elements (TEs) are major players in the evolution of animal and plant genomes [1–3]. The observation of both a complex epigenetic repression of TE expression and a large compartment occupied by TE copies in most sequenced eukaryotic genomes reflects a fine-tuned interaction between TEs and their host genomes [1–4]. TE proliferation in genomes leads to increased genomic diversity through mutations, genomic rearrangements like translocations or inversions [2], and epigenetic modifications [5]. This proliferation can also have a regulatory effect on gene expression that has been proposed to potentially result in adaptive traits [1,6,7].

According to their mode of transposition, TEs are organized into two main classes: retrotransposons (RTs) and DNA transposons (DNA-TEs). RTs multiply using a « copy and paste » strategy mediated by an RNA-intermediate, whereas DNA-TEs use a « cut and paste » mechanism [8]. During their life cycle TEs thus can exist as integrated DNA, mRNA and extrachromosomal linear DNA (S1 Fig). The extrachromosomal linear form, typical of actively proliferating TEs, can be detected by the host and may be circularized by DNA repair processes. The non-homologous end-joining mechanism and/or homologous recombination between flanking repeat sequences have been proposed to promote the circularization of extrachromosomal DNA into extrachromosomal circular DNA (eccDNA) [9–12]. There is no evidence that these eccDNAs can be re-integrated into the plant genome. Thus the formation of eccDNAs by the host could be a mechanism to limit the number of new insertions of active TEs in the genome (S1 Fig). Different types of active TEs have been detected as eccDNAs in plants such as *Tto1* [13], *Mu* [14] and *Ac/Ds* [15], however no genome-wide analysis of these forms has been performed yet. The mobilome consists of all mobile genetic elements in a cell that can be plasmids in prokaryotes or TEs in eukaryotes [16]. We will hereafter refer to the extrachromosomal forms of TEs as the reverse-transcribed mobilome.

Multiple approaches have been used to identify actively proliferating TEs at different steps of their life-cycle: (1) positional cloning of genes altered by a TE insertion (for example in rice the *hAT* DNA-TE [17] or the Long Terminal Repeat RT (LTR-RT) *Houba* [18]), (2) search for TE-insertion polymorphisms using transposon display on candidate TEs (for example rice *mPing* and *Pong* [19]), (3) transcription studies on candidate TEs using primers targeting conserved domains, for example rice LTR-RT *Tos17* [20] or through genome-wide transcriptomic analyses, for example the LTR-RT *Lullaby* in rice calli [21]. Today the most advanced technique to identify actively proliferating TEs in species where the genome sequence is available consists of whole-genome resequencing and detection of TE-associated polymorphisms using paired-end mapping [22–24].

The techniques listed above have important limitations. The analysis of transcripts by RNA-seq allows the description of transcriptionally active retrotransposons but does not take into account their capacity to produce proteins. As transcription is the first step in a retrotransposon life cycle, most copies do not go further this point, either because of post-transcriptional gene silencing activities or because they have accumulated mutations that prevent the translation of mature proteins, although some TEs with non functional proteins might parasitize other TEs [25,26]. The analysis of neo-insertions through genome resequencing is very powerful to reduce the complexity of transcriptionally active TEs to the ones that effectively produce new insertions. This approach detects breakpoints between neo-insertions and a reference genome and thus requires a high sequencing coverage more adapted to small genomes. Furthermore only fixed, transgenerational neo-insertions can be detected with high accuracy. Finally, despite the numerous pipelines developed to characterize these neo-insertions [27,28], only insertions into non repetitive regions of the genome can be accurately detected, leaving a

large part of the structural variations caused by TEs undetectable. Alternative approaches initially developed in mammals, such as retrotransposon-capture sequencing, consist in the enrichment and identification of the flanking sequences of a particular retrotransposon [28–30], but these techniques require prior knowledge of the active TE families in the species of interest. We therefore endeavor to develop a genome-wide strategy that could efficiently track potentially active TEs without full genome resequencing.

We sought to take advantage of the presence of circular forms of active TEs in the eccDNA compartment to identify active TEs in plants. Extrachromosomal DNA circles were identified decades ago in *Drosophila* [9,31] and observed by electron microscopy in *Vigna radiata* [32] and by two-dimensional gel-electrophoresis in carcinogen-treated cells [33] and in plants [34]. These eccDNAs can be formed by homologous recombination between adjacent repeats such as amplified genes [35], tandem repeats (satellite, telomeric, centromeric and ribosomal repeats) [34,36] or they can result from the linear extrachromosomal forms of active TEs [37]. These eccDNAs are ubiquitous elements and heterogeneous populations of eccDNAs seem to be present in all eukaryotic organisms [38]. Recently, sequencing of eccDNAs was experimented in mouse cells where microDNAs originating from chromosomal micro-deletions at specific gene loci [39,40] were identified. Numerous eccDNAs were detected in yeast cells [41,42], although no new active TE could be identified. Therefore, the abundance and identity of eccDNAs specifically resulting from the circularization of extrachromosomal TE DNA in multicellular organisms is not well documented. Here, we used the identification of TE eccDNA as a tool to investigate TE activation in plants and developed a dedicated computational pipeline to address this question.

We analyzed the active mobilomes from the two plant species *Arabidopsis thaliana* and *Oryza sativa*. As a proof of concept, we selected plant material where active TEs had previously been identified: a partially hypomethylated line for *A. thaliana* [43] and a callus tissue for *O. sativa*. Our mobilome-seq analyses clearly identified the two known active LTR-RTs *EVD* [44] and *Tos17* [45], in *A. thaliana* and *O. sativa* samples respectively, in their eccDNA forms. To investigate novel TE activity we applied mobilome-seq to wild type rice seeds and identified *PopRice* LTR-RTs as producing large amounts of eccDNAs specifically in the endosperm tissue. We propose that the mobilome-seq strategy could help identifying mobile TEs in different species to better understand the impact of the active mobilome on the host genome.

Results

Enrichment and sequencing of eccDNAs

In order to isolate and to sequence eccDNAs, total DNA was first extracted from plant tissues (Fig 1). Linear genomic DNA was digested with an exonuclease and the remaining eccDNA molecules were then amplified by rolling circle amplification (RCA) using random primers. This method does therefore not require any *a priori* knowledge on TEs for a given sample. We first performed this experiment on samples from *A. thaliana* Columbia wild type plants as a negative control (Col WT) and on an epigenetic recombinant inbred line (epiRIL12 hereafter called epi12) where an hypomethylated retrotransposon (*EVD/ATCOPIA93*) was detected as actively proliferating [44], as a positive control. Southern blot validation assays using an *EVD* specific probe were performed to analyze the enrichment of eccDNAs before and after the RCA step (Fig 2A). A signal corresponding to digested *EVD* eccDNAs was detected in samples from siliques and flowers from epi12 plants after RCA, but not in samples from WT plants. No signal could be detected in the absence of RCA indicating that most genomic DNA had been degraded after the exonuclease treatment. We used this material for high throughput sequencing.

Detection of *EVD* by mobilome-seq in *A. thaliana* hypomethylated plants

We performed mobilome-seq on Col WT and epi12 siliques samples as shown in Fig 1. After mapping the reads on the reference genome of *A. thaliana* we detected peaks of high coverage in both WT and epi12 mobilome-seq libraries (S2 Fig) corresponding to ribosomal DNA (rDNA) loci that are known to produce eccDNAs [34]. All peaks of high coverage corresponding to TEs in both WT and epi12 are listed in S5 Table. In particular, peaks corresponding to *EVD* were specifically detected in epi12 (Fig 2B, S3A Fig and S5 Table). *EVD* is a 5,3 kilobases (kb)-long LTR-RT present in two full-length copies in the genome of *A. thaliana* ecotype *Columbia*. *EVD* is transcribed and mobilized in *met1*-derived epiRILs [44] and produces eccDNA copies [46]. Due to the repetitive nature of TEs, reads corresponding to *EVD* eccDNA

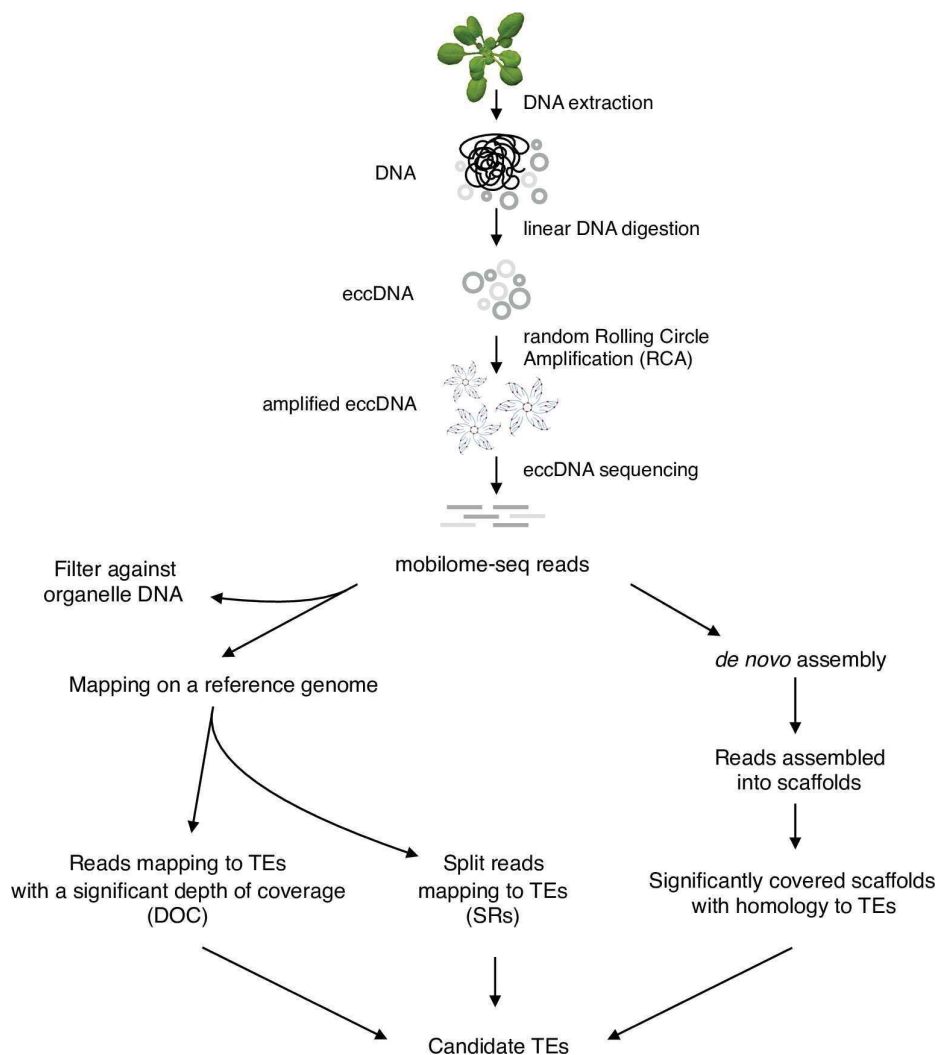


Fig 1. The mobilome-seq approach in plants. A schematic view of the main steps involved in the selection and amplification of the extrachromosomal circular molecules in plants. After DNA extraction, linear DNA molecules are digested and circular molecules are randomly amplified using rolling circle amplification. This DNA material is used for high-throughput sequencing. Mobilome-seq data analysis consists in characterizing the depth of coverage (DOC) of mapped reads and the presence of split reads (SRs) at TE loci and the detection of *de novo* assembled scaffolds corresponding to these TEs.

doi:10.1371/journal.pgen.1006630.g001

can map against full-length and also truncated copies present in the genome explaining why all regions corresponding to *EVD* are more or less covered. Nevertheless, the two full-length copies (on chromosomes 1 and 5) are the most significantly covered with a p-value < 10⁻⁸ (S3B and S3C Fig). The *EVD* locus on chromosome 5 is highly covered in the epi12 mobilome-seq library compared to the WT library, with a depth of coverage (DOC) of 3500X versus 1X, respectively (Fig 2C). To further identify the presence of reads corresponding to eccDNA junctions, we specifically detected split reads (SRs) as paired-reads that are not correctly mapped onto the reference genome (see Methods). We could detect a high number of SRs at both 5' and 3' ends of *EVD* in the epi12 mobilome-seq data compared to WT (Fig 2C and S4 Fig) suggesting the presence of reads spanning the circular junction. A closer examination of some of these reads revealed that they indeed correspond to 2LTR junctions (S5 Fig). While 142 TEs were detected as overexpressed in epi12 at the transcriptional level [47], the mobilome-seq data suggest that only *EVD* produce circular copies (S6 Fig).

Tos17 is highly enriched in the *O. sativa* callus tissue mobilome-seq library

We then analyzed mobilome-seq libraries from *O. sativa* ssp *japonica* cv *Nipponbare*, a species with a larger genome (400Mb) than *A. thaliana* (135Mb) and a three times bigger proportion

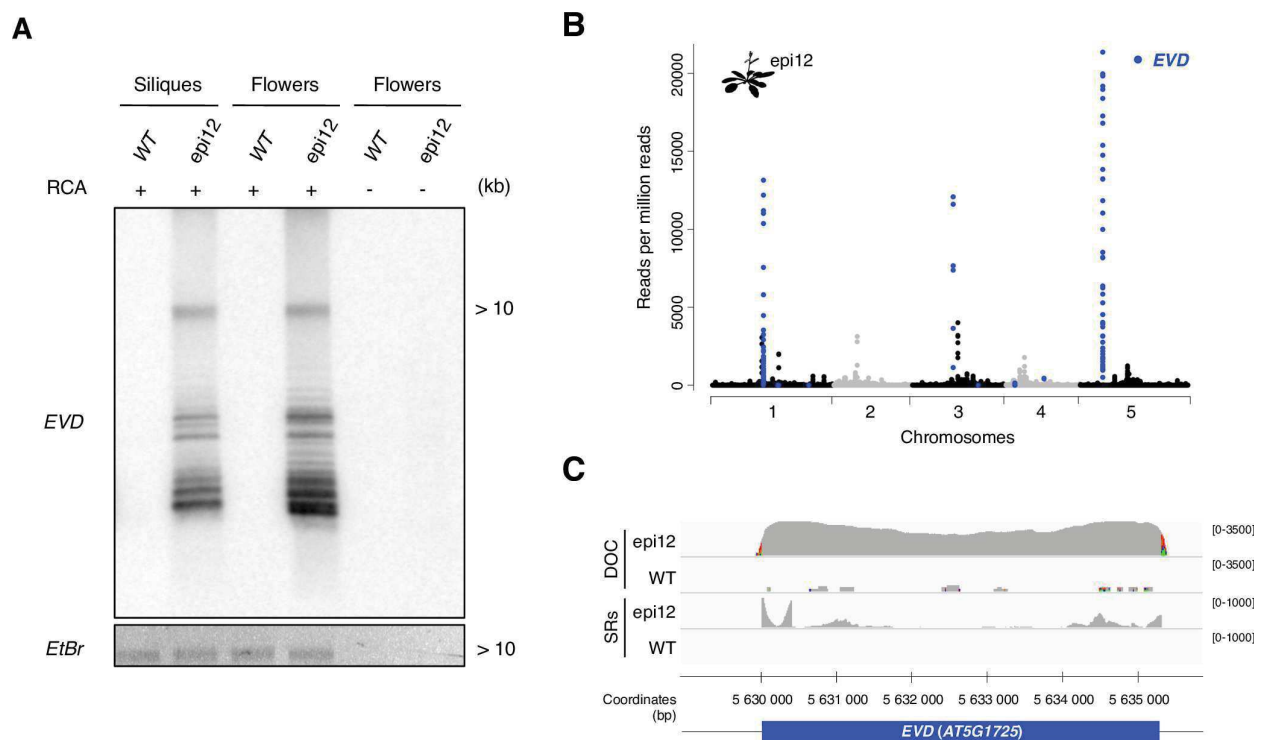


Fig 2. Mobilome-seq detection of *EVD*, a known active retrotransposon in *Arabidopsis*. (A) Southern blot experiment using *Hind*III-digested eccDNAs amplified from *A. thaliana* WT and epi12 flowers and silicles and detected with a probe specific for the *EVD* retrotransposon active in the line epi12. RCA: rolling circle amplification. The ethidium bromide (*EtBr*) gel picture is shown as a loading control. (B) Abundance of reads mapping at TE-annotated loci in the *A. thaliana* epi12 line mobilome-seq library. Each dot represents the normalized coverage per million mapped reads per all TE-containing 100bp windows obtained after aligning the sequenced reads on the *A. thaliana* reference genome. Blue dots indicate the windows corresponding to annotated *EVD* genomic loci. (C) Detail of the depth of coverage of total mapped reads (DOC) and split reads (SR) abundance of the *A. thaliana* epi12 and WT mobilome-seq data at the *EVD* locus on chromosome 5 (blue bar). Grey peaks: read abundance (not normalized), DOC: depth of coverage for all aligned reads, SRs: split reads, WT: wild type silicles, epi12: epi12 silicles. Maximum coverage is indicated on the right. Colors indicate the presence of SNPs.

doi:10.1371/journal.pgen.1006630.g002

of TEs (45% in *O. sativa* against 15% in *A. thaliana*), using both leaf material and callus tissue. TEs with high coverage in *O. sativa* mobilome-seq libraries are listed in S5 Table. More specifically, peaks corresponding to the *Tos17* family were detected in the mobilome-seq libraries of callus tissue but not in leaves (Fig 3A and 3B, S7 Fig). *Tos17* is a 4,1 kb-long LTR-RT present in two copies in the *O. sativa* genome (on chromosomes 7 and 10), the copy on chromosome 7 being active in calli [13]. The DOC analysis indicated a clear enrichment (DOC = 200X) at the *Tos17* locus on chromosome 7 in the callus mobilome-seq library compared to the leaf mobilome-seq library (<1X) (Fig 3B and S7B Fig). SRs were detected on both ends of *Tos17* suggesting the presence of reads spanning the junction. The presence of *Tos17* eccDNAs was confirmed by an inverse PCR assay (Fig 3C) and a closer inspection of SRs identified reads spanning the 2LTR-circle junction (S8A Fig). Moreover we have also analyzed the coverage of *Lullaby*, a LTR-RT active in calli [21]. A low coverage from 10X to 12X was detected in the callus mobilome-seq library and the presence of reads spanning junction of *Lullaby* eccDNAs was confirmed (S8B Fig). Altogether, these results show that known active LTR-RTs could be detected using the mobilome-seq approach, suggesting that this technique can be used to identify new active TEs in plants.

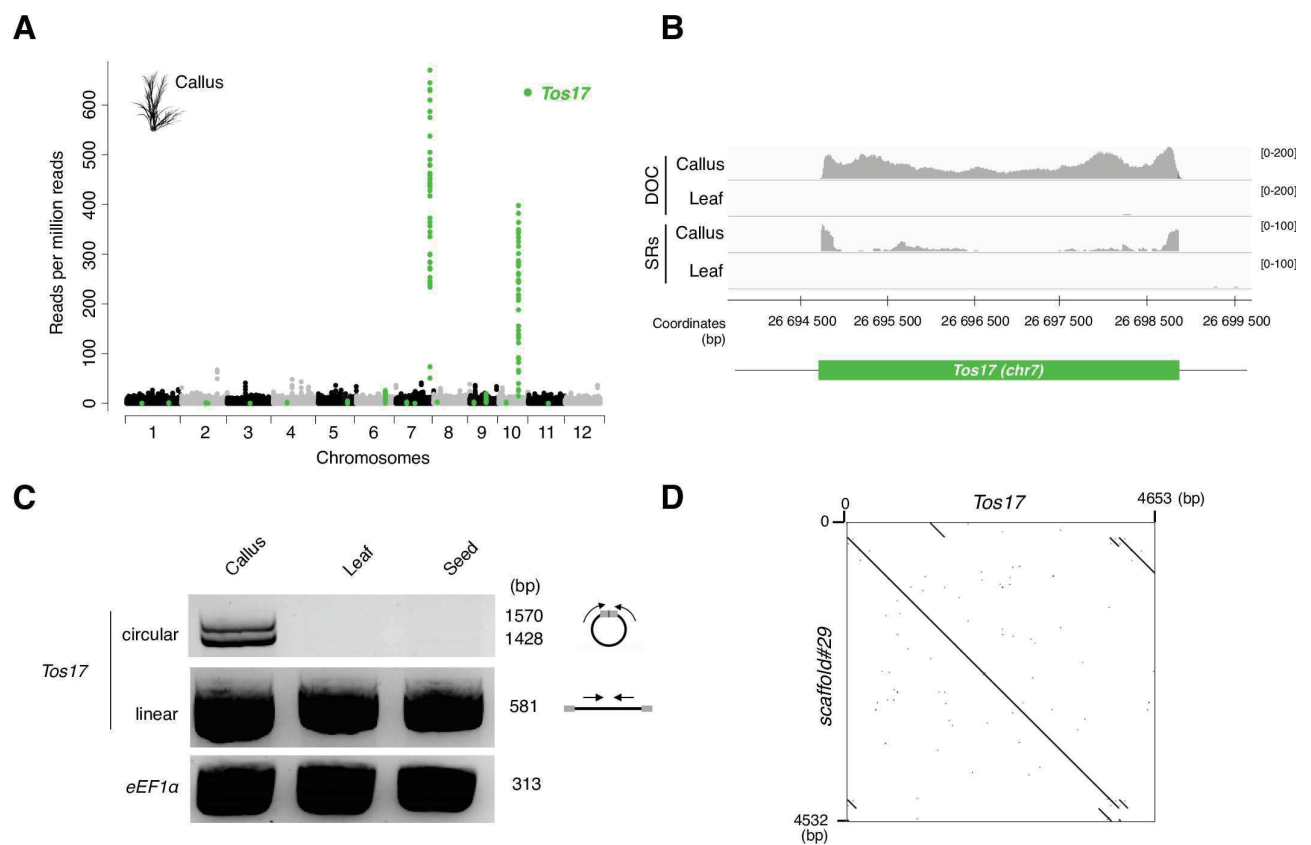


Fig 3. Mobilome-seq detection of *Tos17*, a known active retrotransposon in rice callus. (A) Abundance of reads mapping at TE-annotated loci in the *O. sativa* WT callus mobilome-seq library. Each dot represents the normalized coverage per million mapped reads per all TE-containing 100bp windows obtained after aligning the sequenced reads on the *O. sativa* reference genome. Green dots indicate the windows corresponding to annotated *Tos17* genomic loci. (B) Detail of the depth of coverage of total mapped reads (DOC) and split reads (SRs) abundance of the *O. sativa* WT callus and leaf mobilome-seq library at the *Tos17* locus on chromosome 7 (green bar). Legend as in Fig 2C. (C) Detection of circular forms of *Tos17* using inverse PCR with primers localization depicted on the right (black bar: *Tos17* element, arrows: PCR primers, grey boxes: LTRs). Upper gel: PCR amplification of *Tos17* circles, middle gel: control PCR for *Tos17* detection, lower gel: PCR using *eEF1a* primers as loading control. (D) Dotter alignment of the scaffold #29 obtained after *de novo* assembly of callus mobilome-seq library and *Tos17*.

doi:10.1371/journal.pgen.1006630.g003

Identification of a new active LTR-RT in wild type rice

Epigenomic studies have revealed a release of TE transcriptional silencing during plant development [48–51]. In a first attempt to understand the possible role of TEs reactivation during plant development, we performed mobilome-seq analyses on DNA extracted from whole rice seeds. Some TE regions were significantly highly covered in this mobilome-seq library (Fig 4A and 4B, S9 Fig), most of these regions corresponding to TEs belonging to a single subfamily of *Osr4*. *Osr4* [52] is a large family of 5.7 kb-long LTR-RTs comprising 47 members in the *O. sativa* ssp *japonica* cv *Nipponbare* genome (Fig 4C and S3 Table). To differentiate *Osr4* active

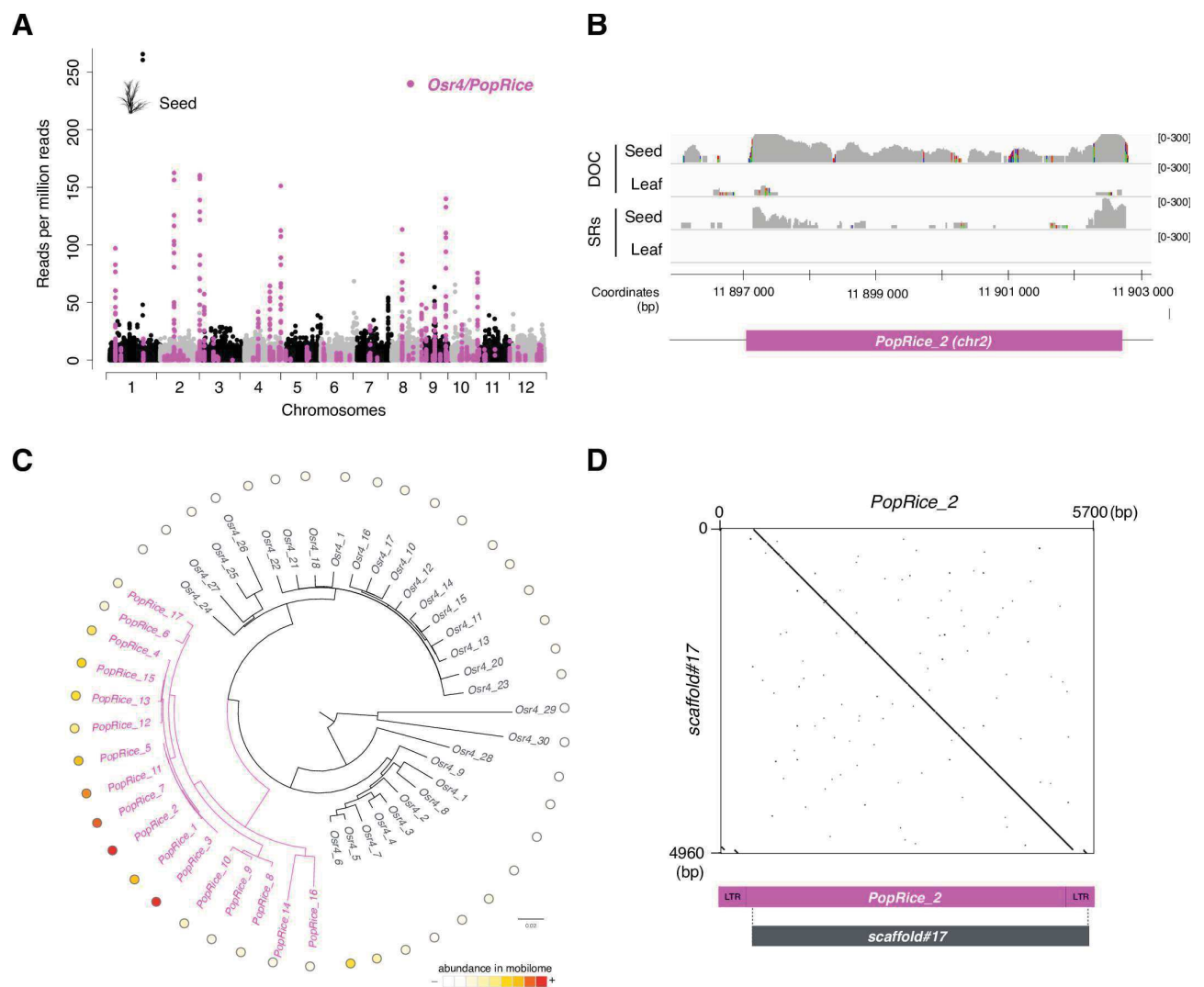


Fig 4. Mobilome-seq detection of a novel active retrotransposon in rice seeds. (A) Genome-wide analysis of mobilome-seq data identifies the *PopRice* retrotransposon family as the most represented active family in WT rice seeds. Legend as in Fig 3A. Pink dots indicate the windows corresponding to *Osr4* and *PopRice* loci. (B) Detail of the depth of coverage of total mapped reads and split reads abundance of the *O. sativa* WT seeds mobilome-seq library at the *PopRice* locus on chromosome 2 (pink bar) for callus and leaf mobilome-seq data. Legend as in Fig 2C. (C) Phylogenetic tree showing that *PopRice* is a distinct subfamily of *Osr4* LTR-RT. The relative DOC calculated from two biological replicates in WT seed mobilome-seq data is indicated as a heatmap. (D) Dotter alignment of the scaffold #17 obtained after *de novo* assembly of WT seed mobilome-seq library and a *PopRice* element.

doi:10.1371/journal.pgen.1006630.g004

and non-active members we hereafter refer to the subfamily enriched in the seed mobilome-seq library as the *PopRice* family. The *PopRice* family is composed of 17 full-length copies in the reference genome. Some of these loci are highly covered in the seed mobilome-seq library with a DOC reaching 300X (Fig 4B), showing that some members of this subfamily are actively producing eccDNAs in wild type rice seeds. We detected SRs located on both 5' and 3' ends of some *PopRice* loci (Fig 4B and S9 Fig). A closer examination of reads spanning junctions has also confirmed the presence of *PopRice* eccDNAs (S10 Fig). Further sequence analyses of *PopRice* family revealed that the most active *PopRice* copies form a subgroup of 5 members (Fig 4C).

de novo assembly can be used to identify the most active LTR-RTs without a reference genome

We performed *de novo* assembly of mobilome-seq libraries to determine whether *EVD*, *Tos17* and *PopRice* could be detected without mapping on a reference genome. We did not detect scaffolds corresponding to *EVD* when performing *de novo* assembly on the WT mobilome-seq library. In the Arabidopsis epi12 mobilome-seq library, *de novo* assembly resulted in three main scaffolds corresponding to *EVD* (S11 Fig). These three scaffolds all result from the assembly of a high number of reads (59,943; 49,098 and 19,424 reads per million (rpm), respectively, p-value < 0.05, negative binomial distribution). In the rice callus mobilome-seq library the most highly covered scaffold (3,906 rpm) with homology to TEs corresponded to *Tos17* (100% identity over 4,532 base pairs (bp), Fig 3D). This suggests that *de novo* assembly can be used to identify active RTs. In the seed mobilome-seq library, the most significantly covered scaffold (3,473 rpm) showed 99% of sequence identity with a *PopRice* consensus sequence over 4,960 bp (Fig 4D). Only the ends of *PopRice* were not assembled in this scaffold, likely due to the repetitive nature of LTR sequences.

Activation of *PopRice* in the endosperm tissue

To further validate the presence of extrachromosomal DNA fragments originating from *PopRice* in WT rice seeds, we performed a Southern blot experiment using non-amplified and non-digested genomic DNA (Fig 5A). Using a *PopRice* specific probe, a signal corresponding to a 5 kb fragment was identified in genomic DNA samples extracted from seeds but not from leaves, revealing a massive accumulation of *PopRice* extrachromosomal copies in wild type seeds. A Southern blot performed on genomic DNA obtained from dissected seed tissues further revealed that *PopRice* extrachromosomal DNA could only be detected in the endosperm tissue but not in the embryo or seed coat (Fig 5B). This result was confirmed by inverse PCR assays (S12 Fig). To characterize the kinetics of *PopRice* activation during plant development we used inverse PCR to detect the presence of *PopRice* eccDNAs at different developmental stages. *PopRice* eccDNAs seemed to be specific of seed tissues from the embryo developmental stage (corresponding to immature seeds, from 3 to 5 days after pollination) to the germination, however eccDNAs were not detected in roots and cotyledons after germination (Fig 5C).

To rule out the possibility that *PopRice* circles could originate from homologous recombination between its endogenous LTR sequences, we identified mobilome-seq reads corresponding to 2LTR junctions in the seed libraries (S10 Fig). The presence of these reads confirmed that non-homologous end-joining of reverse transcription products, and not homologous recombination at the endogenous genomic location, is responsible for the formation of *PopRice* eccDNA molecules. Additionally we analyzed *PopRice* transcription, the mRNAs being the precursors of the eccDNAs. RT-qPCR assays showed that *PopRice* and *Osr4* members are highly transcribed in seeds compared to leaves and flowers (Fig 5D). The level of expression seems higher when all

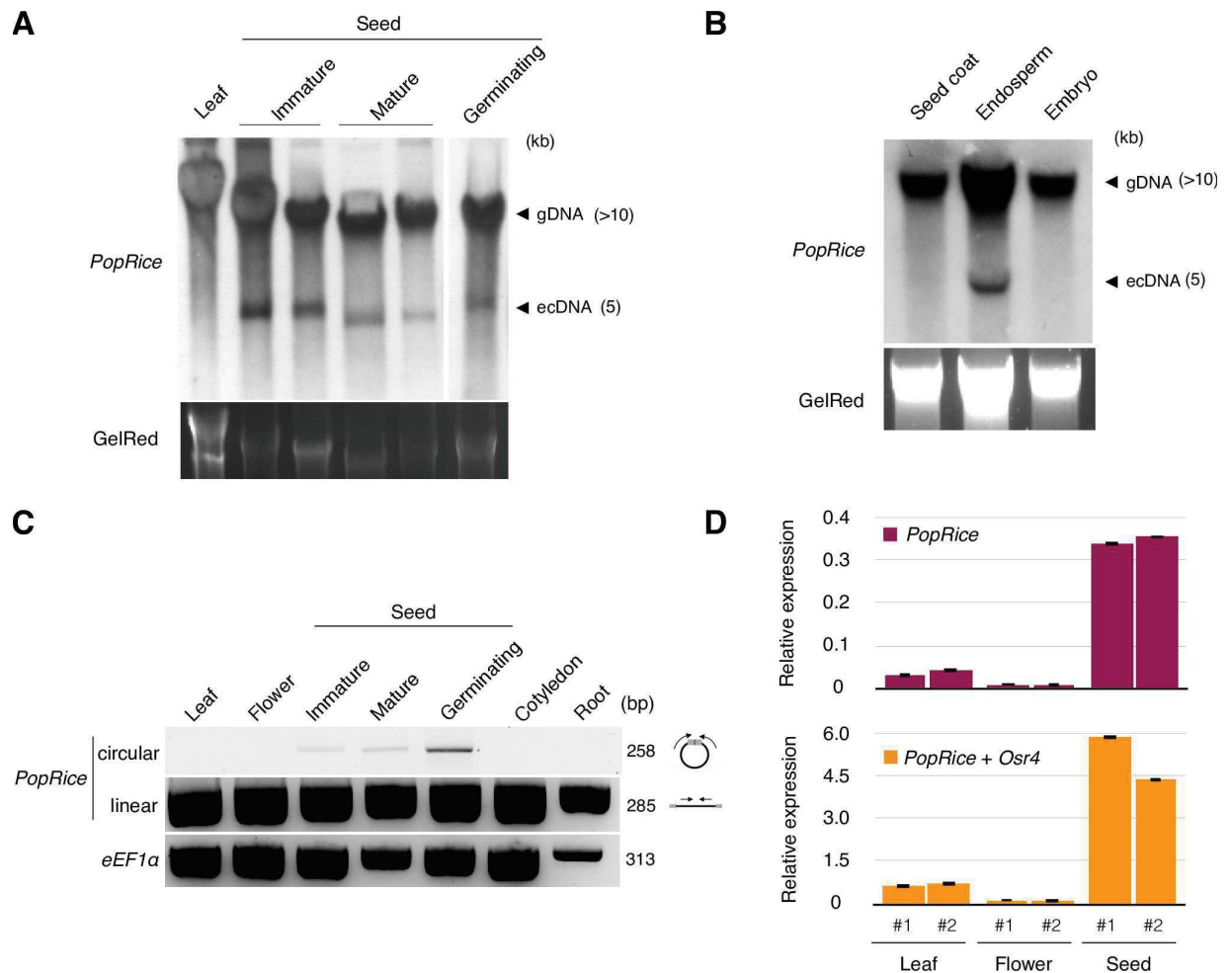


Fig 5. *PopRice* retrotransposons produce extrachromosomal DNA during seed development in wild type rice. (A) Southern blot experiment using non-digested genomic DNA extracted from WT rice leaves and seeds at different stages as indicated and detected with a *PopRice* specific probe (*gDNA*: genomic DNA, *ecDNA*: extrachromosomal DNA). The GelRed gel picture is shown as a loading control. (B) Southern blot experiment using non-digested genomic DNA extracted from dissected rice seed tissues as indicated and detected with a *PopRice* specific probe. Legend as in A. (C) Detection of *PopRice* circular forms using inverse PCR with primers localization depicted on the right (black bar: *PopRice* element, arrows: PCR primers, grey boxes: LTRs). Upper gel: PCR amplification of *PopRice* circles, middle gel: control PCR for *PopRice* detection, lower gel: PCR using *eEF1a* primers as control. (D) qRT-PCR analysis of *PopRice* and *Osr4* transcripts in WT rice leaves, flowers and mature seeds. Two pairs of primers were used: *PopRice* specific primers (top) and primers specific for the whole *Osr4* family (including *PopRice*) (bottom). The relative expression levels were calculated using *eIF-5a* as a reference, error bars indicate technical replicates, two biological replicates are shown for each tissue.

doi:10.1371/journal.pgen.1006630.g005

elements of the *Osr4* family are considered suggesting that the whole *Osr4* family is transcriptionally active, although only *PopRice* eccDNAs could be detected (Fig 4C).

Discussion

In all eukaryotic organisms, eccDNA molecules are ubiquitous elements and constitute an heterogeneous population of circular molecules that can originate from repeats such as rDNA clusters through homologous recombination [34,53] or from active TEs (through circularization of linear extrachromosomal forms). We took advantage of the detection of eccDNAs by next generation sequencing (NGS) to explore the extrachromosomal circular mobilome in plants.

As a proof of concept we analyzed samples from *A. thaliana* and *O. sativa* material for which actively proliferating TEs had previously been characterized [44,45]. Identification of two well-characterized active LTR-RTs, *EVD* and *Tos17* in an *A. thaliana* hypomethylated line and in rice callus tissue, respectively, confirmed that this method is efficient to capture actively proliferating retrotransposons in plants. The detection of rDNA circles validates the enrichment of eccDNAs in our libraries and thus constitutes another positive control. Moreover our observations suggest that some TEs might form different circles where SRs spread into internal regions of TEs reflecting a possible heterogeneity of these extrachromosomal circles. The mobilome-seq strategy exploits the advantages of NGS and requires a low sequencing coverage for each library. Indeed only a minor fraction of the genome is sequenced, opening the future possibility of applying the technique to very large genomes, for which resequencing techniques are not affordable and technically challenging. Furthermore, the *de novo* assembly analyses might represent a precious and powerful method to study the active mobilome of species for which a reference genome is lacking.

Developmental relaxation of TE control has been documented in plant tissues accompanying the gametes: vegetative nucleus for the pollen and endosperm for the ovary [51]. In rice, DNA methylome analyses revealed a global hypomethylation in the endosperm [50,54,55] confirming previous results in Arabidopsis [49,56,57]. This suggests that TE activity could be increased in these tissues; however, to our knowledge, only the proliferation of a DNA-TE Mule element has been documented so far in the *A. thaliana* pollen [48]. Here, our seed mobilome-seq analyses reproducibly revealed that the *PopRice* family of autonomous LTR-RTs produces extrachromosomal copies. These copies can be detected on a Southern blot analysis of untreated genomic DNA showing that *PopRice* extrachromosomal copies accumulate in wild type rice endosperm. Further studies will help evaluating the proliferation of *PopRice* in the endosperm genome. *PopRice* transcripts could be detected in seeds suggesting that eccDNAs indeed originate from reverse transcription of these transcripts. Genomic imprinting could explain the transcriptional activity of some TEs in the endosperm. According to a study by Luo *et al.* [58], only two relatively ancient copies (*PopRice_16* and *Osr4_28*) are localized in introns of paternally imprinted genes (LOC_Os11g09329 and LOC_Os08g24420, respectively) suggesting that imprinting might not be the only trigger for *PopRice/Osr4* transcriptional activation in the endosperm. Recently, Cheng *et al.* have shown that members of the *Osr4* LTR-RT family could retrotranspose in *oscmt3* mutants, affected in a chromomethylase involved in DNA methylation, through genome resequencing [59]. Interestingly all neo-insertions are due to the *PopRice* subfamily suggesting that this subfamily contains all potentially active members and that they are under the epigenetic control of OsCMT3. This transgenerational control is reminiscent of the regulation of *Onsen*, an *A. thaliana* LTR-RT that produces eccDNA molecules after heat stress. However in the case of *Onsen* transgenerational neo-insertions are only detected in mutants affected in the RNA-directed DNA methylation (RdDM) pathway, but not in the CMT3 pathway [60]. The precise role of the RdDM pathway in the transgenerational control of these LTR-RTs neo-insertions is not yet elucidated [61].

Using a newly developed method to sequence and identify eccDNAs originating from TEs, we have characterized a yet unexplored fraction of plant DNA. This study revealed the reactivation of an endosperm-specific LTR-RT in rice. This LTR-RT family seems to be under the control of both epigenetic and post-transcriptional regulation. Furthermore the identification of this only LTR-RT family active in the endosperm suggests that the global hypomethylation occurring in this tissue is not sufficient to trigger a massive reactivation of TEs. By giving an insight into actively proliferating retrotransposons in plants, the mobilome-seq approach is likely to expand our understanding of TE activity in plants and of their putative contribution in response to stress and during plant development.

Materials and methods

Plant material

Arabidopsis thaliana WT ecotype Columbia-0 and epiRIL12 plants from the eighth generation [43] were grown in soil under a 16h/8h (light/dark) cycle after 2 days at 4°C for stratification. Florets and 1–2 cm green siliques were harvested 3 days to 2 weeks after pollination, respectively. *Oryza sativa* ssp. *japonica* cv. *Nipponbare* rice plants were cultivated in a growth chamber (Percival, USA) under a 12h light-dark cycle (12h-28°C/12h-26°C) and with a relative humidity of 80% during the day and 70% during the night. The light intensity varied gradually in 40 min at the beginning and end of the day. Grain material was harvested 3 to 5 to 15 days after pollination for the immature and mature stage, respectively. Seeds were germinated in the dark on a humid Whatman paper for 5 days before harvest. Dissection of seeds was performed under the binocular on mature seeds. Callus material was previously described [21].

DNA extraction

For each plant sample, total DNA was extracted using the plant DNeasy mini kit (Qiagen) according to the manufacturer's instructions. A DNA pre-extraction was performed for rice grains to optimize DNA quantity and quality. Grains were grinded in an extraction buffer (Tris-HCl pH8, NaCl 250mM, EDTA 50mM, 0.2% SDS) and were incubated 30 min at 65°C. DNA samples were precipitated with 0.7 volume of isopropanol and the DNA pellet was directly resuspended in the plant DNeasy mini buffer (Qiagen).

Extrachromosomal circular DNA enrichment

To remove large genomic linear fragments 5µg of genomic DNA for each sample were purified using a GeneClean kit (MPBio) according to the manufacturer's instructions. eccDNA was isolated from 28µl of the GeneClean product using the PlasmidSafe DNase (Epicentre) according to the manufacturer's instructions, except that the 37°C incubation was performed for 17h. The PlasmidSafe exonuclease digests double-stranded linear DNA to deoxynucleotides while leaving circular DNA intact. DNA samples were precipitated by adding 0.1 volume of 3M sodium acetate (pH 5.2), 2.5 volumes of ethanol and 1 µl of glycogen (Fisher) and incubating overnight at -20°C. The precipitated circular DNA was amplified by random RCA using the Illustra TempliPhi kit (GE Healthcare). For this, the DNA pellet was directly resuspended in the Illustra TempliPhi Sample Buffer, and the reaction was performed according to the manufacturer's instructions except that the incubation was performed for 65h at 28°C. One tenth of each amplified DNA sample was digested with restriction enzymes and loaded on an agarose gel electrophoresis to control the DNA quality and amplification. Then, the DNA concentration was determined using the DNA PicoGreen kit (Invitrogen) following the manufacturer's instructions, the fluorescence being read using a LightCycler480 (Roche). The samples were diluted to a final concentration of 0.2 ng/µl in order to prepare the libraries for sequencing.

Libraries preparation and sequencing

One nanogram of DNA from each sample was used to prepare the libraries using the Nextera XT library kit (Illumina) according to the manufacturer's user guide. Each mobilome-seq library was amplified by 12 cycles of PCR using index primers. DNA quality and concentration were determined using a high sensitivity DNA Bioanalyzer chip (Agilent Technologies). Samples were pooled and loaded onto a flow cell and 2x250 nucleotides paired-end sequencing was performed using the MiSeq platform (Illumina). Up to twelve mobilome-seq libraries were

pooled into one run and an average of 1 million reads per library were obtained (S1 and S2 Tables). Illumina reads were collected for the analysis as FASTQ files.

Data analysis

To analyze the sequencing reads we anticipated that the eccDNAs of interest originating from mobile TEs should represent a very small fraction of the genome and consequently that the loci from where these eccDNAs were produced should be highly covered. Furthermore, as these molecules are circular, reads spanning the junction of the circles should not map properly on the reference genome because such junctions do not exist in the chromosomes. However, these reads might map on two different locations (start and end of the element). Thus the eccDNAs of interest should fit to the two following criteria: (1) high DOC and (2) presence of SRs when mapped to a reference genome. Finally, due to the repetitive nature of TEs, we reasoned that the read-mapping coverage could be less sensitive for large TE families as reads could be dispersed amongst related TE copies. Therefore we should be able to identify the most abundant eccDNAs by analyzing highly covered scaffolds after *de novo* assembly.

Read mapping

Quality control of FASTQ files was evaluated using the FastQC tool (version 0.10.1 www.bioinformatics.babraham.ac.uk/projects/fastqc). To remove any read originating from organelle circular genomes, reads were mapped against the mitochondria (NCBI GenBank Y08501.2 for *A. thaliana*; GenBank NC_011033 for *O. sativa*) and chloroplast genomes (GenBank AP000423.1 for *A. thaliana*; GenBank X15901 for *O. sativa*) using the program BOWTIE2 version 2.2.2 [62] with—sensitive local mapping. Unmapped reads were considered for the next analysis and were mapped against a genome of reference, TAIR10 (The Arabidopsis Information Resource, <http://www.arabidopsis.org>) for *A. thaliana*, IRGSP1.0 (International Rice Genome Sequencing Project version 5 <http://rgp.dna.affrc.go.jp/E/IRGSP/Build5.html>) for *O. sativa*. The parameters used for the mapping were as follows:—sensitive local mapping, no multiple-mappings (-k 1) so that only the best hit is kept per read-pair. DNA from both mitochondria and chloroplast genomes are integrated in nuclear genomes. To completely eliminate these regions from our data, sequencing reads were simulated from organelle genome using the dwgsim program (version 0.1.10) and the *fasta* files were mapped against the corresponding reference genome. A total of 816,300 bp and 1,697,400 bp were masked in *A. thaliana* and *O. sativa*, respectively, and TE containing regions cover 24,786,000 bp and 194,224,800 bp in *A. thaliana* and *O. sativa*, respectively. A *.bam* file with all genome regions corresponding to organelle integrated sequences was obtained for each species and was used to filter our alignment files using the intersect module of BEDTools version 2.21.0 (option -v). Finally, for each library, a *.bam* alignment file corresponding to enriched genomic regions was considered for statistical analysis and visualized with the Integrative Genomics Viewer (IGV) software (<https://www.broadinstitute.org/igv/home>) and Circos [63].

Statistical analysis

For each species, the reference genome was split into consecutive windows of 100bp for each library and the coverageBED module of BEDTools was used to determine the read coverage depth of these non-overlapping windows. The coverage data was normalized by the total number of reads which mapped on the genome expressed in rpm and statistical analysis was performed on these files. First we determined covered regions using the Poisson distribution that best fits our data with a p-value $< 10^{-5}$ for each library. All uncovered regions were removed from our coverage files. On the covered regions we applied a negative binomial distribution to

identify peaks of higher coverage with a p-value $< 10^{-3}$. Finally, regions corresponding to the peaks were selected and annotated using *.gff* files (S5 Table). All statistical analysis and graphics were performed using R (Rstudio package version 0.98.1091, www.r-project.org/).

de novo assembly

Mobilome-seq reads were assembled *de novo* using the A5-miseq pipeline [64]. For each library, *.fasta* and *.bam* files were obtained and the *idxstats* module of SAMtools was used to determine the read number corresponding to each assembled scaffold. The coverage data was normalized by the total number of reads used for the *de novo* assembly expressed in million reads (rpm). We applied a negative binomial distribution to identify significantly covered scaffolds (p-value < 0.05). Filtered scaffolds were annotated using a BLAST analysis (-p -m 8) against organelle genomes and a TE database allowing for one hit per scaffold (-b 1 -v 1 options) and for an e-value $< 10^{-2}$. For *A. thaliana* we used the TE database based on TAIR10 annotation and established by H. Quesneville (www.arabidopsis.org/); for *O. sativa* we used an in house curated database (www.panaudlab.org/). Resulting hits were filtered to keep only scaffolds with a HSP ≥ 100 bp.

Split-reads detection

Reads spanning 2 LTR junctions constitute an evidence of a circular TE and were detected using a SR mapping strategy. Reads were aligned against the reference genomes using the *segemehl* software [65] with the following parameters: -S (SR mapping) -A 95 (accuracy of 95%) -U 24 (minimum score of 24) -Z 25 (minimum length of 25) -W 95 (alignment covered on 95% of the read). Split reads were collected from *.bam* files based on the FLAG field of each read, using the *view* module of SAMtools. Therefore, only reads which were not mapped in a proper pair (-f 14) and which have multiple primary alignments (-F 256) were considered as SR. The *coverageBED* module of BEDTools was used to determine the read coverage depth of these SR *.bam* files were visualized with IGV.

Candidate TEs

To determine TE loci of interest in each library, we first filtered the *.bam* files obtained from *bowtie2* mapping (for the DOC) for TEs covered for 90% of their length and with a DOC > 10 rpm. Using the *.bam* files obtained from *segemehl* mapping we selected the TE loci with a SR coverage > 10 rpm. We selected the TE loci fitting both criteria and which length is > 100 bp (S6 Table). The TE loci were visualized with *Circos* [63]. The in-house developed code for DOC and SR detection and for the establishment of the candidate TE list can be accessed upon request.

Southern blot analysis

Total genomic DNA was extracted using the CTAB method [66] and samples were loaded on a 0.8% agarose gel and transferred onto Hybond-N+ nylon membrane (GE Healthcare). The Southern blot on *A. thaliana* material was performed as previously described [44] using a radioactive probe. The Southern blot on *O. sativa* material was performed using a non-radioactive probe and the hybridization signal was detected with the DIG system (Roche) following the manufacturer's instructions. Stringency washes were performed at 65°C in 0.5X SSC. Probes were amplified from genomic DNA by PCR using primers listed in the S4 Table. The Southern blots on Fig 5A and 5B were repeated twice using biological replicates.

Transcription analysis

Total RNAs were isolated from leaves, flowers and seeds using the Tri-reagent (MRC) according to the manufacturer's instructions. RNAs were treated with DNase from RQ1 kit (Promega) and 1.25 μ g were reverse-transcribed into cDNAs using the GoScript kit (Promega). Analyses by quantitative real-time PCR (qRT-PCR) were established using 7 to 35 ng of cDNA. qRT-PCRs were run on a LightCycler 480 (Roche) using Takyon No Rox SYBR MasterMix dTTP Blue Kit (Eurogentec) according to the manufacturer's instructions. The qRT-PCR conditions were the following: a first denaturation step at 95°C for 5 min followed by 40 cycles at 95°C for 15s, an annealing and elongation step at 60°C for 60s, and a melting curve analysis at 95°C for 10s, 60°C for 10s, an increase of 0.04°C per second until 95°C and a final step of cooling at 40°C for 30s. Two biological replicates were analyzed for each tissue. *PopRice* and *Osr4* expression levels relative to *eIF-5a* [67] were calculated using the formula: $2^{-(\text{mean (CT PopRice)} - \text{CT internal references})}$. Primers were used with a concentration of 2 μ M and primers details are given in the S4 Table and S13 Fig.

PCR validations

PCR reactions were performed using 2 μ l of DNA (before or after the RCA amplification) in a final volume of 15 μ l, using the GoTaq polymerase (Promega). All primer pairs were designed using Primer3 (www.primer3.ut.ee) and quality-checked using OligoCalc (www.basic.northwestern.edu/biotools/oligocalc.html) and BLAST (www.ncbi.nlm.nih.gov/BLAST/). Target sequences and primers used are shown in S4 Table. The PCR conditions were the following: a first denaturation step at 95°C for 5 min followed by 30 cycles at 95°C for 30s, an annealing step (temperature details in S4 Table) for 30s, an elongation step at 72°C for 20 seconds, and a final extension step at 72°C for 5 min. 8 μ l of PCR products were deposited on a 1,5% agarose gel and run at 135mV for 30 min. DNA was stained using a GelRed dye (Biotium). Gel pictures were obtained using an UGenius gel imaging system (Syngene). All PCR assays were repeated at least twice using biological replicates.

Identification of *PopRice* subfamily

To determine the evolutionary story of *PopRice* elements, a consensus sequence of *PopRice* was used to detect by BLAST all LTR-RTs from the IRGSP-1.0 reference genome belonging to the same family (HSP>4000bp, minimum of 70% of identity, e-value < e^{-50}). All selected sequences were aligned with MAFFT (<http://mafft.cbrc.jp/alignment/server/>). Alignments were analyzed using SEAVIEW (<http://doua.prabi.fr/software/seaview>) and all incomplete elements were removed. 47 elements were selected for the *Osr4* family (comprising *PopRice* sequences) and a phylogenetic tree was built with PhyML and visualized with FigTree (<http://tree.bio.ed.ac.uk/software/figtree/>).

Accession number

Sequencing data generated in this study have been deposited at the European Nucleotide Archive (ENA, www.ebi.ac.uk/ena) under the accession number PRJEB13537.

Supporting information

S1 Fig. Lifecycle of LTR-RTs. The retrotransposition cycle is composed of different steps. The 5'LTR contains a RNA polymerase II promoter sequence and marks the start of transcription (1) and in contrast the 3' LTR indicates the stop and the polyadenylation signal. LTR transcripts are used both as a matrix for translation (2, 3) and for reverse transcription (6–8). In

the cytoplasm, the polyprotein is self-cleaved into 4 proteins (3): a reverse transcriptase (RT; *green dot*), a RNaseH (*yellow dot*), an aspartic proteinase (AP; *purple dot*) and an integrase (IN; *blue dot*). The interaction between some gag proteins induces the protection of the transcript and of the 4 proteins in a virus-like particle (VLP) (4, 5). The binding of a host tRNA on the primer binding site (PBS) flanking the 3' end of the 5' LTR initiate the reverse transcription of the transcript into DNA via the RT (6). The RNaseH degrades the RNA template (7) and the complementary strand is reverse transcribed (8). The newly synthesized ecDNA copy associated with IN (9) migrates into the nucleus using unknown mechanisms (10). This ecDNA can lead to a new insertion in the host genome (11) or alternately can be recognized by DNA repair mechanisms (either non-homologous end-joining (NHEJ) or homologous recombination) to form eccDNA molecules (12).

(PDF)

S2 Fig. Detection of eccDNAs originating from ribosomal DNA repeats in *A. thaliana*. (A)

Abundance of reads mapping at TE-annotated and rDNA loci in the *A. thaliana* WT mobilome-seq library. Each dot represents the normalized coverage per million mapped reads per all TE-containing (*black circles*) or rDNA containing (*red dots*) 100bp windows obtained after aligning the sequenced reads on the reference genome. (B) Abundance of reads mapping at TE-annotated and rDNA loci in the *A. thaliana* epi12 mobilome-seq library. Legend as in (A).

(PDF)

S3 Fig. Analysis of *A. thaliana* mobilome-seq libraries. (A)

Abundance of reads mapping at TE-annotated loci in the *A. thaliana* WT mobilome-seq library. (B) Statistical analysis of the WT mobilome-seq library presented in (A). (C) Statistical analysis of the epi12 mobilome-seq library presented in Fig 2B. Legend as in Fig 2B.

(PDF)

S4 Fig. Split read (SR) analysis of *A. thaliana* mobilome-seq libraries. (A)

Abundance of SRs mapping at TE-annotated loci in the *A. thaliana* WT mobilome-seq library. (B) Abundance of SRs mapping at TE-annotated loci in the *A. thaliana* epi12 mobilome-seq library. Legend as in Fig 2B.

(PDF)

S5 Fig. EVD forms eccDNAs in *A. thaliana* epi12 line.

Example of a SR identified in the *A. thaliana* epi12 mobilome-seq library spanning the junction of the 2LTR-circle corresponding to EVD aligned with an artificial junction corresponding to the 3' part of the 3'LTR (*red box*) fused to the 5' part of the 5'LTR (*yellow box*).

(PDF)

S6 Fig. Comparison between mobilome and transcriptome data.

A circos plot showing, from outermost to innermost track, scatter plots for (i) split reads coverage per million reads per TE locus (SRs, *red track*), (ii) total coverage per million reads per TE locus (DOC, *blue track*) and (iii) transcriptome coverage at TEs (TR, *green track*). Transcriptome data are presented as the log₂ of fold change in epi12 versus WT at significantly upregulated TE loci [47]. The tracks are scaled separately. The chromosome sizes are indicated in megabase pairs. For mobilome-seq data the names of TE loci that are covered on 90% of their length with both a DOC and SR value >5 reads per million reads are indicated. Data used for this plot are available in S7 Table.

(PDF)

S7 Fig. Analysis of the *O. sativa* WT callus mobilome-seq libraries. (A)

Statistical analysis of the mobilome-seq library presented in Fig 3A. Legend as in Fig 3A. (B) Abundance of SR

mapping at TE-annotated loci in the WT callus mobilome-seq library.
(PDF)

S8 Fig. *Tos17* and *Lullaby* form eccDNAs in rice calli. (A) Example of split read spanning the perfect junction of the 2LTR-circle corresponding to *Tos17*. Legend as in [S5 Fig](#). (B) Example of split read spanning the perfect junction of the 2LTR-circle corresponding to *Lullaby*.
(PDF)

S9 Fig. Analysis of the *O. sativa* seed mobilome-seq libraries. (A) Statistical analysis of the mobilome-seq library presented in [Fig 4A](#). Legend as in [Fig 4A](#). (B) Abundance of split reads mapping at TE-annotated loci in the *O. sativa* WT seed mobilome-seq library.
(PDF)

S10 Fig. *PopRice* retrotransposon forms eccDNAs in rice seeds. (A) Example of a split read spanning the perfect junction of the 2LTR-circle corresponding to *PopRice*. (B) Example of a split read spanning the imperfect junction of the 2LTR-circle corresponding to *PopRice*. A primer binding site (PBS) sequence is highlighted in blue. The PBS is normally found after the 5'LTR in a linear *PopRice*. Legend as in [S5 Fig](#).
(PDF)

S11 Fig. *De novo* assembly of *EVD* eccDNAs. Example of scaffolds obtained after *de novo* assembly of epi12 mobilome-seq library and corresponding to *EVD*. The presence of many scaffolds (and not only one) suggests that *EVD* forms a complex population of circles.
(PDF)

S12 Fig. Detection of *PopRice* eccDNAs by PCR. Circular forms of *PopRice* are specifically detected in the dissected rice endosperm using inverse PCR. Legend as in [Fig 5C](#). PCR using *eEF1 α* primers is used as a loading control.
(PDF)

S13 Fig. Schemes depicting the localization of primers and probes used in this study for the analyzed retrotransposons.
(PDF)

S1 Table. Characteristics of the *A. thaliana* mobilome-seq libraries.
(PDF)

S2 Table. Characteristics of the *O. sativa* mobilome-seq libraries.
(PDF)

S3 Table. Localization of *PopRice* and *Osr4* elements in the *O. sativa* ssp. *japonica* cv. Nipponbare reference genome.
(PDF)

S4 Table. List of primers used in this study.
(PDF)

S5 Table. Significant coverage values. For each library, the number of mapped reads per million per 100bp window is indicated with P-value $< 10^{-3}$ (CHR: chromosome, BP: base pair start coordinate of the 100bp window).
(PDF)

S6 Table. Full mobilome-seq data. For each library, the peaks corresponding to candidate TEs are listed.
(PDF)

S7 Table. Data used for the [S6 Fig](#) circos plot.

(XLS)

S8 Table. Data used for the plot in [Fig 2](#).

(XLS)

Acknowledgments

We would like to thank Jerzy Paszkowski and Jon Reinders for their support in the early development of the mobilome-seq strategy and for epi12 seeds, and our colleagues from the IRD and LGDP laboratories for stimulating discussions. We thank H el ene Vignes at the CIRAD for her technical help with sequencing, Nathalie Picault for kindly providing the rice callus material, Moaine El Baidouri for his help with rice TE annotation, and Pierre Larmande for his help with Circos visualization.

Author Contributions

Conceptualization: MM SL.

Funding acquisition: MM.

Project administration: MM.

Resources: AG OP.

Software: MCC EL SL MM.

Validation: CL EJ SL MM.

Visualization: DRH SL MM.

Writing – original draft: SL MM.

Writing – review & editing: MM SL.

References

1. Lynch VJ, Leclerc RD, May G, Wagner GP. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. *Nat Genet.* 2011; 43: 1154–1159. doi: [10.1038/ng.917](#) PMID: [21946353](#)
2. Rebollo R, Romanish MT, Mager DL. Transposable elements: an abundant and natural source of regulatory sequences for host genes. *Annu Rev Genet.* 2012; 46: 21–42. doi: [10.1146/annurev-genet-110711-155621](#) PMID: [22905872](#)
3. Lisch DR. How important are transposons for plant evolution? *Nat Rev Genet.* 2013; 14: 49–61. doi: [10.1038/nrg3374](#) PMID: [23247435](#)
4. Fedoroff NV. *Plant Transposons and Genome Dynamics in Evolution.* John Wiley & Sons; 2013.
5. Martin A, Troadec C, Boualem A, Rajab M, Fernandez R, Morin H, et al. A transposon-induced epigenetic change leads to sex determination in melon. *Nature.* 2009; 461: 1135–1138. doi: [10.1038/nature08498](#) PMID: [19847267](#)
6. Naito K, Zhang F, Tsukiyama T, Saito H, Hancock CN, Richardson AO, et al. Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. *Nature.* 2009; 461: 1130–1134. doi: [10.1038/nature08479](#) PMID: [19847266](#)
7. Freeling M, Xu J, Woodhouse M, Lisch DR. A Solution to the C-Value Paradox and the Function of Junk DNA: The Genome Balance Hypothesis. *Molecular Plant.* 2015; 8: 899–910. doi: [10.1016/j.molp.2015.02.009](#) PMID: [25743198](#)
8. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet.* 2007; 8: 973–982. doi: [10.1038/nrg2165](#) PMID: [17984973](#)

9. Flavell AJ, Ish-Horowicz D. Extrachromosomal circular copies of the eukaryotic transposable element copia in cultured *Drosophila* cells. *Nature*. 1981; 292: 591–595. PMID: [6265802](#)
10. Flavell AJ, Brierley C. The termini of extrachromosomal linear copia elements. *Nucleic Acids Research*. 1986; 14:3659–3669. PMID: [2423971](#)
11. Kilzer JM, Stracker T, Beitzel B, Meek K, Weitzman M, Bushman FD. Roles of host cell factors in circularization of retroviral dna. *Virology*. 2003; 314: 460–467. PMID: [14517098](#)
12. Li L, Olvera JM, Yoder KE, Mitchell RS, Butler SL, Lieber M, et al. Role of the non-homologous DNA end joining pathway in the early steps of retroviral infection. *EMBO J*. 2001; 20: 3272–3281. doi: [10.1093/emboj/20.12.3272](#) PMID: [11406603](#)
13. Hirochika H, Otsuki H. Extrachromosomal circular forms of the tobacco retrotransposon Tto1. *Gene*. 1995; 165: 229–232. PMID: [8522181](#)
14. Sundaresan V, Freeling M. An extrachromosomal form of the Mu transposons of maize. *Proceedings of the National Academy of Sciences*. 1987; 84: 4924.
15. Gorbunova V, Levy AA. Analysis of extrachromosomal Ac/Ds transposable elements. *Genetics*. 2000; 155: 349–359. PMID: [10790408](#)
16. Siefert JL. Defining the mobilome. *Methods Mol Biol*. 2009; 532: 13–27. doi: [10.1007/978-1-60327-853-9_2](#) PMID: [19271177](#)
17. Moon S, Jung K-H, Lee D-E, Jiang W-Z, Koh HJ, Heu M-H, et al. Identification of active transposon dTok, a member of the hAT family, in rice. *Plant Cell Physiol*. 2006; 47: 1473–1483. doi: [10.1093/pcp/pcl012](#) PMID: [16990289](#)
18. Komatsu M, Chujo A, Nagato Y, Shimamoto K, Kyozuka J. FRIZZY PANICLE is required to prevent the formation of axillary meristems and to establish floral meristem identity in rice spikelets. *Development*. 2003; 130: 3841–3850. PMID: [12835399](#)
19. Jiang N, Bao Z, Zhang X, Hirochika H, Eddy SR, McCouch SR, et al. An active DNA transposon family in rice. *Nature*. 2003; 421: 163–167. doi: [10.1038/nature01214](#) PMID: [12520302](#)
20. Hirochika H, Otsuki H, Yoshikawa M, Otsuki Y, Sugimoto K, Takeda S. Autonomous transposition of the tobacco retrotransposon Tto1 in rice. *Plant Cell*. 1996; 8: 725–734. doi: [10.1105/tpc.8.4.725](#) PMID: [8624443](#)
21. Picault N, Chaparro C, Piegue B, Stenger W, Formey D, Llauro C, et al. Identification of an active LTR retrotransposon in rice. *Plant J*. 2009; 58: 754–765. doi: [10.1111/j.1365-313X.2009.03813.x](#) PMID: [19187041](#)
22. Korb J, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, et al. Paired-end mapping reveals extensive structural variation in the human genome. *Science*. 2007; 318: 420–426. doi: [10.1126/science.1149504](#) PMID: [17901297](#)
23. Sabot F, Picault N, El-Baidouri M, Llauro C, Chaparro C, Piegue B, et al. Transpositional landscape of the rice genome revealed by paired-end mapping of high-throughput re-sequencing data. *Plant J*. 2011; 66: 241–246. doi: [10.1111/j.1365-313X.2011.04492.x](#) PMID: [21219509](#)
24. Hénaff E, Zapata L, Casacuberta JM, Ossowski S. Jitterbug: somatic and germline transposon insertion detection at single-nucleotide resolution. *BMC Genomics*. 2015; 16: 768. doi: [10.1186/s12864-015-1975-5](#) PMID: [26459856](#)
25. Bucher E, Reinders J, Mirouze M. Epigenetic control of transposon transcription and mobility in *Arabidopsis*. *Current Opinion in Plant Biology*. 2012; 15: 503–510. doi: [10.1016/j.pbi.2012.08.006](#) PMID: [22940592](#)
26. Lisch DR, Slotkin RK. Strategies for silencing and escape: the ancient struggle between transposable elements and their hosts. *Int Rev Cell Mol Biol*. 2011; 292: 119–152. doi: [10.1016/B978-0-12-386033-0.00003-7](#) PMID: [22078960](#)
27. Ewing AD. Transposable element detection from whole genome sequence data. *Mob DNA*. 2015; 6: 24. doi: [10.1186/s13100-015-0055-3](#) PMID: [26719777](#)
28. Quadrana L, Bortolini Silveira A, Mayhew GF, Leblanc C, Martienssen RA, Jeddeloh JA, et al. The *Arabidopsis thaliana* mobilome and its impact at the species level. *Elife*. 2016; 5.
29. Upton KR, Gerhardt DJ, Jesuadian JS, Richardson SR, Sánchez-Luque FJ, Bodea GO, et al. Ubiquitous L1 mosaicism in hippocampal neurons. *Cell*. 2015; 161: 228–239. doi: [10.1016/j.cell.2015.03.026](#) PMID: [25860606](#)
30. Baillie JK, Barnett MW, Upton KR, Gerhardt DJ, Richmond TA, De Sapio F, et al. Somatic retrotransposition alters the genetic landscape of the human brain. *Nature*. 2011; 479: 534–537. doi: [10.1038/nature10531](#) PMID: [22037309](#)
31. Flavell AJ, Ish-Horowicz D. The origin of extrachromosomal circular copia elements. *Cell*. 1983; 34: 415–419. PMID: [6616617](#)

32. Bhattacharyya N, Roy P. Extrachromosomal DNA from a dicot plant *Vigna radiata*. *FEBS Letters*. 1986.
33. Cohen S, Lavi S. Induction of circles of heterogeneous sizes in carcinogen-treated cells: two-dimensional gel analysis of circular DNA molecules. *Molecular and Cellular Biology*. 1996; 16: 2002–2014. PMID: [8628266](#)
34. Cohen S, Houben A, Segal D. Extrachromosomal circular DNA derived from tandemly repeated genomic sequences in plants. *Plant J*. 2008; 53: 1027–1034. doi: [10.1111/j.1365-313X.2007.03394.x](#) PMID: [18088310](#)
35. Mukherjee K, Storici F. A mechanism of gene amplification driven by small DNA fragments. *PLoS Genet*. 2012; 8: e1003119. doi: [10.1371/journal.pgen.1003119](#) PMID: [23271978](#)
36. Diaz-Lara A, Gent DH, Martin RR. Identification of Extrachromosomal Circular DNA in Hop via Rolling Circle Amplification. *Cytogenet Genome Res*. 2016; 148: 237–240. doi: [10.1159/000445849](#) PMID: [27160259](#)
37. Gaubatz JW. Extrachromosomal circular DNAs and genomic sequence plasticity in eukaryotic cells. *Mutat Res*. 1990; 237: 271–292. PMID: [2079966](#)
38. Rush MG, Misra R. Extrachromosomal DNA in eucaryotes. *Plasmid*. 1985; 14: 177–191. PMID: [3912782](#)
39. Shibata Y, Kumar P, Layer R, Willcox S, Gagan JR, Griffith JD, et al. Extrachromosomal microDNAs and chromosomal microdeletions in normal tissues. *Science*. 2012; 336: 82–86. doi: [10.1126/science.1213307](#) PMID: [22403181](#)
40. Dillon LW, Kumar P, Shibata Y, Wang Y-H, Willcox S, Griffith JD, et al. Production of Extrachromosomal MicroDNAs Is Linked to Mismatch Repair Pathways and Transcriptional Activity. *Cell Rep*. 2015; 11: 1749–1759. doi: [10.1016/j.celrep.2015.05.020](#) PMID: [26051933](#)
41. Møller HD, Parsons L, Jørgensen TS, Botstein D, Regenberg B. Extrachromosomal circular DNA is common in yeast. *Proceedings of the National Academy of Sciences*. 2015; 112: E3114–22.
42. Møller HD, Larsen CE, Parsons L, Hansen AJ, Regenberg B, Mourier T. Formation of Extrachromosomal Circular DNA from Long Terminal Repeats of Retrotransposons in *Saccharomyces cerevisiae*. *G3 (Bethesda)*. 2015; 6: 453–462.
43. Reinders J, Wulff BBH, Mirouze M, Marí-Ordóñez A, Dapp M, Rozhon W, et al. Compromised stability of DNA methylation and transposon immobilization in mosaic *Arabidopsis* epigenomes. *Genes & Development*. 2009; 23: 939–950.
44. Mirouze M, Reinders J, Bucher E, Nishimura T, Schneeberger K, Ossowski S, et al. Selective epigenetic control of retrotransposition in *Arabidopsis*. *Nature*. 2009; 461: 427–430. doi: [10.1038/nature08328](#) PMID: [19734882](#)
45. Hirochika H. Activation of tobacco retrotransposons during tissue culture. *EMBO J*. 1993; 12: 2521–2528. PMID: [8389699](#)
46. Reinders J, Mirouze M, Nicolet J, Paszkowski J. Parent-of-origin control of transgenerational retrotransposon proliferation in *Arabidopsis*. *Nature Publishing Group*. 2013; 14: 823–828.
47. Mirouze M, Lieberman-Lazarovich M, Aversano R, Bucher E, Nicolet J, Reinders J, et al. Loss of DNA methylation affects the recombination landscape in *Arabidopsis*. *Proceedings of the National Academy of Sciences*. 2012; 109: 5880–5885.
48. Slotkin RK, Vaughn M, Borges F, Tanurdzic M, Becker JD, Feijó JA, et al. Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. *Cell*. 2009; 136: 461–472. doi: [10.1016/j.cell.2008.12.038](#) PMID: [19203581](#)
49. Hsieh T-F, Ibarra CA, Silva P, Zemach A, Eshed-Williams L, Fischer RL, et al. Genome-wide demethylation of *Arabidopsis* endosperm. *Science*. 2009; 324: 1451–1454. doi: [10.1126/science.1172417](#) PMID: [19520962](#)
50. Zemach A, Kim MY, Silva P, Rodrigues JA, Dotson B, Brooks MD, et al. Local DNA hypomethylation activates genes in rice endosperm. *Proceedings of the National Academy of Sciences*. 2010; 107: 18729–18734.
51. Martínez G, Slotkin RK. Developmental relaxation of transposable element silencing in plants: functional or byproduct? *Current Opinion in Plant Biology*. 2012; 15: 496–502. doi: [10.1016/j.pbi.2012.09.001](#) PMID: [23022393](#)
52. Gao L, McCarthy EM, Ganko EW, McDonald JF. Evolutionary history of *Oryza sativa* LTR retrotransposons: a preliminary survey of the rice genome sequences. *BMC Genomics*. 2004; 5: 18. doi: [10.1186/1471-2164-5-18](#) PMID: [15040813](#)
53. Cohen S, Segal D. Extrachromosomal circular DNA in eukaryotes: possible involvement in the plasticity of tandem repeats. *Cytogenet Genome Res*. 2009; 124: 327–338. doi: [10.1159/000218136](#) PMID: [19556784](#)

54. Rodrigues JA, Ruan R, Nishimura T, Sharma MK, Sharma R, Ronald PC, et al. Imprinted expression of genes and small RNA is associated with localized hypomethylation of the maternal genome in rice endosperm. *Proceedings of the National Academy of Sciences*. 2013; 110: 7934–7939.
55. Xing M-Q, Zhang Y-J, Zhou S-R, Hu W-Y, Wu X-T, Ye Y-J, et al. Global Analysis Reveals the Crucial Roles of DNA Methylation during Rice Seed Development. *PLANT PHYSIOLOGY*. 2015; 168: 1417–1432. doi: [10.1104/pp.15.00414](https://doi.org/10.1104/pp.15.00414) PMID: [26145151](https://pubmed.ncbi.nlm.nih.gov/26145151/)
56. Ibarra CA, Feng X, Schoft VK, Hsieh T-F, Uzawa R, Rodrigues JA, et al. Active DNA demethylation in plant companion cells reinforces transposon methylation in gametes. *Science*. 2012; 337: 1360–1364. doi: [10.1126/science.1224839](https://doi.org/10.1126/science.1224839) PMID: [22984074](https://pubmed.ncbi.nlm.nih.gov/22984074/)
57. Pignatta D, Erdmann RM, Scheer E, Picard CL, Bell GW, Gehring M. Natural epigenetic polymorphisms lead to intraspecific variation in Arabidopsis gene imprinting. *Elife*. 2014; 3: e03198. doi: [10.7554/eLife.03198](https://doi.org/10.7554/eLife.03198) PMID: [24994762](https://pubmed.ncbi.nlm.nih.gov/24994762/)
58. Luo M, Taylor JM, Spriggs A, Zhang H, Wu X, Russell S, Singh M, Koltunow A. A genome-wide survey of imprinted genes in rice seeds reveals imprinting primarily occurs in the endosperm. *PLoS Genet*. 2011; 7(6):e1002125. doi: [10.1371/journal.pgen.1002125](https://doi.org/10.1371/journal.pgen.1002125) PMID: [21731498](https://pubmed.ncbi.nlm.nih.gov/21731498/)
59. Cheng C, Tarutani Y, Miyao A, Ito T, Yamazaki M, Sakai H, et al. Loss of function mutations in the rice chromomethylase OsCMT3a cause a burst of transposition. *Plant J*. 2015; 83: 1069–1081. doi: [10.1111/tpj.12952](https://doi.org/10.1111/tpj.12952) PMID: [26243209](https://pubmed.ncbi.nlm.nih.gov/26243209/)
60. Ito H, Gaubert H, Bucher E, Mirouze M, Vaillant I, Paszkowski J. An siRNA pathway prevents transgenerational retrotransposition in plants subjected to stress. *Nature*. 2011; 472: 115–119. doi: [10.1038/nature09861](https://doi.org/10.1038/nature09861) PMID: [21399627](https://pubmed.ncbi.nlm.nih.gov/21399627/)
61. Fultz D, Choudury SG, Slotkin RK. Silencing of active transposable elements in plants. *Current Opinion in Plant Biology*. 2015; 27: 67–76. doi: [10.1016/j.pbi.2015.05.027](https://doi.org/10.1016/j.pbi.2015.05.027) PMID: [26164237](https://pubmed.ncbi.nlm.nih.gov/26164237/)
62. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Meth*. 2012; 9: 357–359.
63. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. *Genome Research*. 2009; 19: 1639–1645. doi: [10.1101/gr.092759.109](https://doi.org/10.1101/gr.092759.109) PMID: [19541911](https://pubmed.ncbi.nlm.nih.gov/19541911/)
64. Coil D, Jospin G, Darling AE. A5-miseq: an updated pipeline to assemble microbial genomes from Illumina MiSeq data. *Bioinformatics*. 2015; 31: 587–589. doi: [10.1093/bioinformatics/btu661](https://doi.org/10.1093/bioinformatics/btu661) PMID: [25338718](https://pubmed.ncbi.nlm.nih.gov/25338718/)
65. Hoffmann S, Otto C, Doose G, Tanzer A, Langenberger D, Christ S, et al. A multi-split mapping algorithm for circular RNA, splicing, trans-splicing and fusion detection. *Genome Biol*. 2014; 15: R34. doi: [10.1186/gb-2014-15-2-r34](https://doi.org/10.1186/gb-2014-15-2-r34) PMID: [24512684](https://pubmed.ncbi.nlm.nih.gov/24512684/)
66. Clarke JD. Cetyltrimethyl ammonium bromide (CTAB) DNA miniprep for plant DNA isolation. *Cold Spring Harbor Protocols*. 2009; 2009: pdb.prot5177.
67. Xu H, Bao JD, Dai JS, Li Y, Zhu Y. Genome-wide identification of new reference genes for qRT-PCR normalization under high temperature stress in rice endosperm. *PLOS ONE*. 2015; 10(11): e0142015. doi: [10.1371/journal.pone.0142015](https://doi.org/10.1371/journal.pone.0142015) PMID: [26555942](https://pubmed.ncbi.nlm.nih.gov/26555942/)

Le séquençage du mobilome a permis l'identification d'ET connus pour être actifs, *EVD* et *Tos17* respectivement chez *Arabidopsis* et chez le riz dans des tissus où l'épigénome est déstabilisé. De plus, le séquençage du mobilome à différents stades de développement chez le riz a mis en évidence l'activité d'un nouvel élément que nous avons appelé *PopRice* dans les grains. Les analyses par Southern blot et par PCR inverse montrent que l'ADNec de *PopRice* est spécifiquement détecté dans les tissus du grain (Figure 13a,b) dès le développement de l'embryon (le stade laiteux : 3 à 5 jours après la pollinisation). La dissection des différents tissus du grain confirme également la présence spécifique de ces cercles dans l'albumen, tissu nourricier du grain (Figure 13b). Par analyse RT-qPCR nous avons montré que *PopRice* et tous les éléments de la famille sont transcrits spécifiquement dans les grains et non dans les feuilles et dans les fleurs (Figure 13d,e).

Dans la littérature, l'hypométhylation de l'albumen de riz par rapport à l'embryon est très bien détaillée, une de ses conséquences étant la réactivation transcriptionnelle massive des ET (Zemach et al. 2010; Xing et al. 2015). Ce changement épigénétique pourrait suggérer également une activité transpositionnelle des ET importante dans ce tissu. Or, nos données de mobilome-seq ont montré que seul l'ADNec de *PopRice* était détecté dans les grains de riz. L'activité de *PopRice* est-elle une conséquence directe du relâchement épigénétique observé dans l'albumen ? Comment alors expliquer que contrairement aux autres familles d'ET, *PopRice* passe la barrière de la transcription et de la transcription réverse ? Quels sont les mécanismes de régulation de *PopRice* ? Quel impact a l'activité de *PopRice* sur le génome de l'albumen de riz ? La suite de nos travaux, présentés ci-dessous dans la deuxième partie de ce chapitre, a consisté à caractériser l'activité de cet élément d'une part au sein du genre *Oryza* puis d'autre part, au sein des céréales. Certains des résultats clés de l'article de *PLoS Genetics* ont pu être étoffés depuis la publication et ils sont repris dans cette partie afin de faciliter la lecture.

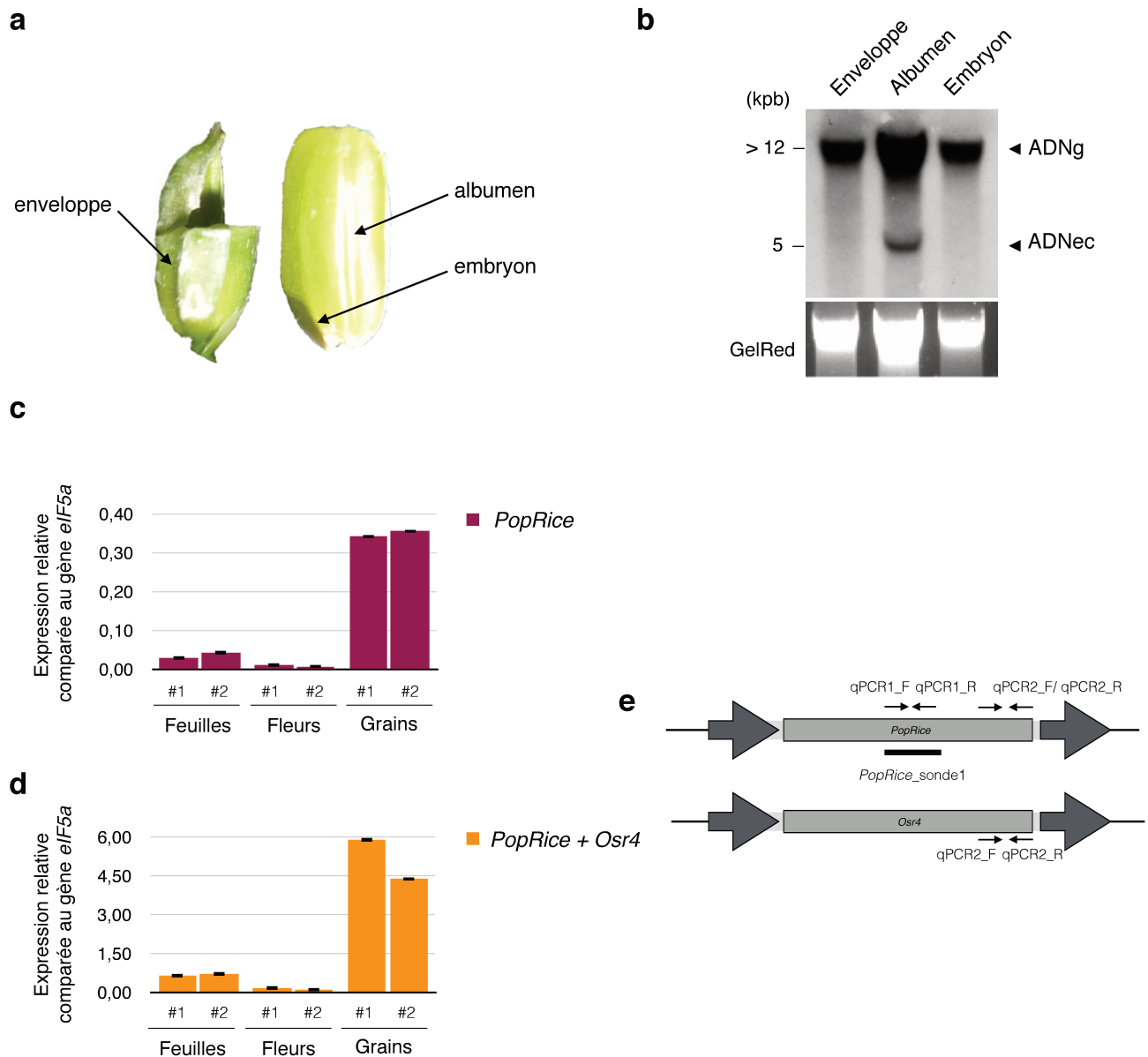


Figure 13. Accumulation du rétrotransposon à LTR *PopRice* sous forme extrachromosomique dans l'albumen du grain de riz. (a) Détail de la structure du grain de riz. L'enveloppe externe protège le grain et assure une barrière entre la graine et l'environnement. L'albumen, constitué de granules d'amidon, se développe jusqu'au stade mature du grain et fournit les nutriments nécessaires à la germination de la graine et au développement de la plantule à partir de l'embryon. (b) Analyse par Southern blot en utilisant de l'ADN génomique (ADNg) non digéré extrait à partir des tissus disséqués de grains de riz matures et hybridés avec la sonde *PopRice_sonde1* (voir (e) et annexes) spécifique des éléments de la famille *PopRice*. Cette analyse montre que les copies intégrées de *PopRice* (ADNg) sont détectées dans tous les tissus du grain contrairement aux formes extrachromosomiques (ADNec) qui ne sont détectées que dans l'albumen du grain. Une image du GelRed est montrée comme contrôle de chargement. (c) Analyse par RT-qPCR des transcrits de *PopRice* dans les feuilles, fleurs et grains de riz matures, en utilisant un couple d'amorces spécifique de la sous-famille *PopRice*. Le niveau d'expression relative a été calculé à partir de l'expression du gène de référence *eIF-5a* (Xu et al., 2015). Les barres d'erreurs ont été calculées à partir de trois réplicats techniques et deux réplicats biologiques sont montrées pour chaque tissu. (d) Analyse par RT-qPCR des transcrits de *PopRice* et *Osr4* dans les feuilles, fleurs et grains de riz matures en utilisant un couple d'amorces spécifique de la famille *Osr4* (incluant *PopRice*). (e) Schéma de la localisation des amorces utilisées pour les analyses b, c et d. Ces figures sont issues de Lanciano et al., 2017.

Partie 2. Caractérisation de l'activité de *PopRice* au sein des grains de céréales.

Introduction

La structure et le développement du grain chez les plantes monocotylédones sont très différents des dicotylédones (Zhou et *al.* 2013). Chez le riz, plante modèle des monocotylédones, les trois tissus principaux qui constituent la graine sont l'embryon, l'albumen et l'enveloppe (Figure 13a). L'embryon résulte de la fécondation entre l'oosphère (1n) du gamétophyte femelle et une cellule spermatique (1n) du gamétophyte mâle alors que l'albumen est issu de la fusion entre la deuxième cellule spermatique du gamétophyte mâle et une des deux cellules centrales (2n) du gamétophyte femelle. L'albumen (3n) est un tissu qui n'est pas transmis à la génération suivante. Le développement du grain se déroule en trois phases majeures : (1) développement de l'embryon, (2) remplissage de l'albumen, (3) déshydratation de la graine. Chez *Arabidopsis* l'albumen est consommé par l'embryon qui se développe, contrairement aux monocotylédones où l'albumen accumule continuellement des nutriments et joue un rôle primordial dans la germination du grain et dans le développement de la plantule en fournissant les nutriments et l'énergie nécessaires. Ces différences majeures expliquent indirectement pourquoi les monocotylédones, et plus précisément les grains de céréales, fournissent les principales ressources énergétiques de l'alimentation animale.

Les différents tissus du grain subissent des changements physiologiques importants ainsi que des changements au niveau épigénétique au cours de la maturation du grain. Les analyses transcriptionnelles effectuées sur les différents tissus du grain à différents stades de développement ont révélé que les facteurs de transcription exprimés dans l'albumen étaient majoritairement associés avec des gènes impliqués dans la régulation du stockage des nutriments et dans la synthèse et la réponse à l'hormone acide abscissique (ABA) (Xue et *al.* 2012). Les hormones, et notamment l'ABA, jouent un rôle central dans la régulation du développement du grain et dans la réponse au stress, ainsi que dans le contrôle de la dormance (arrêt de la croissance et du développement du grain jusqu'à l'arrivée de conditions optimales pour germer) (Liu et *al.* 2014). La méthylation de l'ADN joue également un rôle important dans la régulation des gènes clés du développement du grain (Zemach et *al.* 2010). Comme précédemment présenté dans le chapitre 1 de l'Introduction (voir page 14), l'ADN de l'albumen de riz des deux sous-espèces *Oryza sativa* ssp. *japonica* et *indica* est très fortement déméthylé

Tableau 1. Caractéristiques des éléments de la sous-famille *PopRice*. Les coordonnées réfèrent au génome de référence de Nipponbare (IRGSP1.0). La couverture moyenne de séquençage du mobilome de grains pour chacun des éléments est indiquée en lectures par million (rpm) et est représentée par un code couleur (rouge, les plus représentés, blanc les moins représentés). Cette couverture est calculée à partir de deux réplicats biologiques. La taille des deux LTR est indiquée en paires de bases ainsi que leur identité. L'analyse des polymorphismes d'insertions et des délétions entre les deux sous-espèces *O. sativa* ssp. *japonica* et *indica* a été réalisée par E. Lasserre. La localisation dans un intron de gène est indiquée, ainsi que la fonction du gène le cas échéant. Les domaines protéiques VHS et GAT sont impliqués dans le transport membranaire et le gène OsCBL1 est une protéine calcineurine qui se lie au calcium.

ID	coordonnées (IRGSP1.0)	couverture moyenne mobilome-seq (rpm)	taille 5'LTR (bp)	taille 3'LTR (bp)	% identité LTR	polymorphe japonica vs. indica	intron gène	fonction du gène
<i>PopRice_1</i>	chr01:4776239-4781940	346	362	362	100	oui	LOC_Os01g09384	non déterminée
<i>PopRice_2</i>	chr02:11897051-11902751	609	362	362	100	oui	non	
<i>PopRice_3</i>	chr02:34010137-34015809	602	334	362	92	oui	non	
<i>PopRice_4</i>	chr03:1910824-1916518	197	362	362	100	non	LOC_Os03g04169	non déterminée
<i>PopRice_5</i>	chr04:21858331-21863990	345	368	364	98	oui	non	
<i>PopRice_6</i>	chr04:29764342-29769992	50	393	360	72	oui	non	
<i>PopRice_7</i>	chr04:31205979-31211679	509	362	362	100	oui	LOC_Os04g52479	non déterminée
<i>PopRice_8</i>	chr06:2568077-2573729	49	332	350	94	oui	non	
<i>PopRice_9</i>	chr06:3207031-3212695	26	350	350	100	oui	non	
<i>PopRice_10</i>	chr07:11227404-11233065	82	350	348	91	non	non	
<i>PopRice_11</i>	chr08:9051840-9057543	440	363	363	100	oui	non	
<i>PopRice_12</i>	chr09:1229148-1234789	157	334	334	100	oui	LOC_Os09g02729	phospholipase
<i>PopRice_13</i>	chr09:8572145-8577847	246	363	363	100	oui	non	
<i>PopRice_14</i>	chr10:13174222-13179845	30	352	328	93	non	LOC_Os10g25487	résistance aux maladies
<i>PopRice_15</i>	chr10:22300481-22306180	298	363	363	99,7	non	LOC_Os10g41510	similaire à la calcineurine OsCBL1
<i>PopRice_16</i>	chr11:4997937-5003672	32	405	364	89	oui	LOC_Os11g09329	domaines protéiques VHS et GAT
<i>PopRice_17</i>	chr11:26689013-26694658	58	301	301	99	non	non	

dans les trois contextes de méthylation CG, CHG et CHH (Zemach et al. 2010; Xing et al. 2015). Cette hypométhylation semble également responsable de l'activation transcriptionnelle des ET dans l'albumen (Zemach et al. 2010) mais ne semble pas suffisante pour activer transpositionnellement les ET. En effet nos données de mobilome ont montré que seul le rétrotransposon *PopRice* semble avoir la possibilité de former des ADNec dans l'albumen du riz.

Ce premier résultat suscite des interrogations sur les mécanismes génétiques et épigénétiques qui régulent *PopRice*. Si l'activité de *PopRice* est le simple résultat d'un relâchement épigénétique contrôlé par la plante, d'autres familles d'ET devraient répondre de la même manière. Or, nous avons montré que ce n'était pas le cas. La singularité de *PopRice* pourrait indiquer que l'activité de *PopRice* a un impact sur le développement ou la germination du grain et a donc été potentiellement sélectionnée au cours de l'évolution. Si notre hypothèse se révèle juste, l'activité de *PopRice* pourrait également être conservée dans les grains d'autres espèces de céréales. En effet, le développement du grain chez les monocotylédones est un processus relativement conservé (Evers et Millar 2002). J'ai ainsi consacré une partie de ma thèse à analyser les séquences régulatrices de *PopRice* chez le riz puis à caractériser l'activité de *PopRice* dans les grains de plusieurs espèces de céréales.

Résultats

2.1 Régulations génétiques et épigénétiques de *PopRice*

PopRice est une sous-famille de rétrotransposons appartenant à la grande famille d'*Osr4* (*O. sativa* LTR retrotransposon 4) (McCarthy et al. 2002) et qui est composée de 17 éléments (Tableau 1). Les 17 copies intégrées dans le génome de référence ont en moyenne entre 85% et 99% d'identité au niveau nucléotidique. Les résultats de séquençage du mobilome ont montré que certaines copies de *PopRice* étaient plus couvertes que d'autres ce qui suggère que toutes les copies ne feraient pas de cercles et que seules certaines copies seraient actives dans les grains (Tableau 1). Par ailleurs, les autres membres de la famille d'*Osr4* ne sont pas présents ou couverts dans le mobilome du grain et ne semblent donc pas actifs.

Huit copies de *PopRice* ont leurs 2 LTR identiques à 100% (Tableau 1). Comme précédemment exposé dans le chapitre 1 de l'Introduction (voir page 41), le pourcentage d'identité des LTR est un témoin de l'âge de l'insertion : plus les LTR sont identiques, plus l'insertion est récente et il apparaît donc que la sous-famille *PopRice* est récente. Les LTR d'un rétrotransposon ont

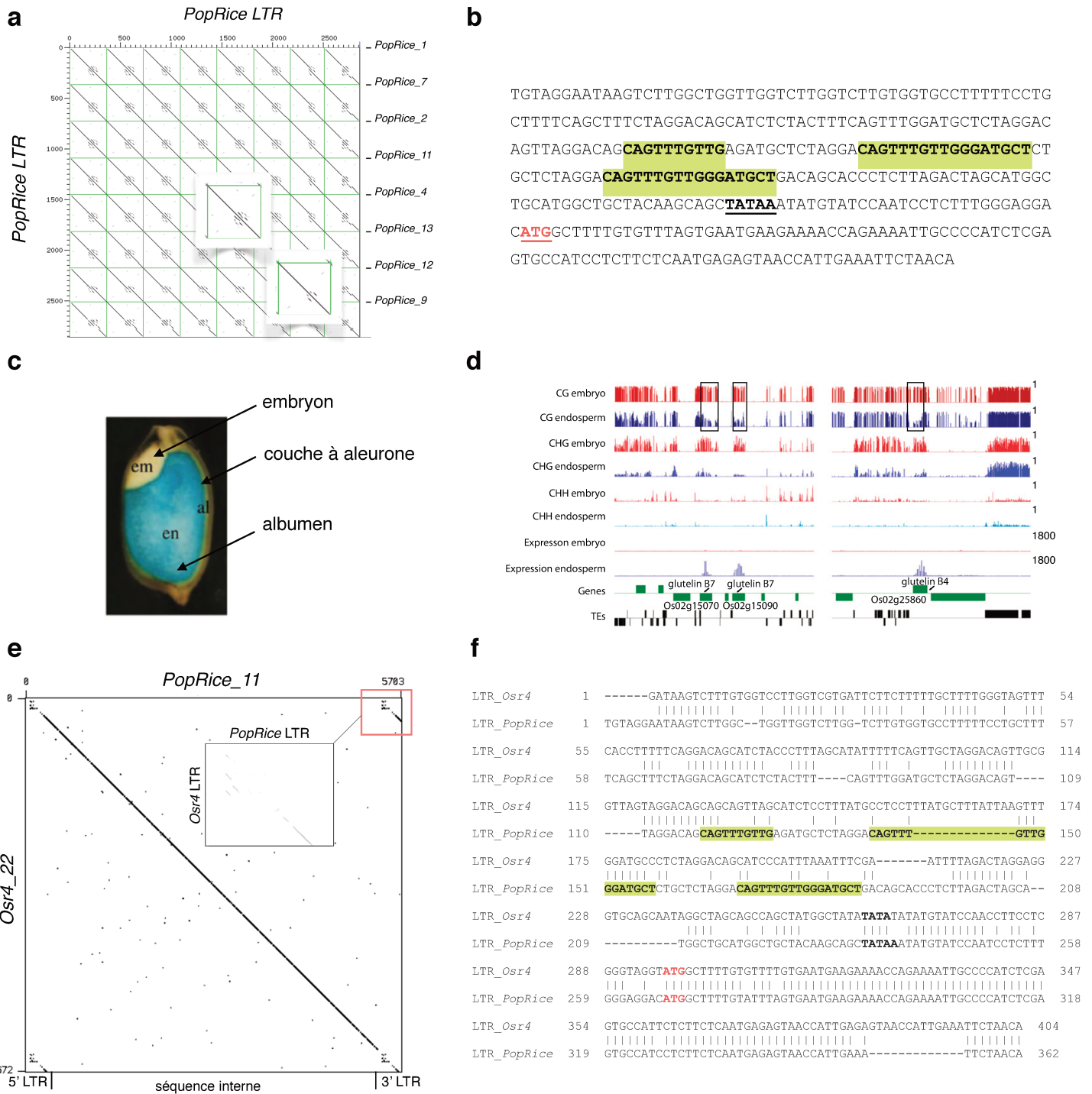


Figure 14. Présence de motifs répétés dans le promoteur de *PopRice*. (a) Analyse par dot-plot de huit LTR de *PopRice*. La séquence nucléotidique de chaque LTR de *PopRice* est graphiquement comparée à elle-même et aux sept autres (logiciel gepard). La ligne principale correspond à la séquence alignée contre elle-même. Les répétitions en tandem sont représentées comme des lignes parallèles à la ligne principale. Ce dot-plot montre que *PopRice* comporte des séquences répétées en tandem dans ses LTR. (b) Séquence nucléotidique du LTR de *PopRice*. Les motifs répétés du LTR de *PopRice* sont surlignés en vert, la TATA box est représentée en caractère gras et le codon *start* en rouge. (c) Analyse du profil d'expression du gène GUS sous l'activité du promoteur du gène *GluC* (gène de la famille des glutélines chez le riz; Qu et al. 2008). (d) Niveaux de méthylation et d'expression de trois gènes glutélines dans l'embryon et dans l'albumen de riz (chr02:8425000-8495000) (Figure de Zemach et al. 2010). Les encadrés soulignent les zones déméthylées dans l'albumen par rapport à l'embryon, en contexte CG. (e) La séquence nucléotidique de *PopRice_11* est comparée à la séquence nucléotidique de *Osr4_22*. Les séquences internes de *PopRice* et de *Osr4* sont très fortement conservées (96,4% d'identité) contrairement aux LTR de *PopRice* et de *Osr4* (59,3%). Un agrandissement de l'alignement des séquences au niveau du LTR 3' de *PopRice* et du LTR 5' de *Osr4*, encadré en rouge, est montré. (f) L'analyse de l'alignement du LTR de *PopRice* contre le LTR de *Osr4* montre que les motifs répétés du LTR de *PopRice*, surlignés en vert, ne sont pas conservés dans le LTR de *Osr4*.

également un rôle de promoteur et l'analyse de ces séquences promotrices chez *PopRice* a mis en évidence la présence de boîtes répétées (Figure 14a,b). Parmi ces boîtes, nous avons identifié le motif CAGTTTGTG qui est une séquence *cis* régulatrice appelée g-box présente dans les promoteurs des gènes de la glutéline, gènes codant pour des protéines de stockage, déméthylés dans l'albumen (Figure 14d) (Zemach et al. 2010) et exprimés spécifiquement dans l'albumen (Figure 14c) (Qu et al. 2008)). Par ailleurs, les motifs identifiés dans le LTR de *PopRice* sont absents dans le LTR des autres membres d'*Osr4* (Figure 14e,f). Les séquences internes de *PopRice* et *Osr4* sont très conservées (96% d'identité) contrairement aux séquences des LTR (59%) (Figure 14e,f). Seuls quelques exemples d'ET sont connus pour avoir acquis des séquences régulatrices bien identifiées (voir page 16). La présence de ces séquences *cis* régulatrices répétées dans le promoteur de *PopRice* pourrait expliquer la spécificité tissulaire de l'activité de *PopRice* et suggérerait donc une régulation génétique de *PopRice*. Or, précédemment par analyses RT-qPCR nous avons montré que *PopRice* et *Osr4* étaient fortement exprimés dans les grains comparés aux feuilles et aux fleurs (Figure 13c,d). Par conséquent la présence des motifs de type g-box chez *PopRice* ne semble pas suffisante pour expliquer les différences de représentation dans le mobilome observée entre *PopRice* et *Osr4*.

Comme précédemment mentionné, l'albumen de riz est fortement hypométhylé par rapport à l'embryon (Zemach et al. 2010; Xing et al. 2015). De plus, le reséquençage du mutant *oscm3* affecté dans une méthyltransférase à ADN chez *Oryza sativa* (Cheng et al. 2015) a montré une rétrotransposition de deux éléments de la famille *Osr4* (*PopRice_6* et *PopRice_9* selon notre nomenclature). Ces données suggèrent que *PopRice* est sous contrôle épigénétique. Pour confirmer cette hypothèse, le niveau de méthylation de *PopRice* et de *Osr4* dans les feuilles, les grains (albumen et embryon compris), ainsi que dans l'albumen seul a été évalué à l'aide de l'enzyme de restriction *HpaII* (Figure 15a,b). *HpaII* est une enzyme sensible à l'état de méthylation du site de restriction, autrement dit *HpaII* coupe l'ADN (au site CCGG) seulement si le motif CG est déméthylé. Ces résultats montrent que seulement une partie de l'ADN a été digérée dans les feuilles et dans les grains, suggérant que dans ces deux échantillons, certains éléments sont déméthylés et d'autres non. Cependant la digestion de l'ADN dans l'albumen seul est totale et indique donc que les éléments de *PopRice* et *Osr4* sont totalement déméthylés dans ce tissu. Cette technique étant peu résolutive, l'analyse *in silico* des données de séquençage bisulfite publiées par Zemach et al. (2010) a été effectuée. Néanmoins, notre analyse a montré que la couverture de séquençage de ce jeu de données était insuffisante pour conclure. En effet, par exemple sur le chromosome 2, 70% des cytosines sont couvertes dans l'albumen contre

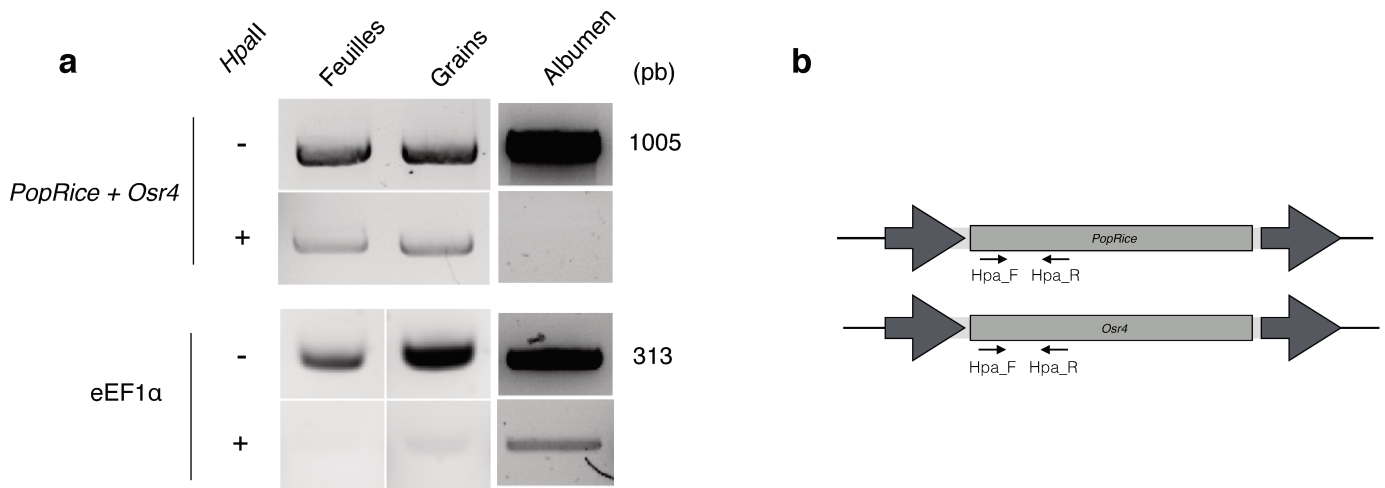


Figure 15. Régulation épigénétique de *PopRice* et *Osr4*. (a) Niveau de méthylation de *PopRice* et de *Osr4* dans les feuilles, les grains et l'albumen après digestion de l'ADN génomique par l'enzyme de restriction sensible à la méthylation *HpaII*. eEF1α est montré comme contrôle de digestion non méthyliée. (b) Schéma de la localisation des amorces utilisées pour les analyses en (a) (voir Annexes).

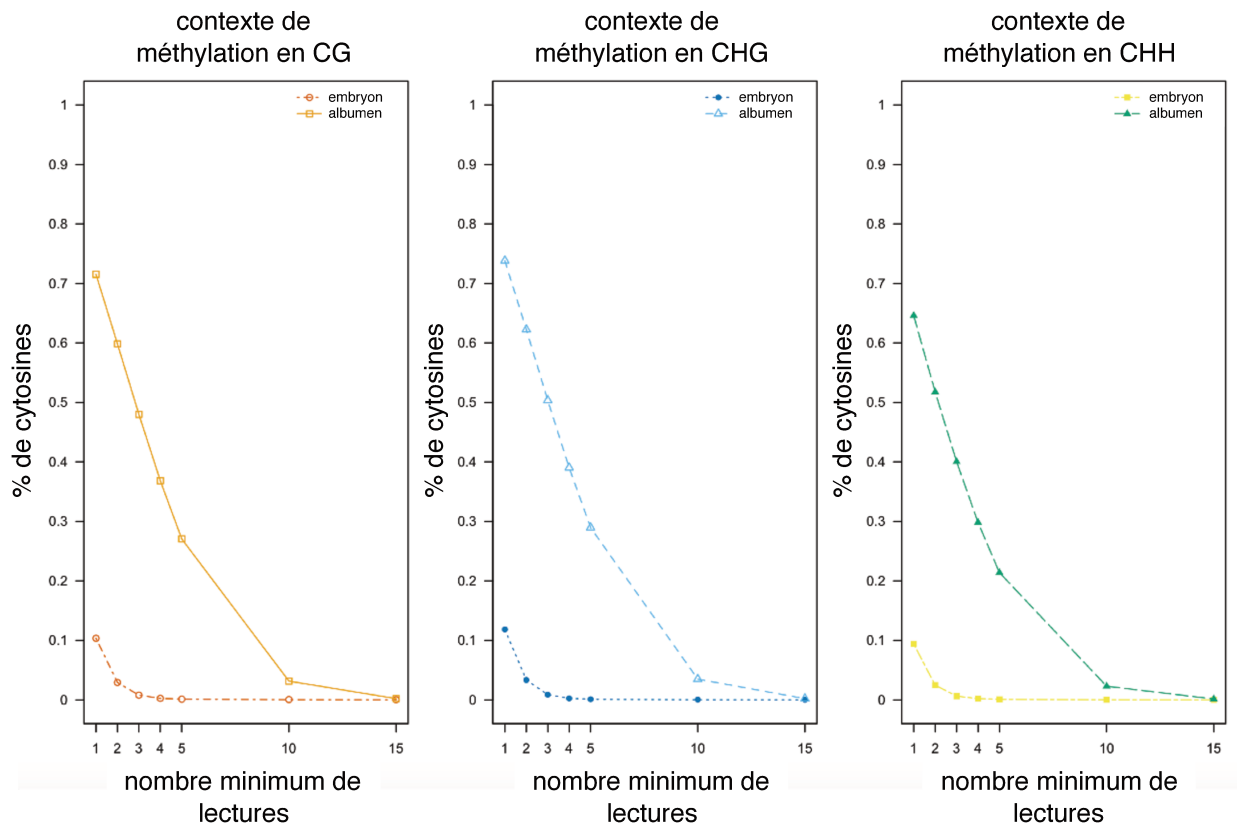


Figure 16. Analysis *in silico* de données de méthylation chez l'embryon et l'albumen de riz. Couverture de séquençage de chaque cytosine du chromosome 2 dans les trois contextes de méthylation CG, CHG et CHH (où H est A, T ou C) des données bisulfite de Zemach et al., (2010). Figure réalisée à partir du package R DMRcaller. Ces plots montrent que pour les trois contextes par exemple, 70% des cytosines sont couvertes et seulement 30% ont 5 lectures ou plus.

seulement 10% dans l'embryon (Figure 16). De plus, dans l'albumen, seules 30% des cytosines sont soutenues par au moins 5 lectures or un minimum de 10 lectures est nécessaire pour obtenir des résultats statistiquement significatifs (Becker et *al.* 2011). Finalement, ces résultats d'expression et de méthylation confirment l'activité de *PopRice* dans l'albumen mais en revanche ils ne permettent pas d'expliquer les différences d'activité au sein de la grande famille d'*Osr4*.

2.2 Activité transpositionnelle de *PopRice*

Les différentes expériences menées jusqu'ici ont montré que dans l'albumen de riz, les éléments de *PopRice* étaient transcrits, reverse-transcrits et présents sous forme extrachromosomique (Figure 13). Si le séquençage du mobilome permet d'identifier les ET actifs, il ne permet pas en revanche de savoir si de l'ADNc provenant de l'ET actif se ré-insère dans le génome et crée une néo-insertion. Afin d'évaluer l'impact de l'activité de *PopRice* dans le génome de l'albumen, les grains de riz de la variété Nipponbare pour laquelle nous disposons d'un génome de référence de qualité, ont été reséquencés. Les grains ont été disséqués sous loupe binoculaire, l'ADN d'albumen a été extrait et reséquencé par la technologie Illumina (voir page 25) avec deux profondeurs différentes, un séquençage léger (environ 10X) puis un séquençage plus profond (environ 100X) (Figure 18a). Les données de séquençage ont été analysées à partir d'un pipeline développé par M-C Carpentier (Figure 17 et Matériel et Méthodes). En résumé, les lectures pairées de 250 paires de bases (bp) ont été alignées sur la séquence de *PopRice*. Les lectures non-alignées des paires alignées/non-alignées ont ensuite été alignées par BLAST sur le génome de référence (IRGSP1.0) afin de localiser la région d'insertion de l'élément. Le nombre de lectures qui soutient chaque néo-insertion détectée a été comptabilisé mais aucun seuil de lecture minimum n'a été établi. Chaque néo-insertion identifiée a ensuite été validée par dot-plot et par PCR.

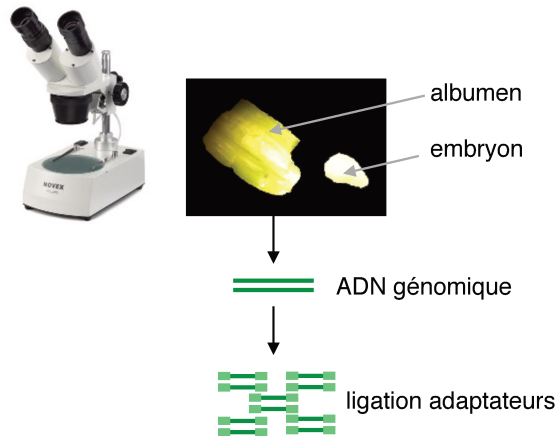
La couverture moyenne de séquençage a été calculée en prenant comme référence la couverture de 72 séquences uniques réparties sur l'ensemble du génome (voir Matériel et Méthodes page 74; Tableau S2 en Annexes) et varie entre 23X +/- 9 pour le séquençage en 10X et 82X +/- 27 pour le séquençage en 100X (Figure 18a). La variation de la couverture moyenne de séquençage s'explique par le fait que certaines régions du génome sont plus difficilement séquencées que d'autres. Or le nombre de néo-insertions détectées par le pipeline varie en fonction de la

a

① dissection du grain sous loupe binoculaire

② extraction d'ADN au CTAB

③ préparation des banques Illumina (Novogene Co.)



b

④ lectures de séquençage paires (2x250bp)

⑤ alignement des lectures contre la séquence de *PopRice*

⑥ BLAST contre le génome de référence

⑦ nombre de lectures qui soutiennent la néo-insertion

⑧ analyse de la néo-insertion détectée par dot-plot

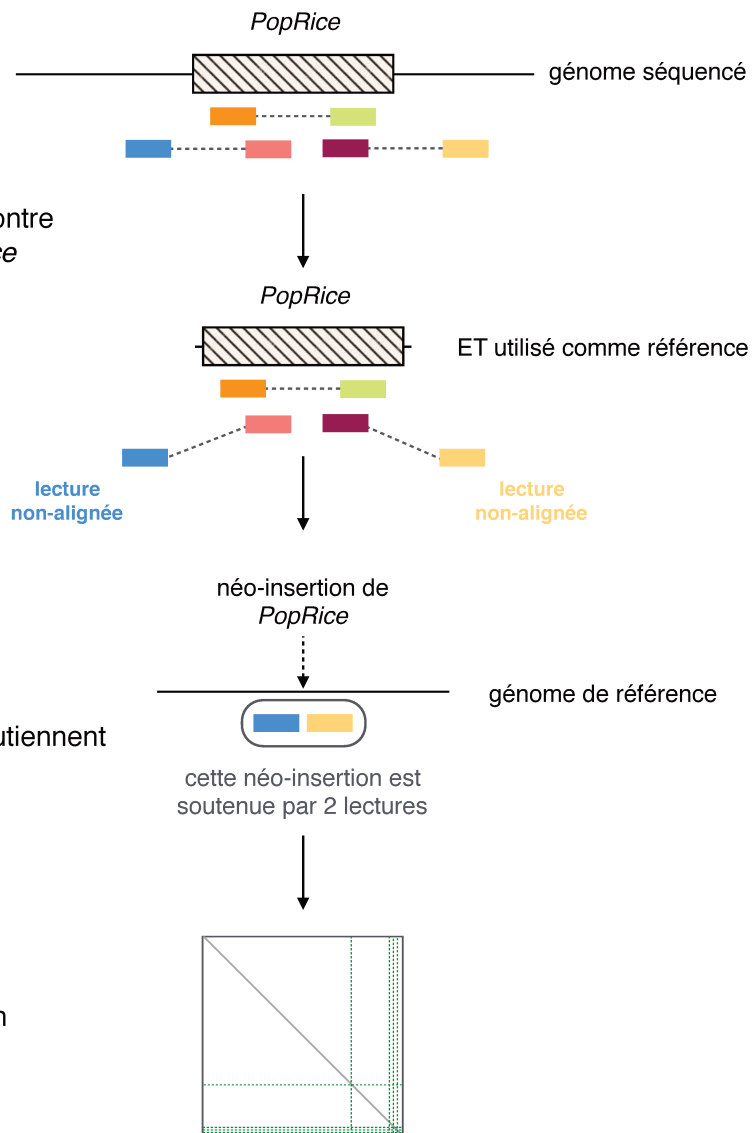


Figure 17. Expérience de reséquençage profond du génome de l'albumen de riz pour analyser l'activité transpositionnelle de *PopRice*. (a) Dissection et préparation de l'ADN génomique à séquencer. (b) Pipeline d'analyse développé par M-C Carpentier et utilisé pour l'analyse du reséquençage profond.

profondeur de séquençage. En outre le pipeline utilisé pour détecter les insertions de *PopRice* ne permet pas de localiser une copie insérée dans une région répétée du génome (voir Matériel et Méthodes page 74). C'est pourquoi, sur les 47 éléments endogènes d'*Osr4* (copies de *PopRice* incluses) seulement 41 insertions ont été détectées pour le séquençage en 100X. L'avantage de ce séquençage à double profondeur est de pouvoir évaluer la fréquence des néo-insertions de *PopRice*. Autrement dit, si les néo-insertions de *PopRice* sont un événement rare dans le génome, la probabilité de détecter des néo-insertions avec une couverture de séquençage basse est faible et plus la couverture de séquençage est importante, plus les probabilités de trouver des néo-insertions augmentent. En revanche, si ces néo-insertions sont très fréquentes dans le génome, même avec une faible couverture de séquençage, les probabilités de trouver des néo-insertions sont bonnes. Notre analyse montre qu'à partir d'un séquençage d'environ 23X, 16 néo-insertions de *PopRice* ont été identifiées par le pipeline contre 60 néo-insertions pour le séquençage en 100X. En d'autres termes, ce résultat suggère que la fréquence des insertions de *PopRice* dans l'albumen semble importante.

Dans le but de valider ces néo-insertions, les séquences de la région d'insertion, de l'ET et des lectures pairées qui couvrent la néo-insertion sont comparées entre elles par dot-plot. Pour qu'une néo-insertion soit validée, la première lecture d'une paire doit aligner sur la région d'insertion (qui ne présente aucune homologie avec l'ET) alors que la seconde lecture doit aligner sur l'ET. C'est le cas pour les insertions 8 et 11 (Figure 18b,c).

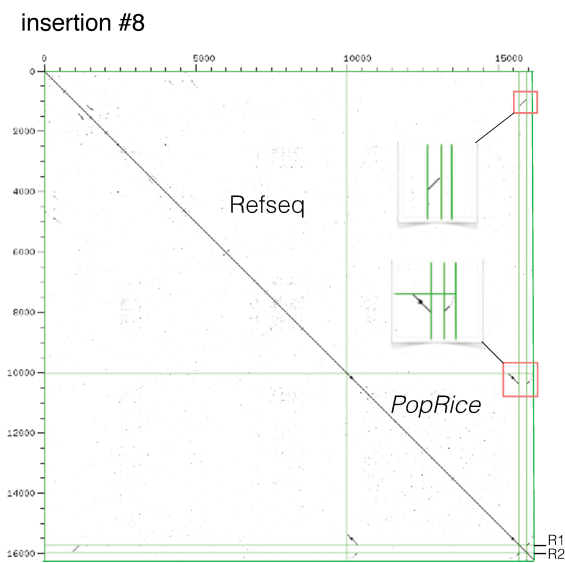
Chacune des insertions identifiées est soutenue par une seule lecture. Or l'activité de *PopRice* dans l'albumen étant somatique, chaque néo-insertion de *PopRice* est unique et peut être présente dans un nombre limité de cellules. Ceci peut expliquer la faible couverture de séquençage des néo-insertions et illustre la difficulté de détecter des insertions somatiques à partir des données de séquençage Illumina. Nous avons ensuite tenté de valider ces deux insertions par PCR. Actuellement, nous avons testé 5 néo-insertions par PCR sans résultat positif. Étant donné la difficulté de valider par PCR des insertions somatiques, d'autres approches devront être envisagées (capture par exemple) afin de conclure sur le niveau de transposition de *PopRice*.

Nos données de mobilome suggéraient que seul *PopRice* était fortement actif dans l'albumen. Pour tester si d'autres familles étaient actives d'après nos données de reséquençage, deux familles de rétrotransposons à LTR, *Tos17* et *Houba*, ont été étudiées pour leurs néo-insertions. Le pipeline a été appliqué sur ces deux familles (Figure 19a). *Tos17* est présent en 2 copies

a

	Insertions de <i>PopRice</i>	
	Albumen (10X)	Albumen (100X)
Couverture des régions uniques	23X ± 9	82X ± 27
Copies natives détectées (<i>PopRice</i> + <i>Osr4</i>)	39/47	41/47
Néo-insertions détectées par le pipeline	22	60
Néo-insertions validées par <i>dotplot</i>	16	54

b



c

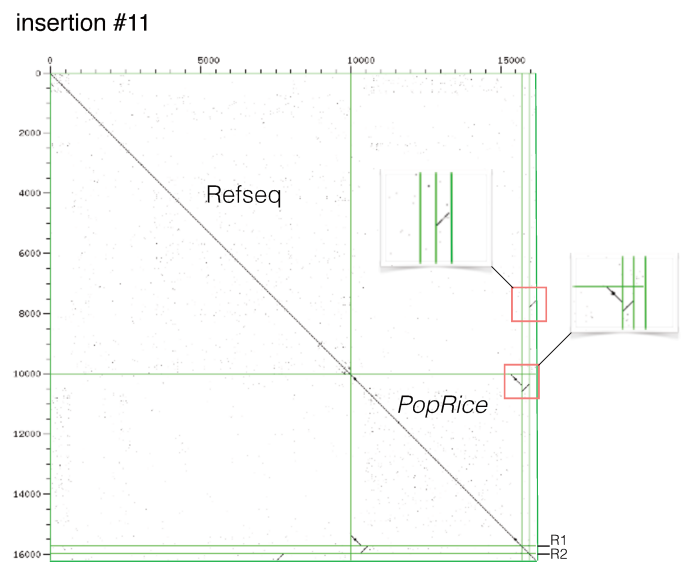


Figure 18. Activité transpositionnelle de *PopRice* dans l'albumen de riz. (a) Tableau des résultats du reséquençage. (b) et (c) Validations par dot-plot des néo-insertions #8 et #11 détectées par le re-séquençage de l'albumen de riz. Les séquences de la région d'insertion, de l'ET et des lectures paires qui couvrent la néo-insertion sont comparées entre elles. Pour qu'une néo-insertion soit validée, la première lecture d'une paire doit aligner sur la région d'insertion (Refseq, qui ne présente aucune homologie avec l'ET) alors que la seconde lecture doit aligner sur l'ET. Ces dot-plots montrent que l'insertion 8 et l'insertion 11 sont validées.

a

	<i>Tos17</i>	<i>Houba</i>
Copies natives	2	650 copies > 250nt 145 copies > 5kb
Néo-insertions détectées par le pipeline	1	504 à tester (copies natives comprises)
Néo-insertions testées par dotplot	1	1
Neo-insertions validées par dotplot	1	1

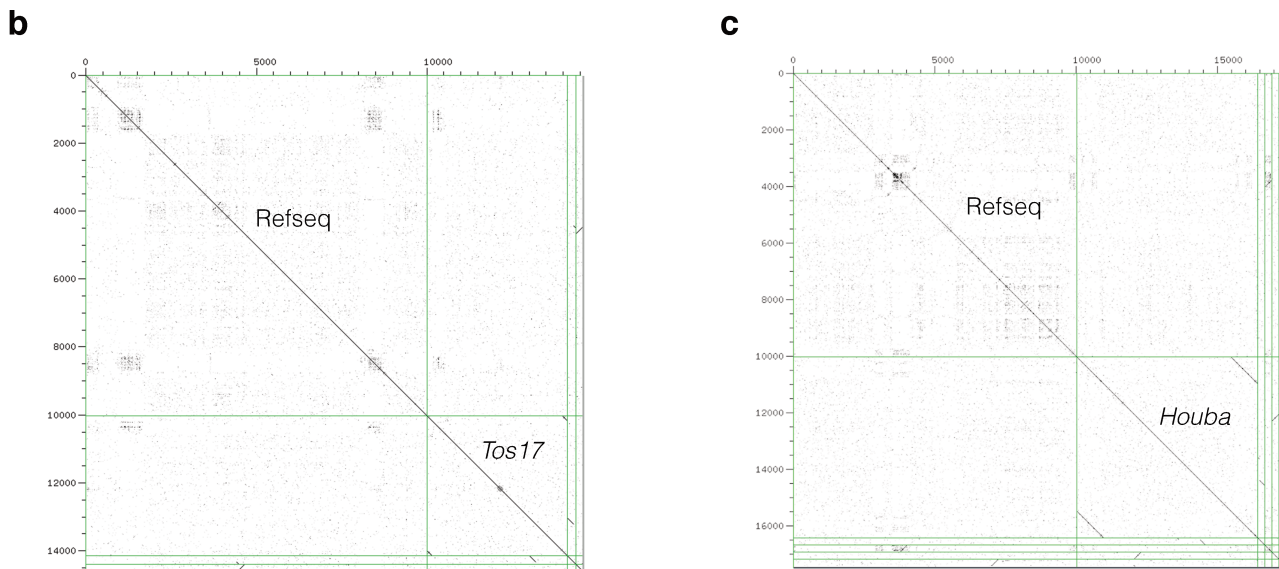


Figure 19. Activité transpositionnelle de *Tos17* et *Houba* dans l'albumen de riz. (a) Tableau des résultats pour le séquençage profond (100X) de l'albumen de riz. (b) Validations par *dotplot* des néo-insertions détectées par le re-séquençage de l'albumen de riz.

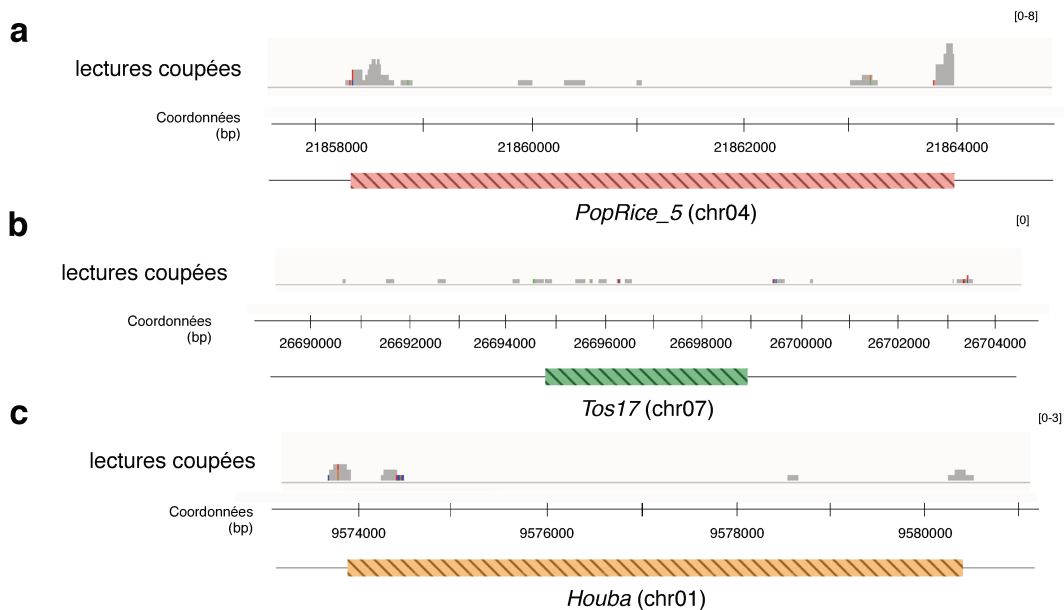


Figure 20. Détection des néo-insertions de *PopRice* par analyse des lectures coupées. Abondance des lectures coupées au locus de *PopRice* sur le chromosome 4 (a), de *Tos17* sur le chromosome 7 (b), de *Houba* sur le chromosome 1 (c). La couverture maximale est indiquée sur la droite. Image obtenue à partir du logiciel IGV. Couverture grise: abondance des lectures (non normalisée); les SNP sont représentés en couleur. De nombreuses lectures coupées sont détectées aux extrémités 5' et 3' de *PopRice* (a) comparé à *Tos17* ou *Houba*. La présence de lectures coupées aux extrémités de l'élément suggère la présence de lectures qui couvrent des néo-insertions de l'élément.

entières dans le génome de Nipponbare contre environ 145 copies entières pour *Houba*. Une nouvelle insertion de *Tos17* a été identifiée et validée par dot-plot (Figure 19b). Pour *Houba*, 504 insertions sont à valider par *dotplot*. Pour une famille de l'ampleur d'*Houba*, la procédure de validation manuelle est conséquente. Dans un premier temps, il faudra tester un échantillon de ces 504 insertions afin d'évaluer s'il est intéressant d'étendre la validation ou non. Au vu de son grand nombre de copies dans le génome, nous pouvons supposer que le nombre d'artefacts dus au pipeline sera conséquent.

En parallèle du pipeline pour détecter les néo-insertions, nous avons développé une autre analyse basée sur la détection de lectures dites coupées. Comme mentionné dans le chapitre 1 de l'Introduction (Figure 9, voir Introduction page 26), les lectures coupées sont des lectures qui alignent à deux endroits différents du génome. En effet, une lecture coupée représente l'extrémité (5' ou 3') de l'ET et la région flanquante de la néo-insertion. La détection de ce type de lectures permet d'obtenir avec plus de précision, à la base près, la localisation de la néo-insertion. L'analyse a été effectuée avec le programme *segemehl* (Coil et al. 2015, voir Matériel & Méthodes page 75) et les résultats ont été visualisés à l'aide du logiciel IGV. Comme indiqué sur la Figure 20, des lectures coupées sont détectées spécifiquement aux extrémités 5' et 3' de *PopRice* contrairement aux extrémités 5' et 3' de *Houba* et de *Tos17*. Cette analyse suggère une fois de plus la présence de lectures qui couvrent des néo-insertions de *PopRice*. De plus, le fait que les lectures coupées soient détectées uniquement aux extrémités de l'ET, indique que ce ne sont pas des artefacts. Néanmoins, le nombre de lectures coupées est faible (8 lectures au locus de *PopRice* contre 0 au locus de *Tos17* et 3 lectures au locus d'*Houba*). Une analyse plus approfondie devra être réalisée pour identifier les régions flanquantes couvertes par ces lectures et pour pouvoir conclure sur le taux de transposition de ces ET dans l'albumen.

2.3 Analyse comparative de *PopRice* au sein des céréales

2.3.1 Histoire évolutive de *PopRice* au sein des céréales

Le genre *Oryza* possède 24 espèces dont 2 espèces cultivées et 22 espèces sauvages. Il représente environ 15 millions d'années d'évolution (Figure 21a) et constitue un vaste réservoir de diversité génétique. Les deux espèces cultivées ont été domestiquées de manière indépendante une première fois, il y a environ 10000 ans en Asie (*O. sativa* var. *japonica* et *indica* dit « riz asiatique »), et une deuxième fois, il y a environ 3000 ans en Afrique (*O.*

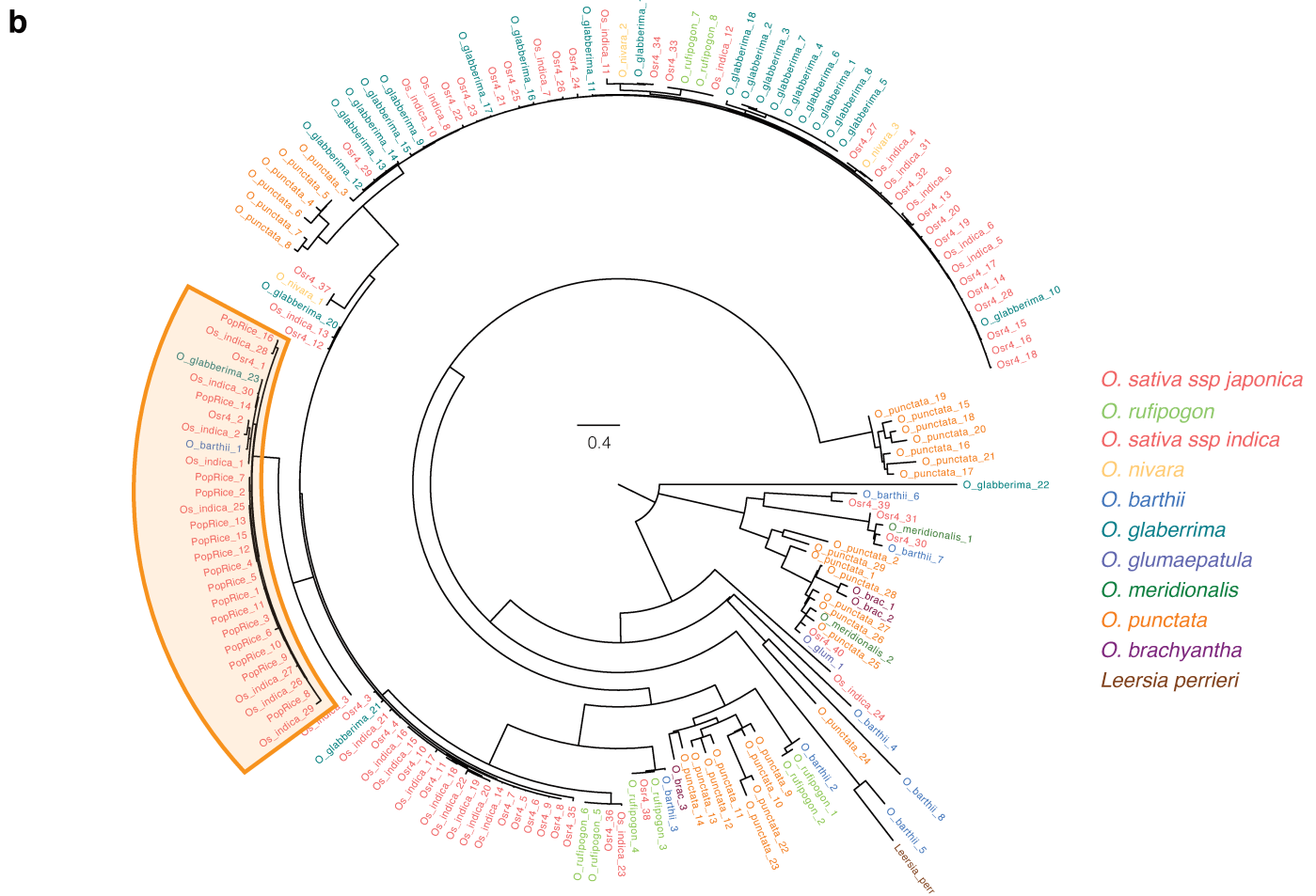
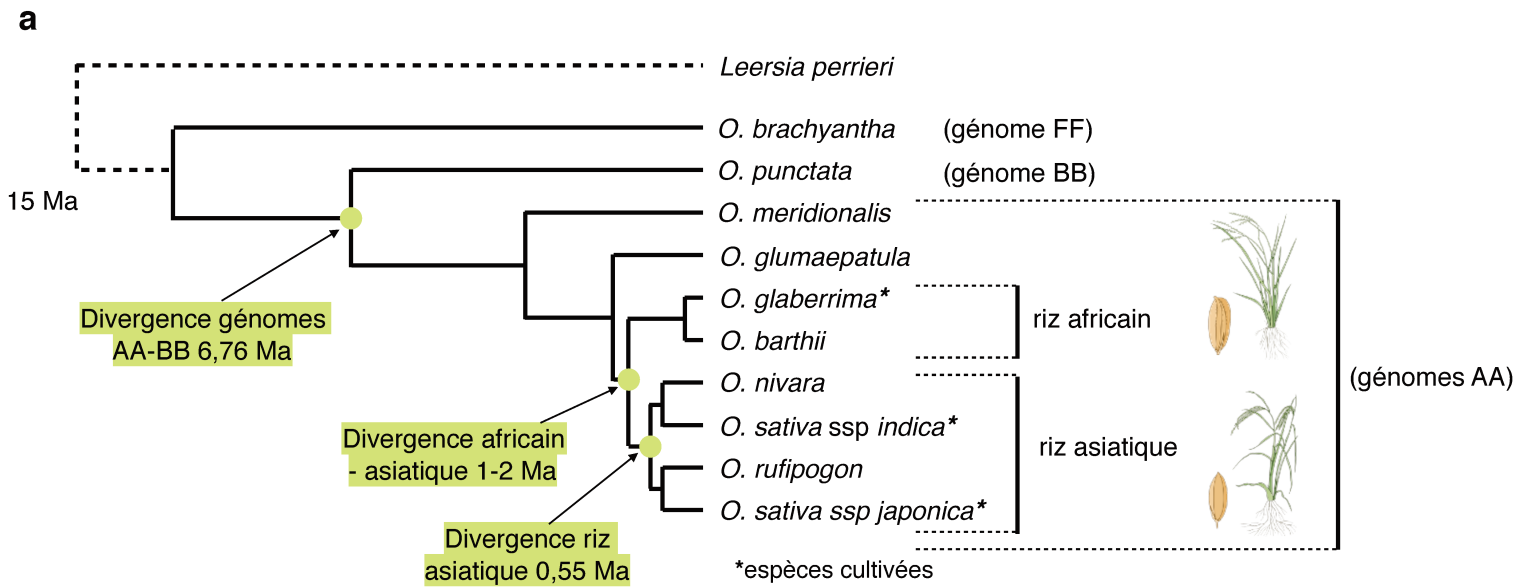


Figure 21. Analyse évolutive des éléments de *PopRice* au sein du genre *Oryza*. (a) Arbre phylogénétique des espèces du genre *Oryza* (adapté de D. Zwickl, Sanderson lab). Ma : millions d'années. (b) Arbre phylogénétique des éléments appartenant à la famille de *Osr4* présents dans les 10 génomes du genre *Oryza* et *Leersia perrieri*. L'arbre a été réalisé à partir des séquences des LTR de chaque élément et le clade surligné en orange correspond aux éléments *PopRice* et *PopRice-like*.

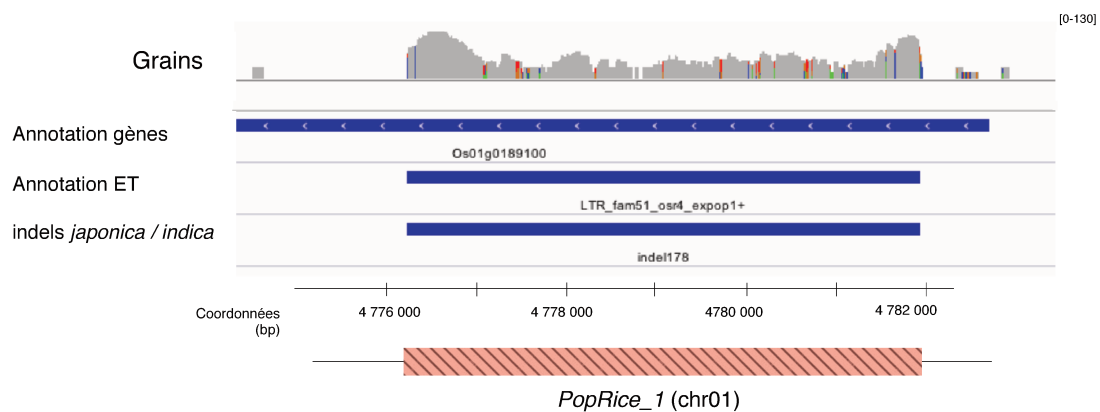


Figure 22. Polymorphisme d'insertion au locus de *PopRice_1* entre *japonica* et *indica*. L'analyse des insertions et des délétions (indels) entre les deux sous-espèces *japonica* et *indica* révèle que certaines insertions de *PopRice* sont polymorphes (travaux d'E. Lasserre). Par exemple, l'insertion de *PopRice_1* (insérée dans l'intron du gène *Os01g0189100*/LOC_Os01g09384) est absente du génome d'*indica* et est identifiée comme l'indel 178. La couverture de séquençage du mobilome de grains de Nipponbare est montrée en exemple. La couverture maximale est indiquée sur la droite. Image obtenue à partir du logiciel IGV. Couverture grise: abondance des lectures (non normalisée); les SNP sont représentés en couleur.

glaberrima dit « riz africain ») (Vaughan et al. 2008). Les relations phylogénétiques entre les espèces d'*Oryza* sont finement étudiées et l'histoire évolutive du riz est très bien caractérisée. Un nombre important de ressources génétiques et génomiques sont aujourd'hui disponibles notamment pour *Oryza sativa* (Ohyanagi et al. 2016), ce qui en fait un intéressant modèle pour étudier les processus d'évolution à court et long terme. Les espèces d'*Oryza* sont divisées en différents types de génomes dont les génomes diploïdes de type AA, BB et FF (Figure 21a). Au sein de l'équipe d'Olivier Panaud, nous disposons de 11 génomes séquencés et assemblés (dont 10 espèces d'*Oryza* et une espèce *outgroup* *Leersia perrieri* de la tribu des *Oryzaceae*) qui couvre l'histoire des *Oryza* (The International *Oryza* Map Alignment Consortium, *en révision*).

Afin de retracer l'histoire évolutive de *PopRice*, une analyse par BLAST sur ces 11 génomes a permis de montrer que la famille *Osr4* était présente dans toutes les espèces *Oryza* étudiées. Afin de distinguer les éléments *PopRice* de la famille *Osr4*, des analyses phylogénétiques ont été réalisées à partir des séquences des LTR (partie la plus divergente entre *PopRice* et les autres membres d'*Osr4* chez la référence *O. sativa japonica Nipponbare*) de chaque élément détecté chez les 11 génomes (Figure 21b). Différents groupes se distinguent suggérant une grande diversité au niveau des LTR au sein de la famille *Osr4* chez les espèces d'*Oryza*. Par ailleurs, le clade qui regroupe les éléments de *PopRice* regroupe également des éléments provenant d'autres espèces ou sous-espèces que *O. sativa ssp. japonica* indiquant ainsi la présence d'éléments proches de *PopRice* chez la sous-espèce *indica* et chez le riz africain *O. glaberrima* et *O. barthii*. De plus, en collaboration avec Éric Lasserre (Université de Perpignan) nous avons montré que 12 des 17 loci de *PopRice* étaient polymorphes entre les deux sous-espèces *japonica* et *indica* ce qui témoigne d'une activité récente de *PopRice* au sein des variétés asiatiques (Figure 22).

De par son importance économique et culturelle, les ressources génétiques chez le riz ne cessent de se multiplier et le consortium *The 3,000 rice genomes project* a sélectionné et séquencé 3000 accessions de riz d'*O. sativa* provenant de 89 pays différents (2014). Cette ressource est disponible, à l'état pendant de séquences brutes (lectures non assemblées). Les accessions sont classées en 5 groupes variétaux (Figure 23a) regroupant 1067 variétés traditionnelles et 1933 variétés issues de l'amélioration variétale. Au sein de mon équipe d'accueil, un projet a été développé sur l'activité transpositionnelle des ET d'une part au sein des 1067 variétés traditionnelles puis au sein des 3000 génomes (travaux de thèse de M-C Carpentier). Le pipeline d'analyse utilisé pour cette analyse est celui utilisé pour la détection des néo-insertions de

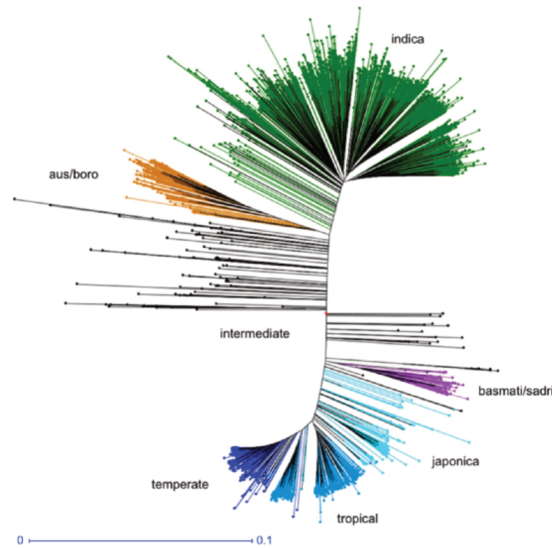
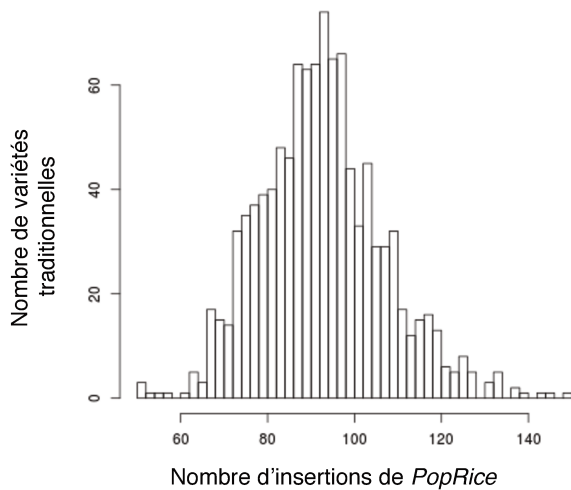
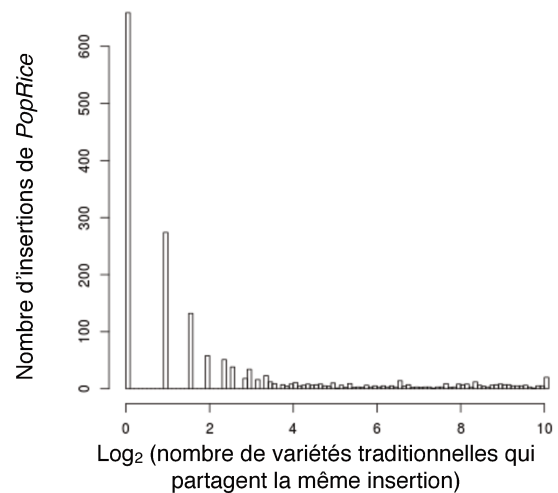
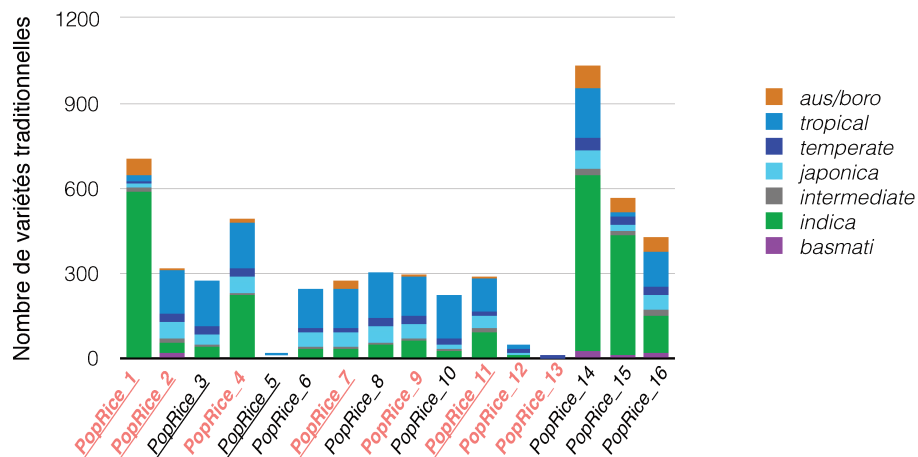
a**b****c****d**

Figure 23. Analyse évolutive des éléments de *PopRice* au sein de 3000 accessions de riz (a) Classification des 3000 accessions de riz regroupées en 5 groupes variétaux distincts réalisée à partir des 18.9 millions de variants SNP (*The 3,000 rice genomes project*) (b) Nombre d'insertions de *PopRice* présent dans les variétés traditionnelles séquencées dans le projet des 3000 génomes (analyse et figure réalisées par M-C Carpentier). (c) Nombre de variétés qui partagent la même insertion en Log2 (analyse et figure réalisées par M-C Carpentier). Par exemple, sur les 2327 insertions détectées dans les 1067 variétés traditionnelles, plus de 600 insertions sont présentes dans une seule variété et témoignent de l'activité récente de *PopRice*. (d) Nombre de variétés traditionnelles par groupe variétale qui partage les insertions présentes dans notre génome de référence (IRGSP1.0). Les copies en rouges sont celles qui ont 100% d'identité entre leurs LTR et les copies soulignées sont fortement couvertes dans le mobilome de Nipponbare.

PopRice dans les données de reséquençage (Figure 17) Dans le cadre de cette analyse, pour chaque insertion détectée, le nombre de lectures qui soutient cette insertion est comptabilisé et un minimum de 5 lectures est requis pour valider l'insertion. Le nombre d'insertions pour chaque famille d'ET testée est comptabilisé pour chacune des accessions. Cette étude a notamment été réalisée sur les insertions de *PopRice* pour lesquelles 1671 insertions ont été détectées chez 1067 variétés traditionnelles. Plus précisément, le nombre d'insertions de *PopRice* par accession est compris entre 51 et 149 (Figure 23b). Or, environ 700 insertions de *PopRice*, soit près de la moitié de ces insertions, sont uniques et donc présentes dans une seule accession (Figure 23c) indiquant que l'activité de *PopRice* dans ces génomes est très récente.

Notre analyse s'est ensuite concentrée sur les insertions de *PopRice* présentes dans le génome de référence Nipponbare et pour lesquelles la structure de l'insertion est connue (identité des LTR, domaines codants conservés) et dont l'activité est caractérisée notamment grâce au séquençage du mobilome (Figure 23d, Tableau 1). Par exemple, *PopRice_14* semble être une des plus anciennes copies de *PopRice*, 96% des variétés traditionnelles partageant cette insertion. Dans le génome de Nipponbare cette copie présente 93% d'identité entre ses LTR et fait partie des copies les plus divergentes de la sous-famille de *PopRice* et ne semble donc plus être une copie active (Figure 23). En revanche, et de façon plus surprenante, *PopRice_1* semble être une insertion relativement ancienne (partagée par plus de 700 variétés) or il semble que cette copie soit active chez Nipponbare car elle présente 100% d'identité entre ses LTR et est fortement couverte dans le mobilome. Le fait que cette copie soit conservée dans de nombreuses accessions suggère une pression de sélection qui pourrait être due à sa localisation dans le génome (présent dans l'intron du gène LOC_Os01g09384, Tableau 1).

2.3.2 Activité de *PopRice* chez différentes variétés de riz

Afin de caractériser l'activité de *PopRice* dans différentes espèces et variétés d'*Oryza*, nous avons dans un premier temps développé une collaboration avec le laboratoire LMI-Rice (Hanoi, Vietnam). Le LMI-Rice est un laboratoire mixte international qui regroupe l'IRD, l'Institut de Génétique Agronomique du Vietnam, l'Académie des Sciences Agricoles du Vietnam, l'Université des Sciences et Techniques d'Hanoi ainsi que l'Université de Montpellier. Le LMI possède une collection de 188 variétés vietnamiennes de riz issues des deux sous-espèces *japonica* et *indica* (Figure 24a) (Phung et al. 2014). Il existe au sein de cette collection une riche diversité génétique et une large étude de phénotypage a été menée par le LMI sur ces 188 variétés (Phung et al. 2014; 2016) ce qui fait de cette collection un intéressant modèle d'étude

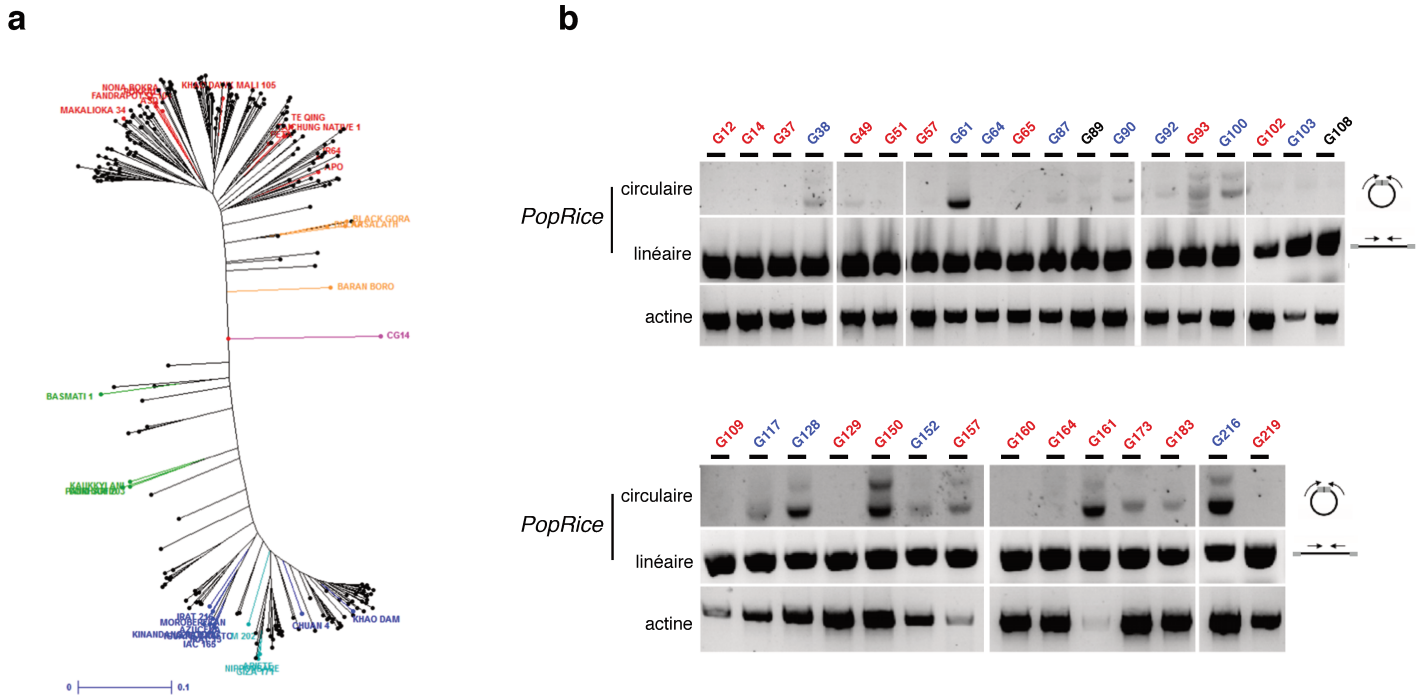


Figure 24. Activité de *PopRice* dans les grains des variétés de la collection vietnamienne de riz *Oryza sativa*. (a) Arbre phylogénétique de 271 accessions de riz dont la collection de l'Université de Hanoï (Phung et al., 2014). En rouge sont représentées les accessions *indica*, en jaune les accessions *aus/boro*, en vert les accessions *sedri/basmati*, en bleu foncé les accessions *japonica* tropical, en bleu clair les accessions *japonica* tempéré. CG14 (en rose), accession d'*O. glaberrima*, est utilisé comme *outgroup*. Chaque point noir correspond à une accession vietnamienne. (b) Analyse de la présence de formes circulaires de *PopRice* par PCR inverse (amorces voir Annexes) sur de l'ADN génomique extrait de grains de 34 variétés vietnamiennes. Un couple d'amorces qui amplifie spécifiquement les formes extrachromosomiques des éléments de *PopRice* au niveau de la jonction des cercles (représenté par les flèches noires) a été utilisé. Les amplifications PCR avec des amorces spécifiques des formes linéaires de *PopRice* et du gène actine sont montrées comme contrôles. Les variétés en rouge correspondent à des accessions *indica* et les variétés en bleu à des accessions *japonica*. Les variétés en noir ne sont pas encore définies de façon certaine.

de l'activité de *PopRice*.

Dans une première expérience pilote, une quarantaine de variétés a été sélectionnée afin de d'analyser l'activité de *PopRice* dans cet échantillon de la collection (Figure 24b). Pour ce faire nous avons extrait l'ADN de grains puis réalisé des PCR inverses avec l'utilisation d'un couple d'amorces qui amplifie spécifiquement les ADN_{Necc} de *PopRice* au niveau de la jonction des cercles. Notre étude montre que les ADN_{Necc} de *PopRice* sont abondants dans certaines variétés telles que G61, G128, G216 alors qu'ils semblent totalement absents chez des variétés telles que G12, G65 et G160 par exemple (Figure 24b). Il est également intéressant de noter que l'activité de *PopRice* ne semble pas dépendre de la sous-espèce de la variété, l'ADN_{Necc} de *PopRice* étant détecté dans les deux variétés *japonica* et *indica*. Dans le but de valider ces premiers résultats, quatre variétés (2 variétés *indica* et 2 variétés *japonica*) ont été sélectionnées pour l'analyse par mobilome-seq, deux variétés dites « positives » G61 et G161 (présence d'ADN_{Necc} de *PopRice*) et deux variétés négatives G64 et G65 (absence d'ADN_{Necc} par PCR) (Figure 25). La présence d'ADN_{Necc} de *PopRice* est confirmée par mobilome-seq dans les deux variétés G61 et G161 alors que les cercles de *PopRice* sont absents chez les deux autres variétés G64 et G65. Ces résultats indiquent que l'activité de *PopRice* n'est pas conservée chez toutes les variétés de riz.

2.3.3 Activité de *PopRice* chez différentes espèces de riz

Dans un deuxième temps, pour aller plus loin dans la caractérisation de l'activité de *PopRice*, une analyse mobilomique comparative de grains chez les deux espèces de riz africains, l'espèce sauvage *Oryza barthii* et l'espèce cultivée *Oryza glaberrima*, a été réalisée. Les banques de mobilome ont été réalisées à partir d'ADN de grains issus de ces deux espèces et ont été séquencées. L'analyse de la couverture de séquençage au locus de *PopRice* dans les banques de mobilome de *O. glaberrima* et de *O. barthii* montre que ce locus est enrichi dans les deux banques, suggérant que *PopRice* est présent sous forme d'ADN_{Necc} dans les trois espèces du genre *Oryza* (Figure 26). Un nombre important de SNP est observé et s'explique par le fait que les lectures de séquençage ont été alignées contre le génome de référence de *Nipponbare*. Néanmoins, étant donné que la couverture de séquençage est relativement faible comparée à la banque de mobilome de *Nipponbare* (couverture d'environ 70X pour *O. glaberrima* et *O. barthii* contre 130X pour *Nipponbare*), ce résultat est préliminaire et devra être confirmé par Southern blot.

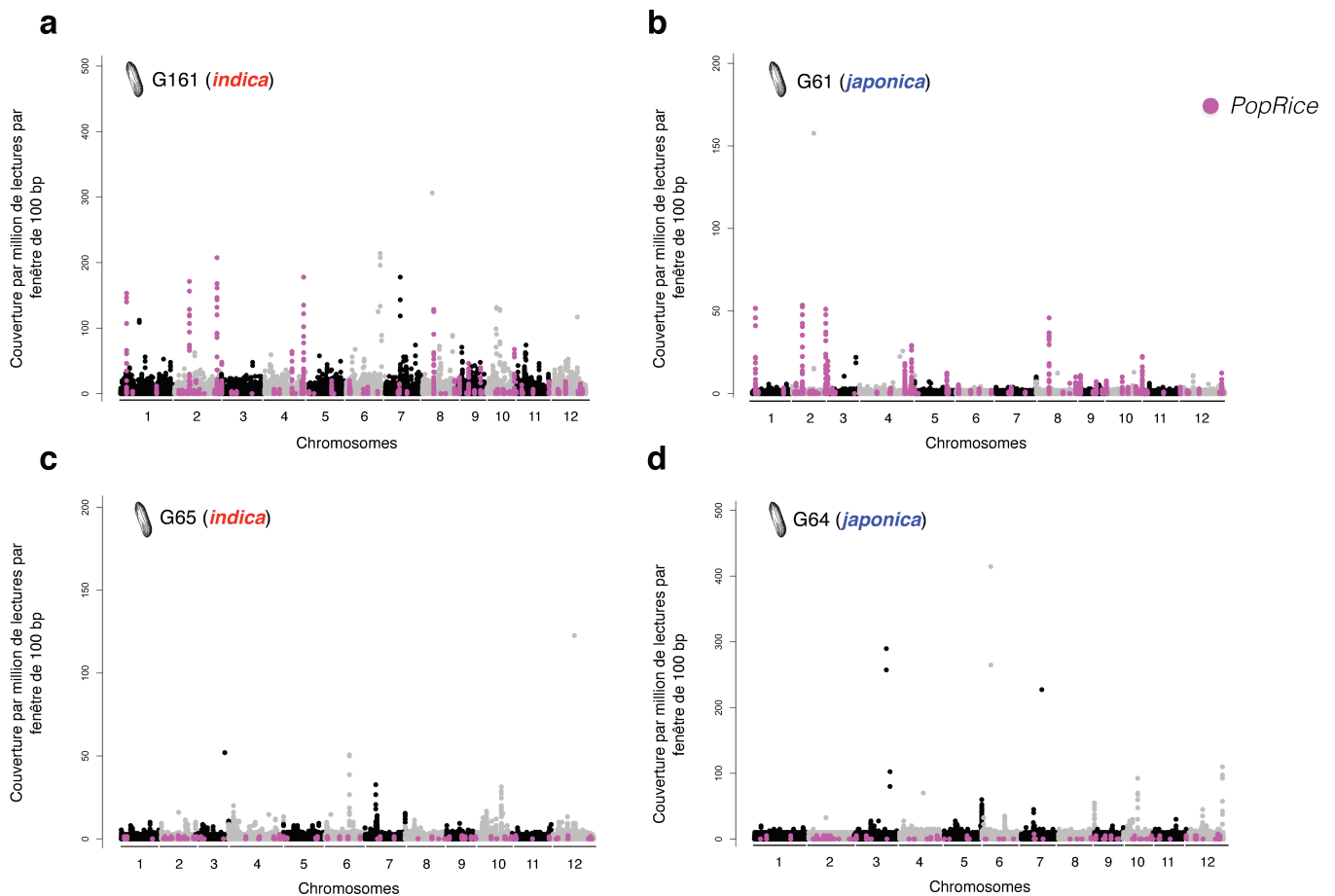


Figure 25. Mobilome des grains de riz des variétés vietnamiennes *Oryza sativa*. Couverture de séquençage des données de mobilome de 4 accessions vietnamiennes à l'échelle du génome. Chaque point représente la couverture obtenue par fenêtre de 100 paires de bases. Toutes les fenêtres contenant un ET annoté sont représentées dans ces figures, de façon alternative en noir et gris pour les douze chromosomes de riz. Les points roses représentent les fenêtres correspondant au loci de *PopRice*.

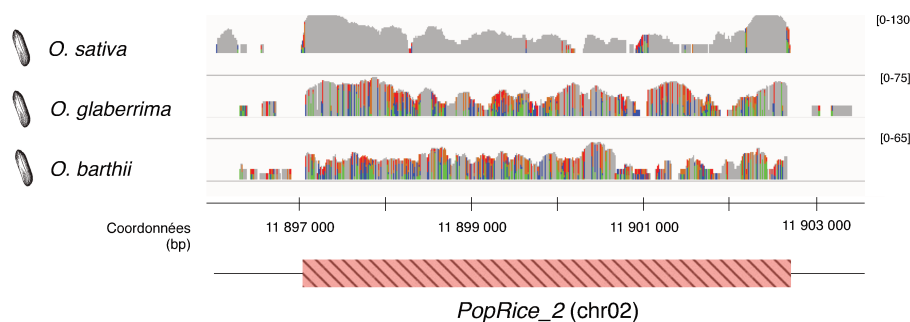


Figure 26. Activité de *PopRice* dans les grains au sein du genre *Oryza*. (a) Couverture de séquençage au locus de *PopRice* sur le chromosome 2 dans le mobilome de grains de 3 espèces du genre *Oryza* : *O. sativa*, *O. glaberrima* et *O. barthii*. La couverture maximale est indiquée sur la droite. Image obtenue à partir du logiciel IGV. Couverture grise: abondance des lectures (non normalisée); les SNP sont représentés en couleur.

Ces résultats remettent néanmoins en question notre hypothèse de départ qui suggérait un rôle de *PopRice* dans le développement du grain dans le genre *Oryza*. L'absence d'ADNecc de *PopRice* dans certaines variétés vietnamiennes suggère que *PopRice* ne semble pas nécessaire au bon développement du grain. Cependant, l'activité conservée de *PopRice* dans des espèces très éloignées comme *O. glaberrima* ou *O. barthii* suscite toutefois des interrogations. Très récemment, nous avons développé une analyse de type *genome-wide association study* (GWAS) (analyse effectuée en collaboration avec Ernandes Manfroi, Université UFRGS, Brésil, en accueil au laboratoire) sur toute la collection variétale vietnamienne en utilisant la présence ou l'absence d'ADNecc de *PopRice* comme trait phénotypique (Figure 27). L'analyse est actuellement en cours et je présente ci-dessous uniquement les résultats préliminaires que nous avons obtenus.

Une analyse GWAS permet de rechercher des associations entre des marqueurs génétiques, ici des SNP, et un ou plusieurs phénotypes étudiés, ici la présence ou l'absence d'ADNecc de *PopRice*. Pour notre analyse nous avons utilisé la matrice de 25971 SNP obtenue par séquençage GBS (*Genotyping By Sequencing*) de 185 lignées de la collection (Phung et al. 2014). Pour le phénotypage 182 variétés ont été analysées par PCR inverse : 73 variétés avec le phénotype « 1 » (présence d'ADNecc de *PopRice*), 65 variétés avec le phénotype « 0 » (absence d'ADNecc de *PopRice*) et 44 variétés avec un phénotype non déterminé. Différents tests statistiques ont été effectués afin de s'assurer de la robustesse des résultats d'association. Par exemple, un des paramètres importants à vérifier lors d'une analyse GWAS est la structure de la population étudiée et ses liens de parenté. En effet, si les fréquences alléliques et les taux de déséquilibre de liaison sont très variables au sein de la population étudiée, ils peuvent induire de nombreux faux positifs. Les diagrammes quantile-quantile (Q-Q plot) sont un moyen efficace pour étudier ce paramètre (Figure 27a). Les Q-Q plots représentent la distribution des valeurs-p ou *p-values* observées contre la distribution des *p-values* attendues (représentée par une diagonale rouge) et permet de visualiser des déviations de la distribution observée par rapport à la distribution attendue. Sur ce type de graphique, pour qu'une association soit significative la déviation par rapport à la courbe rouge doit être uniquement à l'extrémité de la distribution, c'est-à-dire pour les *p-values* les plus significatives. Autrement dit, la structure de notre population semble correcte pour espérer obtenir des résultats d'association significatifs. Les résultats de l'analyse d'association des SNP et du phénotype sont représentés dans un Manhattan plot (Figure 27b). Le Manhattan plot représente les *p-values* d'association de chaque SNP avec le phénotype, en fonction de leur position sur le génome. Deux régions du génome

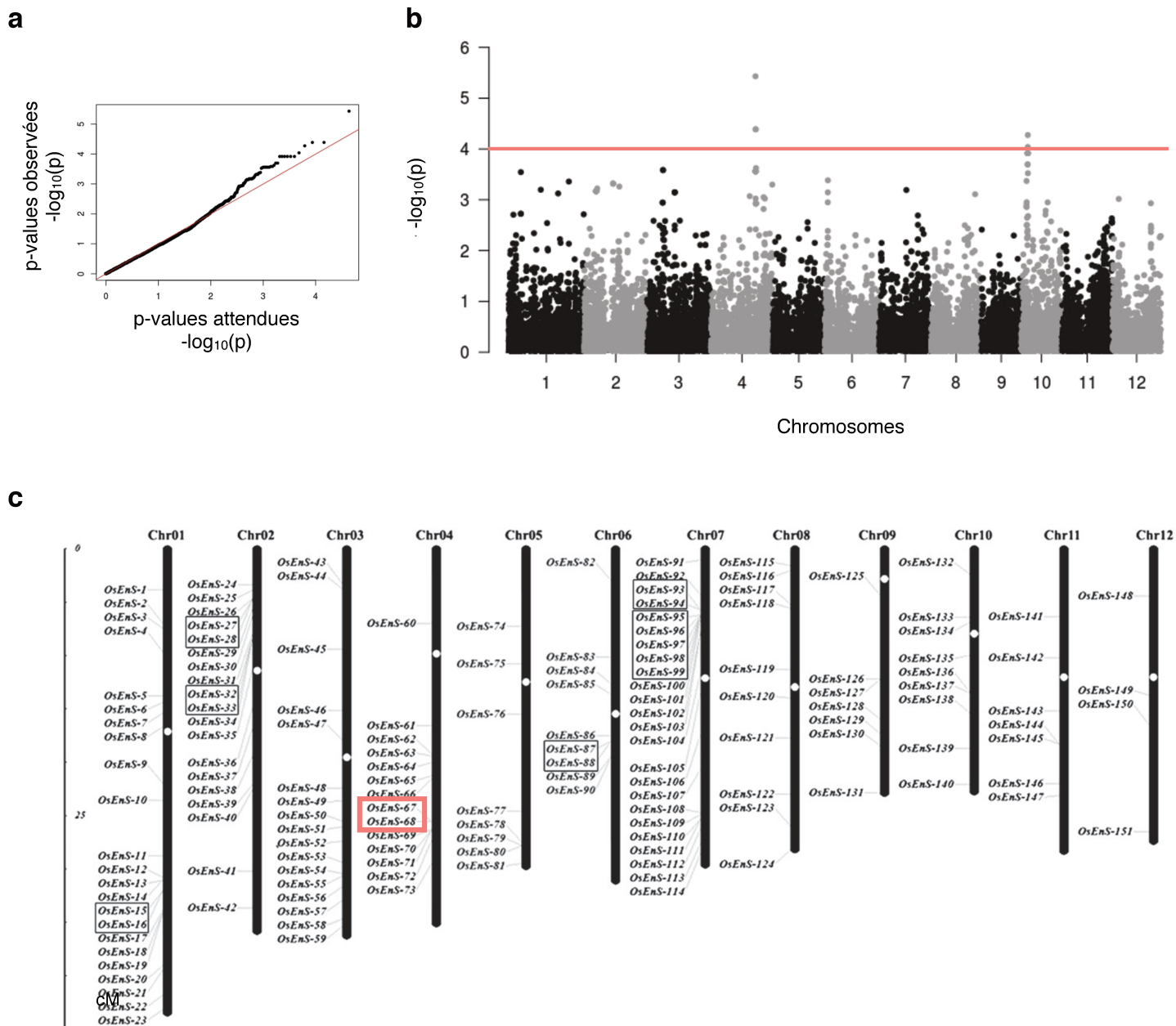


Figure 27. Analyse par GWAS des variétés vietnamiennes de *Oryza sativa*. (a) Quantile-Quantile plot (QQ-plot) qui représente la distribution des *p-values* observées (ici représentées par des points noirs) contre la distribution des *p-values* attendues (ici représentées par une diagonale rouge). La déviation des *p-values* observées par rapport aux *p-values* attendues indique la significativité des résultats de l'analyse d'association. (b) Manhattan plot qui représente les résultats d'association entre la matrice de SNP et un phénotype (ici présence ou absence d'ADNec de *PopRice*). Chaque point représente un SNP avec sa *p-value* associée au phénotype. La ligne rouge délimite le seuil de significativité. Tous les points qui se situent au dessus sont des SNP qui sont statistiquement associés au phénotype. Deux régions (chromosomes 4 et 10) semblent statistiquement associées à la présence (ou l'absence) d'ADNec de *PopRice*. Les analyses et les figures ont été réalisées par E. Manfroi. (c) Localisation des 151 gènes spécifiques de l'albumen chez *Oryza sativa* (Nie et al. 2013). Les deux gènes encadrés en rouge, *OsENS-67* et *OsENS-68*, sont localisés dans la région du chromosome 4 identifiée par l'analyse GWAS.

sur le chromosome 4 et 10 semblent être statistiquement associées au phénotype (Figure 27b). Les SNP qui présentent une *p-value* significative sur le chromosome 4 entourent une région d'environ 250 kb qui regroupe 29 gènes et notamment deux gènes « spécifiques de l'albumen » *LOC_Os04g43160* et *LOC_Os04g43170* (Figure 27c). Ces deux gènes semblent être deux candidats intéressants pour expliquer la présence des ADNecc de *PopRice* dans l'albumen. En revanche, les gènes présents dans la région du chromosome 10, environ 100 kb, ne semblent pas avoir de lien établi avec un rôle dans l'albumen ou la graine et demandent une analyse plus approfondie.

2.3.4 Étude comparative de l'activité de *PopRice* chez les céréales

Afin de déterminer si la famille *PopRice* est spécifique au genre *Oryza* ou si *PopRice* est présent et actif dans d'autres céréales une analyse *in silico* a été réalisée à partir des génomes séquencés de céréales. Par recherche BLAST, des éléments similaires à *PopRice* (*PopRice like elements* PLE) chez différentes céréales ont été identifiés. L'analyse par dot-plot révèle que les PLE identifiés chez le sorgho ou le maïs présentent la même structure que *PopRice* au niveau des LTR. Autrement dit, les PLE ont des boîtes répétées dans leur LTR (Figure 28a). Il est également intéressant de noter que le nombre de répétitions semble varier en fonction de l'espèce et que leur séquence dépend de l'espèce. Par exemple, les éléments détectés chez le maïs semblent avoir acquis un nombre très important de répétitions en comparaison avec les éléments du riz ou du sorgho. Effectivement, les séquences des LTR des PLE des 3 espèces (riz, sorgho, maïs) sont plus divergentes (55% d'identité moyenne) que leur région interne (75% d'identité moyenne) (Figure 28b). Les motifs répétés dans le LTR semblent différer entre les espèces. Néanmoins, si les taux moyens de similarité sont comparés à la conservation génique observée entre céréales (El Baidouri et al. 2014), les résultats révèlent que les taux de similarité des PLE inter-espèces sont plus élevés que les taux de similarité entre gènes (Figure 28b). Par exemple, entre le riz et le sorgho, la conservation génique représente 37% alors que le taux de similarité entre *PopRice* et *SbPLE* de sorgho (*Sorghum bicolor* PLE) représente 76%. Ces observations suggèrent que l'insertion de la famille de *PopRice* est plus récente que l'ancêtre commun et que la prolifération de *PopRice* dans ces génomes pourrait être due à un transfert horizontal. Afin d'évaluer l'activité de *PopRice* chez d'autres céréales que le riz, une analyse mobilomique a été effectuée chez 3 espèces : maïs (*Zea mays*), blé (*Triticum aestivum*) et sorgho (*Sorghum tricolor*). Les banques de mobilome ont été réalisées à partir d'ADN de grains issus de ces 3 espèces et ont été séquencées. L'analyse de ces données a montré que les ADNecc de

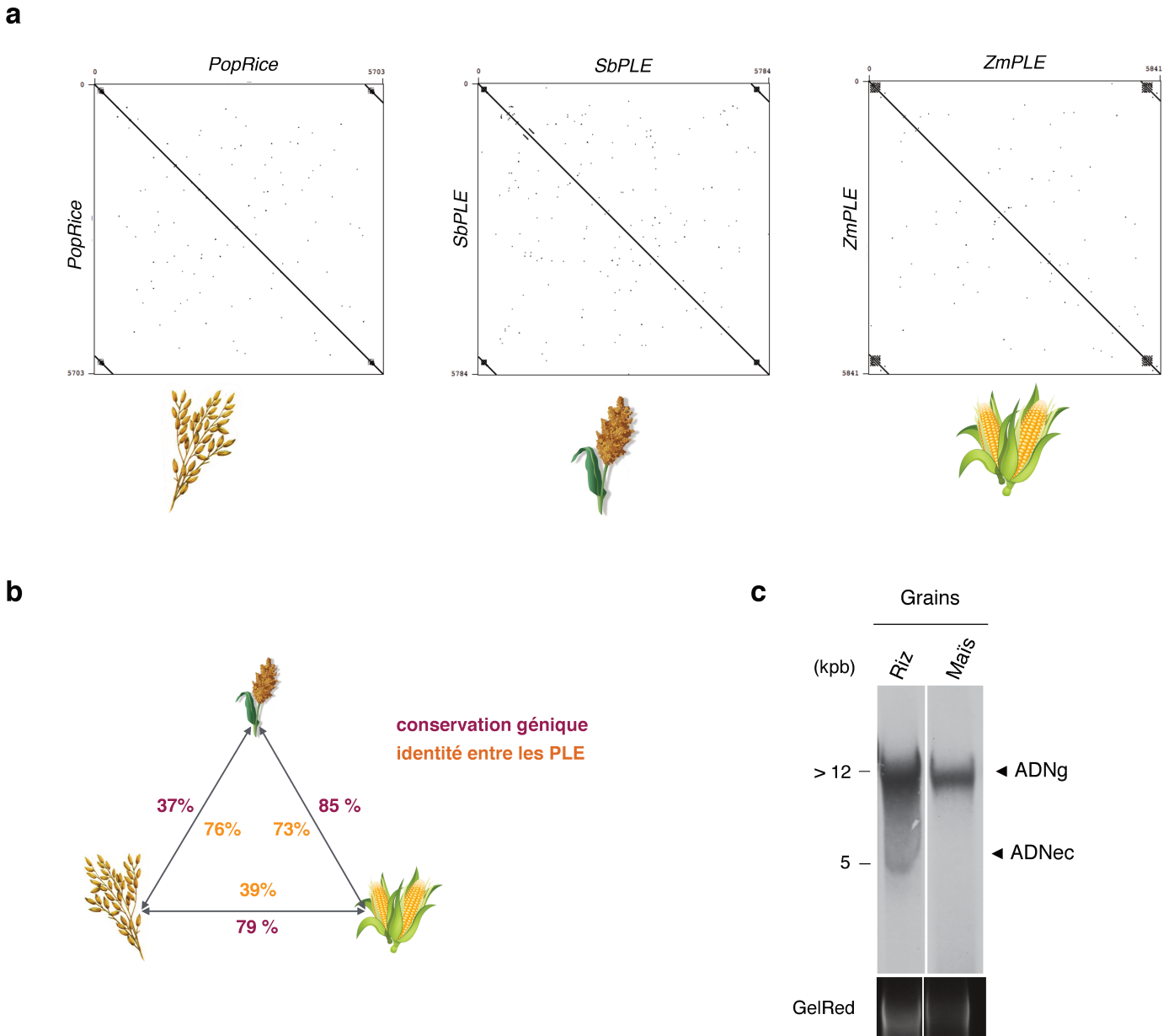


Figure 28. Analyse d'éléments *PopRice*-like chez trois céréales. (a) Analyse par dot-plot des éléments « cousins » de la famille de *PopRice* chez le maïs (*Zea mays ssp. mays var. B73*) et le sorgho (*Sorghum bicolor*). Ces dot-plots montrent que la famille *PopRice* est présente chez plusieurs espèces de céréales. Les éléments chez le maïs (*Zea mays PopRice-like element ZmPLE*) (Chr4:219291566-219291218) et le sorgho (*Sorghum bicolor PopRice-like element SbPLE*) (clone SB_BBc0109L12) ont conservé des motifs répétés dans leurs LTR. (b) Les PLE présentent une forte similarité de séquences (en moyenne 75%). Le taux moyen de conservation au niveau des séquences codantes des gènes (CDS) entre les espèces (El Baidouri et al., 2014) est donné à titre de comparaison. (c) Analyse par Southern blot en utilisant de l'ADN génomique (ADNg) non digéré extrait à partir de grains matures de riz (*O. sativa ssp japonica*) et de grains de maïs (*Zea mays ssp. mays var. B73*) hybridés avec une sonde spécifique de la famille *PopRice*, *PopRice_sonde2* (voir Annexes). Cette analyse montre que les formes extrachromosomiques (ADNec) de *PopRice* sont spécifiquement détectées chez le riz. Le marquage du gel au GelRed est montré comme contrôle de chargement.

PopRice n'étaient pas détectés dans ces tissus (non montré). Ces résultats ont ensuite été confirmés par Southern blot chez le maïs (Figure 28c) à l'aide d'une sonde qui hybride au niveau de la région interne de *PopRice* et montre que les formes extrachromosomiques ne sont pas détectées dans les grains de maïs. En définitive, la famille de *PopRice* a proliféré dans les génomes des céréales maïs néanmoins, l'activité de *PopRice* dans les grains semble spécifique au genre *Oryza*.

2.4 Activité de *PopRice* en condition de stress chez Nipponbare

Des éléments de réponse aux hormones ont été détectés dans les promoteurs des gènes fortement transcrits dans le grain (Xue et al. 2012) indiquant un lien étroit entre la signalisation hormonale et le développement des grains. Dans le but de comprendre si les hormones pouvaient jouer un rôle sur l'activité de *PopRice* nous avons testé l'effet de l'ABA sur la présence d'ADNec de *PopRice* en condition de culture *in vitro*. Comme précédemment mentionné, l'ABA joue un rôle important dans le développement du grain. De plus, la culture de cals chez le riz induit des changements au niveau de la méthylation de l'ADN (Stroud et al. 2013) qui conduit à la réactivation d'ET tel que *Tos17* (Hirochika et al. 1996) et qui en fait un tissu intéressant pour regarder l'activité de *PopRice*. D'après nos données de mobilome, *PopRice* n'est pas actif dans les cals. Dans le but de faire une expérience pilote, 5 μ M d'ABA ont été appliqués directement sur des cals de 10 semaines et l'ADN a été extrait à partir des cals traités et non traités à l'ABA. Par PCR inverse nous avons testé la présence d'ADNec de *PopRice* (Figure 29a). Les résultats montrent de multiples amplifications pour les cals traités à l'ABA pouvant indiquer une activité de *PopRice* et des recombinaisons éventuelles entre les ADNec. Afin de tester cette hypothèse, nous avons effectué une analyse par Southern blot en utilisant de l'ADN génomique non digéré et une sonde spécifique à *PopRice* (Figure 29b). La présence d'ADNec de *PopRice* n'a pas été détectée, remettant en cause notre hypothèse.

Discussion

Selon le modèle proposé pour expliquer le maintien du *silencing* des ET au cours des générations, une réactivation transitoire des ET s'observe dans les tissus accompagnant les gamètes (cellule centrale et cellule végétative) ainsi que dans l'albumen de la graine (Fultz et

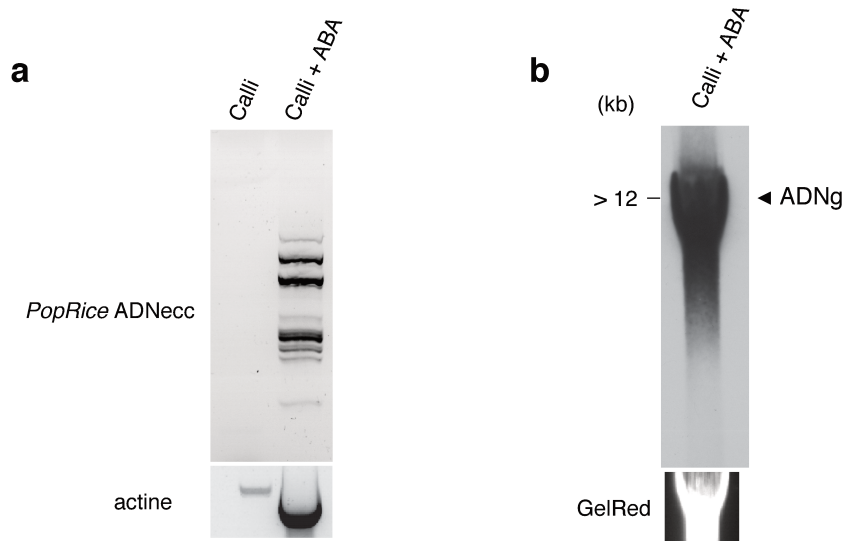


Figure 29. Activité de *PopRice* en condition de stress. (a) PCR inverse (amorces PRecc_F/R, voir Annexes) pour la détection des formes d'ADNecc de *PopRice* avec de l'ADN extrait à partir de cals de riz exposés ou non à un traitement à l'ABA ($5\mu\text{M}$) durant 7 jours. Les résultats montrent que le traitement à l'ABA semblent induire des formes d'ADNecc. (b) Analyse par Southern blot en utilisant de l'ADN génomique (ADNg) non digéré extrait à partir de cals de riz exposés ou non à un traitement à l'ABA, hybridés avec une sonde spécifique de la famille *PopRice*, *PopRice_sonde1* (voir Annexes). Cette analyse montre que seules les copies intégrées de *PopRice* sont détectées. Le marquage du gel au GelRed est montré comme contrôle de chargement.

al. 2015). Or, nos données ont révélé que seul le rétrotransposon *PopRice* semblait actif dans l'albumen de riz. L'activité prédominante de *PopRice* comparé aux autres RT a suscité des interrogations sur les mécanismes génétiques et épigénétiques qui s'exerçaient sur l'élément.

L'analyse des LTR de *PopRice* a révélé la présence de motifs de type g-box présents également dans les gènes glutélines (Yoshihara *et al.* 1996) spécifiquement exprimés dans l'albumen (Qu *et al.* 2008). Il apparaît que l'activation de certains rétrotransposons à LTR résulte de la combinaison entre un relâchement du contrôle épigénétique et le recrutement au niveau des LTR de facteurs de régulation spécifiques (Grandbastien 2015). Aujourd'hui, nos résultats sur la méthylation et l'expression de *PopRice* confirment l'hypométhylation et l'activité transcriptionnelle de *PopRice* dans l'albumen mais en revanche, aucune corrélation n'a été établie entre la présence de séquences *cis* régulatrices dans la séquence promotrice de *PopRice* et une augmentation de la transcription de *PopRice* dans l'albumen. Or, la présence de ces boîtes est la principale différence nucléotidique qui permet de dissocier les éléments de *PopRice* avec les autres éléments de la famille *Osr4*. De plus, nous n'avons jusqu'à présent étudié la transcription de *PopRice* et de *Osr4* qu'au début de la maturation du grain (14 jours après la pollinisation) et nous n'avons pas déterminé si les éléments *Osr4* et *PopRice* avait le même profil de transcription au cours du développement ou si le profil était différent entre les éléments. Une analyse transcriptomique à différents stades de développement du grain pourrait être envisagée afin de répondre à la question. L'analyse *in silico* des données de méthylome (Zemach *et al.* 2010) n'a pas été concluante et la caractérisation du niveau de méthylation pour chacun des éléments de la famille pourrait apporter des réponses sur la distinction entre les éléments. Le reséquençage des copies de *PopRice* et *Osr4* après traitement au bisulfite pourrait donc être envisagé.

Par ailleurs, les analyses par RT-qPCR et par PCR inverse ont révélé, respectivement, une faible activité transcriptionnelle et une absence d'ADNecc de *PopRice* dans les fleurs (avant pollinisation) suggérant que *PopRice* est activé après la pollinisation dans le tissu de l'albumen. Or, une étude récente a montré que chez *Arabidopsis* et chez le riz, l'hypométhylation observée dans l'albumen est initiée dans la cellule centrale du gaméophyte femelle (Park *et al.* 2016). Il est donc possible que le niveau de transcrits de *PopRice* soit trop faible pour être détecté à l'échelle de la fleur entière. Des futures analyses sur les différents tissus du gaméophyte femelle pourraient nous éclairer sur les facteurs génétiques et épigénétiques qui initient l'activité de *PopRice*.

Le nombre considérable de ressources phylogénétiques, génétiques et génomiques disponibles chez le riz nous a permis de caractériser l'histoire évolutive de *PopRice* au sein du genre *Oryza* d'une part et au sein des céréales d'autre part. En effet, l'analyse des insertions de *PopRice* dans 3000 accessions de riz asiatiques (travaux de M-C Carpentier) a permis de mettre en évidence l'activité récente de *PopRice* dans ces génomes. Nos données préliminaires ont également montré que de l'ADNecc de *PopRice* semblait être présent dans les grains de riz africains *O. glaberrima* et *O. barthii*. En revanche, si la famille de *PopRice* a largement proliféré dans les génomes des *Poaceae*, les PLE ne semblent pas actifs dans les grains provenant d'espèces telles que le maïs ou le sorgho. Il est toutefois intéressant de noter que les PLE présentent de forts pourcentages d'identité entre les espèces (>70%) ce qui suggère que leur prolifération pourrait être due à des transferts horizontaux (El Baidouri et al. 2014). Les transferts horizontaux d'ET *via* différents types de vecteurs paraissent être un moyen efficace pour l'ET de coloniser les génomes et ainsi assurer sa survie (Panaud 2016). En ce qui concerne les LTR, le nombre de motifs répétés accumulés chez les PLE semble varier entre espèces et les séquences de LTR divergent (environ 55% d'identité entre deux LTR de deux espèces différentes). Néanmoins, pour une même espèce, les éléments identifiés présentent 100% d'identité entre leurs deux LTR, indiquant une activité récente. Finalement, selon les espèces, la famille de *PopRice* pourrait avoir acquis différentes spécificités et un large spectre d'activité. L'étude du mobilome de ces espèces dans différentes conditions et tissus pourrait nous permettre de tester cette hypothèse.

À l'instar de HERV dans le développement du placenta chez les mammifères (voir Introduction page 21), l'activité conservée de *PopRice* dans des grains provenant d'espèces lointaines de riz sauvages et domestiquées nous a conduit à envisager l'hypothèse que *PopRice* puisse jouer un rôle dans le développement du grain ou dans le processus de germination. Toutefois, nos analyses mobilomiques sur une collection de variétés de riz vietnamiens (Phung et al. 2014) ont montré que l'ADNecc de *PopRice* n'était pas présent dans toutes les variétés. L'absence d'ADNecc de *PopRice* dans certaines variétés présentant un développement normal, suggère que *PopRice* n'est pas nécessaire au bon développement du grain. En revanche, l'utilisation de la présence ou de l'absence d'ADNecc de *PopRice* comme trait phénotypique a permis par une analyse GWAS de mettre en évidence chez ces variétés deux gènes candidats qui semblent liés à l'activité de *PopRice*. Ces deux gènes « spécifiques de l'albumen » sont impliqués respectivement, dans le métabolisme des lipides et dans la signalisation hormonale (Nie et al. 2013). Leur localisation dans le génome ne corrèle pas avec la présence proche d'une copie de *PopRice* dans le génome de Nipponbare. Nous ne pouvons néanmoins pas exclure que dans les

génomomes de la collection vietnamienne, il y ait une insertion de *PopRice* dans cette région. Pour ce faire le pipeline des néo-insertions devra être testé sur les données de re-séquençage des lignées de la collection quand celles-ci seront disponibles. Les données transcriptomiques chez la variété Zhenshan 97 (*O. sativa ssp japonica*) montrent que les deux gènes candidats sont tous deux exprimés dans l'albumen et présentent dans leur promoteur des éléments *cis* régulateurs de type g-box ou CAAT box (Nie et al. 2013). Toutefois, ces résultats sont préliminaires et devront être affinés. Afin qu'une analyse GWAS ait une puissance statistique suffisante, nous devons améliorer notre phénotypage, notamment pour les 44 variétés non déterminées. De même, il est évident que plus le phénotypage sera précis, meilleurs seront les résultats d'association. Nous envisageons d'évaluer par q-PCR l'abondance des ADNec de *PopRice* au sein des variétés vietnamiennes afin d'établir différentes catégories d'intensité de l'activité de *PopRice* et ainsi affiner notre phénotypage. De plus, pour valider l'association entre ces deux gènes et l'ADNec de *PopRice*, une étude de mutant est nécessaire, soit en analysant des mutants déjà existant pour ces deux gènes, soit en générant les mutants à l'aide de la technologie CRISPR (Shan et al. 2013).

Enfin, si *PopRice* ne joue pas de rôle biologique dans le tissu de la graine et qu'aucune pression de sélection ne s'exerce sur cette famille, son activité conservée au cours de l'évolution pourrait s'expliquer par une répllication égoïste de l'élément permise par l'accumulation de boîtes à glutélines lui permettant d'être actif dans les albumens de riz. Toutefois, la relation entre l'activité de gènes « spécifiques de l'albumen » et l'activité de *PopRice* dans l'albumen suscite des interrogations sur les mécanismes impliqués dans sa réactivation. De même, la présence d'insertions récentes de *PopRice* dans les génomes de riz alors que cette famille ne semble active que dans un tissu non transmis à la génération suivante paraît contradictoire. Ces résultats suggèrent que *PopRice* pourrait être actif dans d'autres conditions que celles que nous avons testées jusqu'à présent. L'albumen est un tissu complexe qui subit des changements physiologiques importants au cours de son développement (Zhou et al. 2013) et l'activité de *PopRice* pourrait être le reflet d'une réponse à un stress ou à une hormone. L'hormone ABA joue un rôle crucial dans le développement du grain chez le riz (Liu et al. 2015) et nous avons supposé que l'ABA pourrait donc influencer l'activité de *PopRice*. Notre expérience pilote a montré que l'application d'ABA sur des cals de riz semblait induire une certaine perturbation au niveau de la détection d'ADNec de *PopRice* par PCR inverse. Cependant aucune forme extrachromosomique de *PopRice* n'a été détectée par Southern blot. Par ailleurs, nous n'avons pas évalué l'état de méthylation de *PopRice* en conditions *in vitro*. Or, Cheng et al. (2015) ont

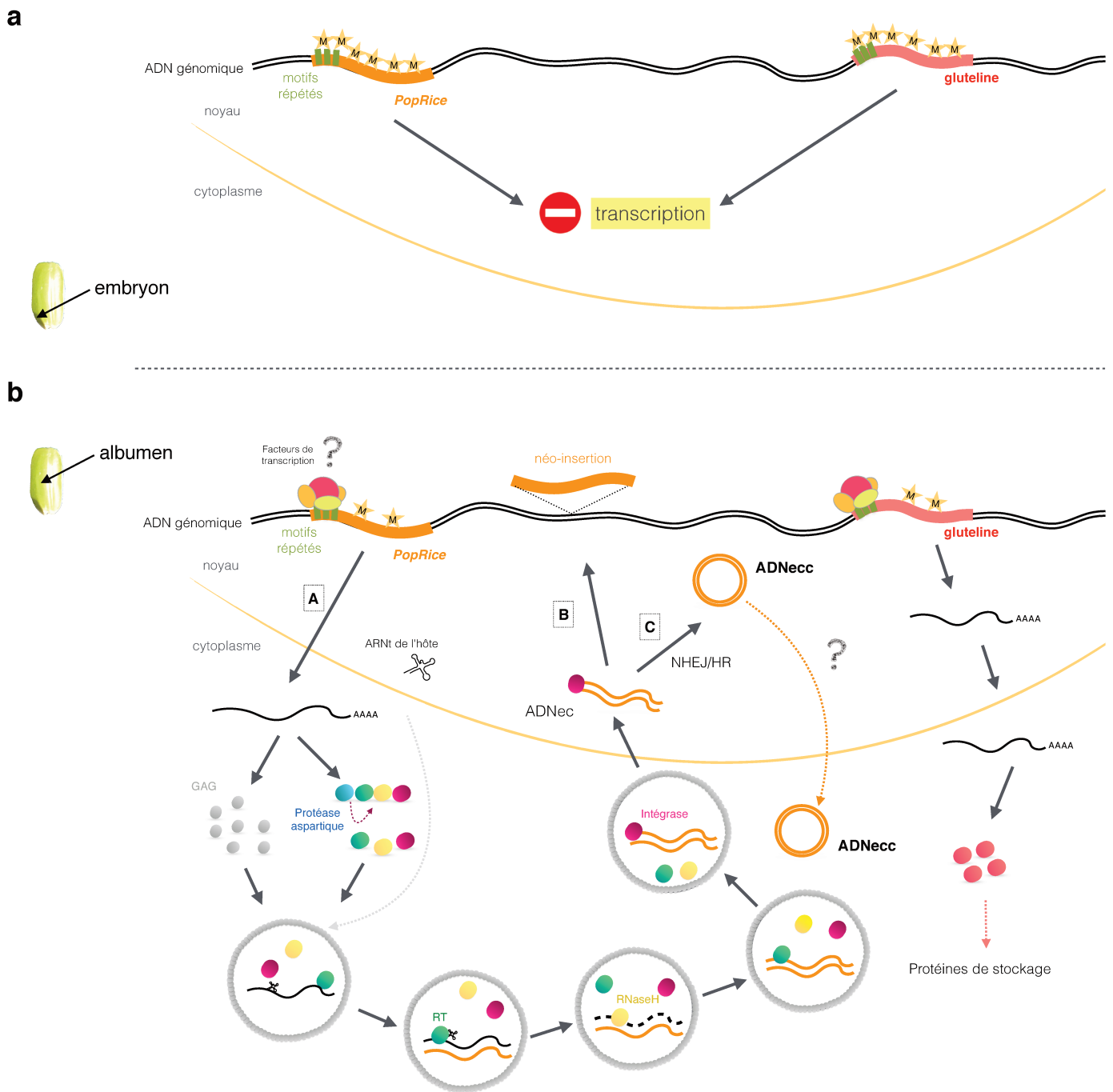


Figure 30. Modèle pour le cycle de vie de *PopRice* dans le grain *Oryza sativa ssp. japonica var. Nipponbare* (a) Dans le tissu embryonnaire, les gènes de la famille des glutélines et les éléments de la famille de *PopRice* sont méthylés inhibant leur transcription (étoiles: méthylation). (b) Dans l'albumen, la déméthylation des gènes de la famille des glutélines et de la famille *PopRice* et la présence de motifs répétés dans leur promoteur induiraient le recrutement de facteurs de transcription permettant l'initiation de leur transcription (A). Les transcrits migreraient dans le cytoplasme pour être traduits en protéines. *PopRice* sous forme d'ADNec double brin rejoindrait ensuite le noyau où nous avons détecté de nouvelles insertions somatiques (B) et des formes circulaires extrachromosomiques (C). Les étapes A, B et C schématisent les résultats que nous avons obtenus.

récemment montré que *PopRice* était sous contrôle épigénétique et que la simple application d'une hormone ou une condition de stress pourrait ne pas suffire à activer *PopRice*. Nos résultats préliminaires actuels ne nous permettent donc pas de conclure sur l'effet de l'ABA sur l'activité de *PopRice*. Toutefois, il serait intéressant de continuer dans cette voie et de tester différentes applications d'hormones et différentes conditions de stress dans des plantes où l'épigénome est perturbé.

Pour finir, le reséquençage du génome de l'albumen a mis en évidence la présence de néo-insertions de *PopRice* dans l'albumen. C'est la première fois que la présence d'insertions somatiques de rétrotransposon est détectée par ce biais chez les plantes, à notre connaissance. Il est toutefois difficile de différencier les faux positifs et les insertions réelles montrant que le reséquençage de génomes ne semble pas être un moyen suffisant à ce jour pour étudier l'activité somatique des ET contrairement au séquençage du mobilome qui est un puissant outil de détection en somatique. Par ailleurs, les insertions de *PopRice* semblent fréquentes dans le génome de l'albumen et, de façon surprenante, *PopRice* ne semble donc pas sujet à des mécanismes de répression. L'acquisition d'éléments *cis* régulateurs pourrait lui avoir donné l'avantage de transposer avec une forte intensité sans qu'aucun mécanisme de *silencing* n'intervienne. Si toutefois, par PCR inverse, aucun cercle d'ADNecc de *PopRice* n'est détecté dans l'embryon, de nombreux échanges se font entre l'albumen et l'embryon (Lopato et al. 2014) et il serait donc intéressant d'étudier plus en détail le mobilome de ce tissu pour regarder si quelques ADNecc de *PopRice* auraient la possibilité de contourner les barrières cellulaires afin de circuler d'un tissu à un autre.

Nos différents résultats obtenus jusqu'ici nous ont permis d'établir un modèle sur le cycle de vie de *PopRice* dans le grain de riz (Figure 30). Dans le tissu embryonnaire, les gènes « spécifiques de l'albumen » et les éléments de la famille de *PopRice* seraient méthylés et non transcrits. En revanche, dans l'albumen, la déméthylation de ces mêmes gènes et des éléments de la famille de *PopRice* et la présence de motifs répétés dans leur promoteur induiraient le recrutement de facteurs de transcription permettant l'initiation de leur transcription. Les transcrits migreraient dans le cytoplasme où ils seraient traduits en protéines. *PopRice* sous forme d'ADNec rejoindrait ensuite le noyau où il pourrait d'une part se réinsérer dans le génome et créer une néo-insertion ou d'autre part être circularisé par le mécanisme de NHEJ. La localisation précise de ces ADNecc n'est pas connue.

Aujourd'hui il est clairement établi que les ET jouent un rôle central dans l'évolution des

génomomes (Lisch 2013; Chuong et al. 2016a). La présence d'éléments *cis* régulateurs, la transposition dans un fond sauvage en condition normale et la conservation de l'activité dans des espèces lointaines font de *PopRice* un ET singulier. À vrai dire, l'activité de *PopRice* ressemble à peu d'autre ET aujourd'hui décrits dans la littérature si ce n'est au rétrotransposon *ONSEN* chez *Arabidopsis* (Ito et al. 2011; Cavrak et al. 2014). Comme mentionné précédemment, *ONSEN* possède des éléments *cis* régulateurs dans son promoteur, il est actif en réponse à un stress thermique et en transposant affecte la transcription des gènes voisins. Très récemment, Nozawa et al. (2017) ont réalisé une analyse phylogénétique chez les plantes crucifères (telles que le chou, le brocoli, etc...) sur l'histoire évolutive des *ONSEN-like elements* (OLE). De façon intéressante, les OLE ont été détectés dans toutes les espèces testées dans l'étude et l'activité transpositionnelle semble conservée chez certaines espèces. Finalement, l'étude d'ET tels qu'*ONSEN* ou *PopRice* ouvre de nouvelles possibilités pour étendre nos connaissances sur la biologie des ET et sur les stratégies mises en place par les ET pour échapper aux contrôles épigénétiques et maintenir leur activité dans les génomes hôtes.








	fond sauvage	épigénome déstabilisé	stress biotique	stress abiotique	candidats
 <i>Arabidopsis thaliana</i>	[blurred]	zébularine + α -amanitin		stress thermique	✓
 <i>Drosophila melanogaster</i>	[blurred]		virus à ARN		✓
		tissu ovarien			✓
 <i>Oryza sativa</i>	[blurred]		virus à ARN		
		mutant <i>dmd1</i>			
	[blurred]	zébularine + α -amanitin			✓
 <i>Zea mays</i>	[blurred]	grains			
 <i>Triticum aestivum</i>	[blurred]			stress thermique	
	[blurred]			stress hydrique	
 <i>Populus tremula alba</i>		mutants RNAi <i>dmd1</i>			✓
		mutants RNAi <i>dmd1</i>		stress hydrique	✓
 <i>Picea abies</i>	[blurred]				✓

Tableau 2. Schématisation des résultats obtenus par nos travaux en collaborations.

Les cases colorées en bleu symbolisent les caractéristiques pour chacun des échantillons étudiés.

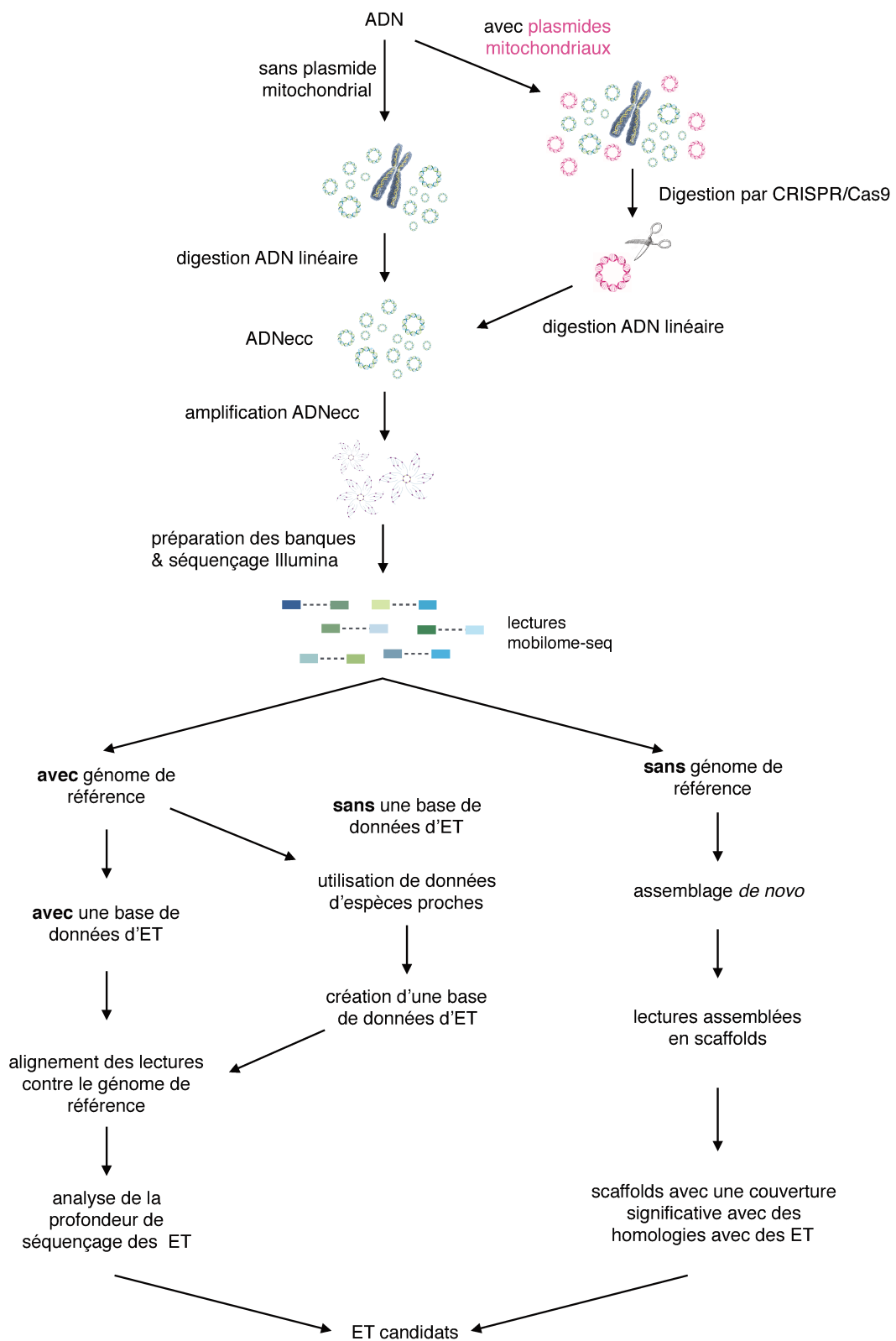


Figure 31. Stratégies d'analyse des données de mobilome-seq en fonction des caractéristiques du génome étudié.

Partie 3. Caractérisation du mobilome chez différents organismes eucaryotes.

La méthode du séquençage du mobilome que nous avons développée est un outil dédié à l'identification des ET actifs et applicable à un grand nombre d'organismes. Si mes premiers travaux ont principalement ciblé l'activité des ET au cours du développement du riz, au fil de nos collaborations, nous avons étendu notre étude du mobilome à d'autres modèles et nous disposons aujourd'hui d'un nombre important de données de mobilome sur une dizaine d'organismes différents dans des fonds mutants ou dans des conditions de stress (Tableau 2 et Tableau S1 – Annexes). Lors du transfert de la technique à d'autres organismes nous avons, dans certains cas, été confrontés à des difficultés particulières à chaque génome, ce qui nous a permis de mettre au point des améliorations (Figure 31). Ces travaux nous ont également permis d'apporter des réponses ou de proposer des pistes pour répondre à différentes questions biologiques que nous exposerons dans ce chapitre. Nous détaillerons notamment deux résultats issus de travaux en collaboration sur le riz et sur la drosophile qui ont fait l'objet de publications.

3.1 Les cousins de *PopRice* (PLE) sont-ils actifs chez le maïs ?

Précédemment, l'analyse de Southern blot (Figure 28c) a confirmé l'absence d'ADNec de ZmPLE dans les grains de maïs. Or, l'analyse par BLAST et par dot-plot des séquences des PLE dans les génomes des céréales a toutefois révélé la répétition de boîtes dans les LTR des PLE (Figure 28a). Si ces boîtes ne semblent pas similaires à des boîtes à glutélines, la présence de ces boîtes suggère néanmoins que les ZmPLE pourraient avoir acquis un profil d'activité différent de *PopRice*. Nous avons appliqué la méthode du mobilome-seq chez le maïs pour tester cette hypothèse.

Nos premiers séquençages du mobilome de maïs ont révélé que 80% des données de séquençage provenaient de plasmides mitochondriaux (Figure 32a,b). La présence de plasmides mitochondriaux de 1,9kb dans les cellules de maïs a été décrite précédemment par Ludwig et al. (1985). Afin d'éliminer ces séquences parasites, nous nous sommes inspirés de la technique CRISPR-DASH (Gu et al. 2016) qui consiste à utiliser l'enzyme Cas9 afin de digérer des séquences abondantes avant un séquençage (Figure 31).

L'enzyme Cas9 est une endonucléase à ADN, guidée par un ARN guide, et capable de cibler et couper la séquence complémentaire de cet ARN guide. Cette protéine a été découverte chez

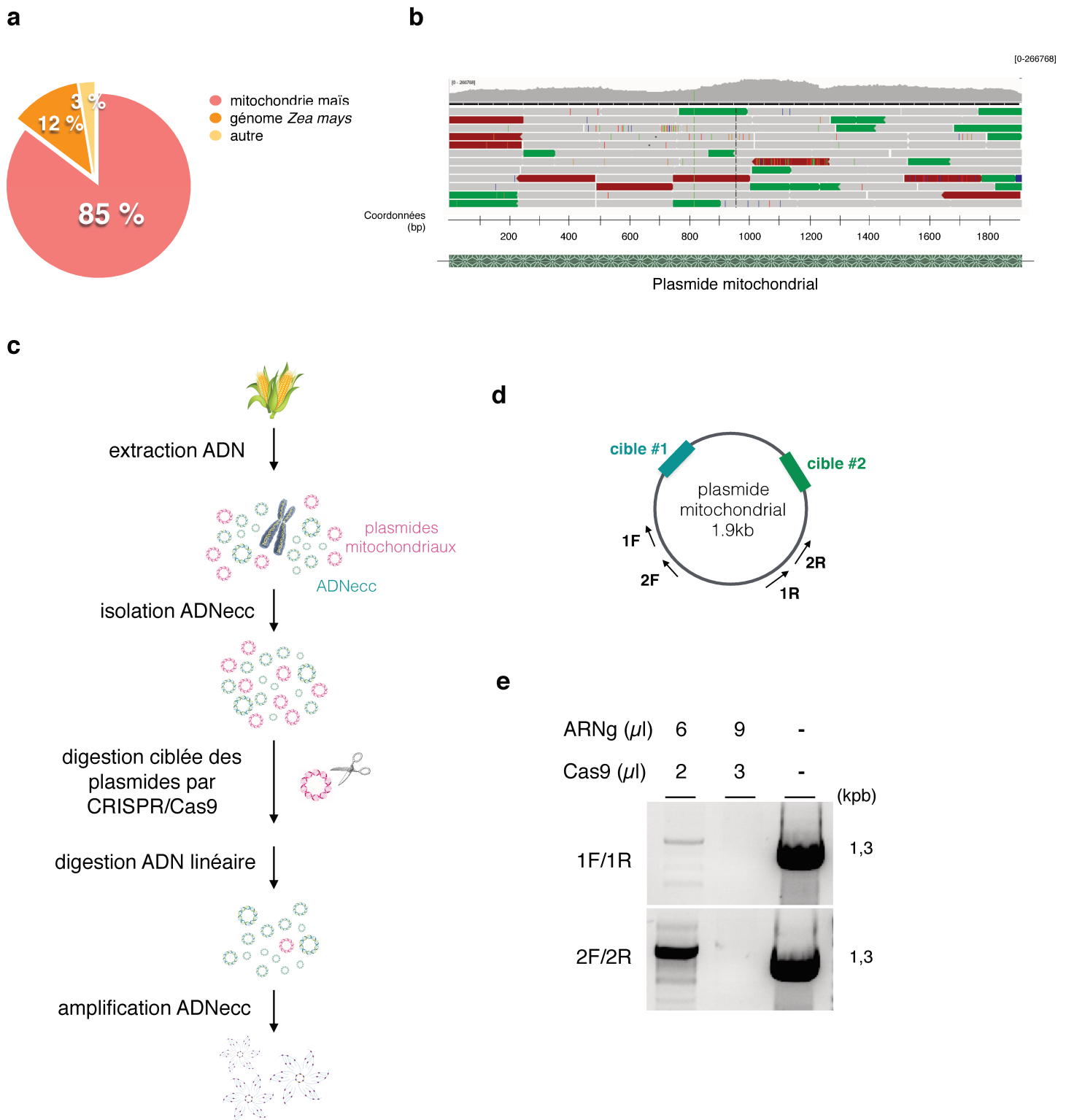


Figure 32. Digestion des plasmides mitochondriaux de maïs par CRISPR/Cas9. (a) Pourcentage de lectures d'une banque de mobilome-seq de feuilles de maïs qui alignent sur les différents génomes (génomme de *Zea mays*, génomme mitochondrial et autres). 85% de la banque correspond à du génomme mitochondrial. (b) Détail de la profondeur de couverture sur la séquence du plasmide mitochondrial. (c) Schéma des différentes étapes réalisées pour la digestion des plasmides mitochondriaux. (d) Positions sur le plasmide mitochondrial des amorces utilisées pour la validation (voir Annexes). (e) Validation par PCR de la présence de plasmides mitochondriaux dans les échantillons digérés et non digérés (piste de droite). Deux concentrations d'ARN guide et d'enzyme Cas9 ont été testées comme indiqué.

Streptococcus pyogenes et participe au système immunitaire adaptatif de type II de CRISPR (*Clustered Regularly Interspaced Short Palindromic Repeats*). Chez les bactéries, le système CRISPR lutte contre les attaques d'ADN étranger (bactériophage, plasmides, virus) en les éliminant spécifiquement en fonction de leurs séquences. La découverte de ce système a totalement révolutionné la biologie et la médecine et offre de puissantes perspectives pour éditer de manière spécifique le génome. Dans notre cas nous avons développé une expérience de digestion *in vitro* des plasmides mitochondriaux à l'aide de l'enzyme CRISPR Cas9. La digestion par CRISPR Cas9 a été réalisée à partir d'ADNecc extraits de feuilles de maïs. Les différentes étapes de l'expérience sont représentées dans la Figure 32c. Afin de maximiser les chances d'éliminer totalement les plasmides mitochondriaux, deux régions du plasmide ont été ciblées par des ARN guides (ARNg). Pour la digestion des plasmides mitochondriaux il est important d'utiliser un ratio 20/20/20/1 (ARNg1/ARNg2/Cas9/ADN cible) pour obtenir une efficacité optimale de l'enzyme Cas9. Néanmoins, étant incapable d'estimer la quantité de plasmides mitochondriaux présents, différents tests de quantité d'ARN guide ont été réalisés. Une réaction (1) a été effectuée avec 6 µl de chaque ARN guide et 2 µl d'enzyme Cas9 pour 1,45 µg d'ADN et une réaction (2) avec 9 µl d'ARN guide et 3 µl d'enzyme Cas9 pour 1,45 µg d'ADN. Suite à la digestion une seconde réaction de digestion des ADN linéaires a été réalisée dans le but d'éliminer les séquences digérées par Cas9. L'ADN circulaire a ensuite été amplifié de la même façon que pour les autres expériences de mobilome-seq.

Après la digestion par CRISPR Cas9 de l'ADN et l'amplification aléatoire des ADNecc, des PCR inverses ont été réalisées pour déterminer si les plasmides mitochondriaux avaient été ciblés, digérés et éliminés de notre échantillon. Deux couples d'amorces différents ont été utilisés pour évaluer avec précision la présence d'ADNecc provenant de ces plasmides mitochondriaux (Figure 32d). Les résultats obtenus montrent que la digestion par CRISPR Cas9 a fonctionné et que la quantité d'ADNecc a diminué dans les deux échantillons digérés comparés à l'échantillon non digéré (Figure 32e). En comparant les deux tests expérimentaux, nous pouvons remarquer que l'ADNecc provenant des plasmides mitochondriaux n'est pas détecté par PCR et semble totalement éliminé de l'échantillon pour lequel une plus grande quantité d'ARN guide (9 µl) et d'enzyme (3 µl) a été utilisée. Ce résultat suggère une grande abondance de ces ADNecc dans le mobilome de maïs. La digestion *in vitro* par CRISPR Cas9 nous a permis d'éliminer spécifiquement des ADNecc qui parasitaient notre analyse et ainsi d'améliorer notre technique du mobilome-seq afin d'étendre l'analyse à des échantillons de

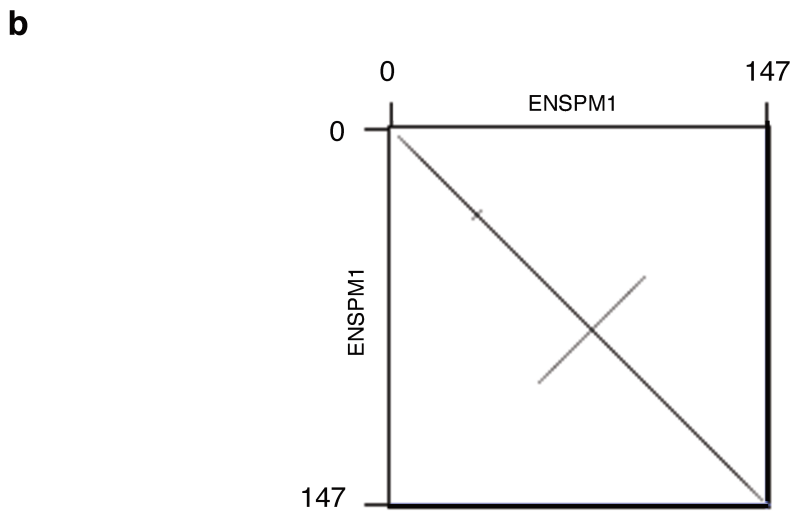
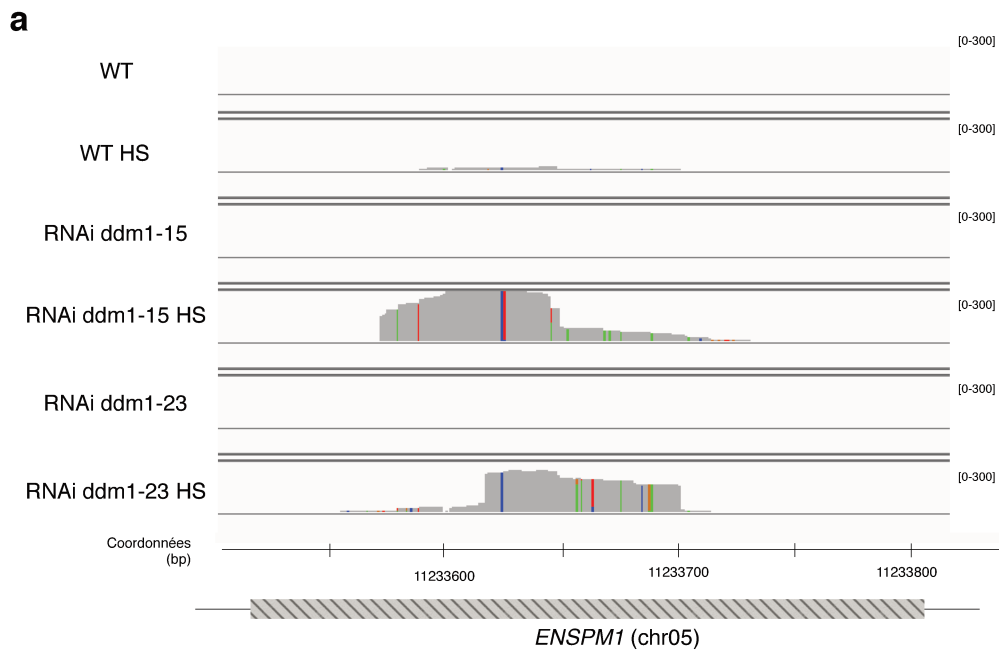


Figure 33. Détection d'ET actif dans les mutants RNAi DDM1 en réponse à un stress hydrique chez le peuplier. (a) Couverture de séquençage au locus de *ENSPM1* sur le chromosome 5 dans les mobilomes de peuplier chez le sauvage (WT) et dans deux lignées RNAi du gène DDM1 (lignées 15 et 23) en condition normale ou en réponse à un stress hydrique (HS). La couverture maximale est indiquée sur la droite. Image obtenue à partir du logiciel IGV. Couverture grise: abondance des lectures (non normalisée); les SNP sont représentés en couleur. **(b)** La séquence nucléotidique de l'élément *ENSPM1* est graphiquement comparée à elle-même. La ligne principale correspond à la séquence alignée contre elle-même. Les répétitions inversées sont représentées comme des lignes perpendiculaires à la ligne diagonale.

mais afin de tester notre hypothèse sur le profil d'activité de *ZmPLE*. Cette étape de digestion peut également être appliquée chez d'autres organismes possédant le même type de plasmides ou d'autres types d'ADNcc parasites. Les banques de mobilome de maïs seront réalisées à partir des échantillons digérés afin de confirmer l'élimination de ces plasmides et d'étudier l'activité des *ZmPLE*.

3.2 La déstabilisation de l'épigénome chez les plantes induit-elle une réactivation massive du mobilome ?

La réponse à cette question avait été précédemment suggérée avec l'analyse du mobilome de grains de riz qui montrait que seul *PopRice* était actif malgré une hypométhylation globale dans l'albumen. Au cours de ma thèse et à deux reprises nous avons pu à nouveau tester cette hypothèse par l'analyse des mobilomes de deux mutants *ddm1* chez deux espèces : le riz et le peuplier *Populus tremula alba* (Tableau 2).

Si chez le riz, aucun candidat intéressant n'a été retenu (non montré), chez le peuplier l'analyse des mobilomes de deux lignées indépendantes de mutants RNAi a permis l'identification de deux transposons à ADN actif chez le mutant contrairement au sauvage. L'étude a également été effectuée après l'application d'un stress hydrique et a révélé l'activité de deux transposons à ADN, différents de la première analyse (Figure 33a). Un des deux transposons à ADN identifiés semble appartenir à la famille *ENSPM* (Figure 33b). Cet élément présente une structure typique dite « en croix » autrement dit des répétitions inversées au sein de sa séquence interne (Figure 33b). Ces résultats sont en cours de validation moléculaire. De plus l'effet de la diminution de l'expression du gène *DDMI* sur le niveau de méthylation global du génome est également en cours d'analyse (Stéphane Maury, INRA d'Orléans). Si aujourd'hui ces résultats restent préliminaires, l'application de la technique du mobilome-seq chez le peuplier nous a permis le développement d'une expertise en amont de l'analyse même des données de mobilome-seq. Effectivement, différentes données d'annotations étaient disponibles pour l'espèce *Populus trichocarpa* contrairement à l'espèce que nous avons étudiée (*P. tremula alba*). Dans un premier temps, nous avons donc utilisé les données de cette espèce proche pour construire une base de données d'ET pour l'espèce *P. tremula alba* (Figure 31) (voir Matériel et Méthodes page 76). La base de données de l'espèce proche est alignée contre le génome de référence de l'espèce étudiée par BLAST et différents

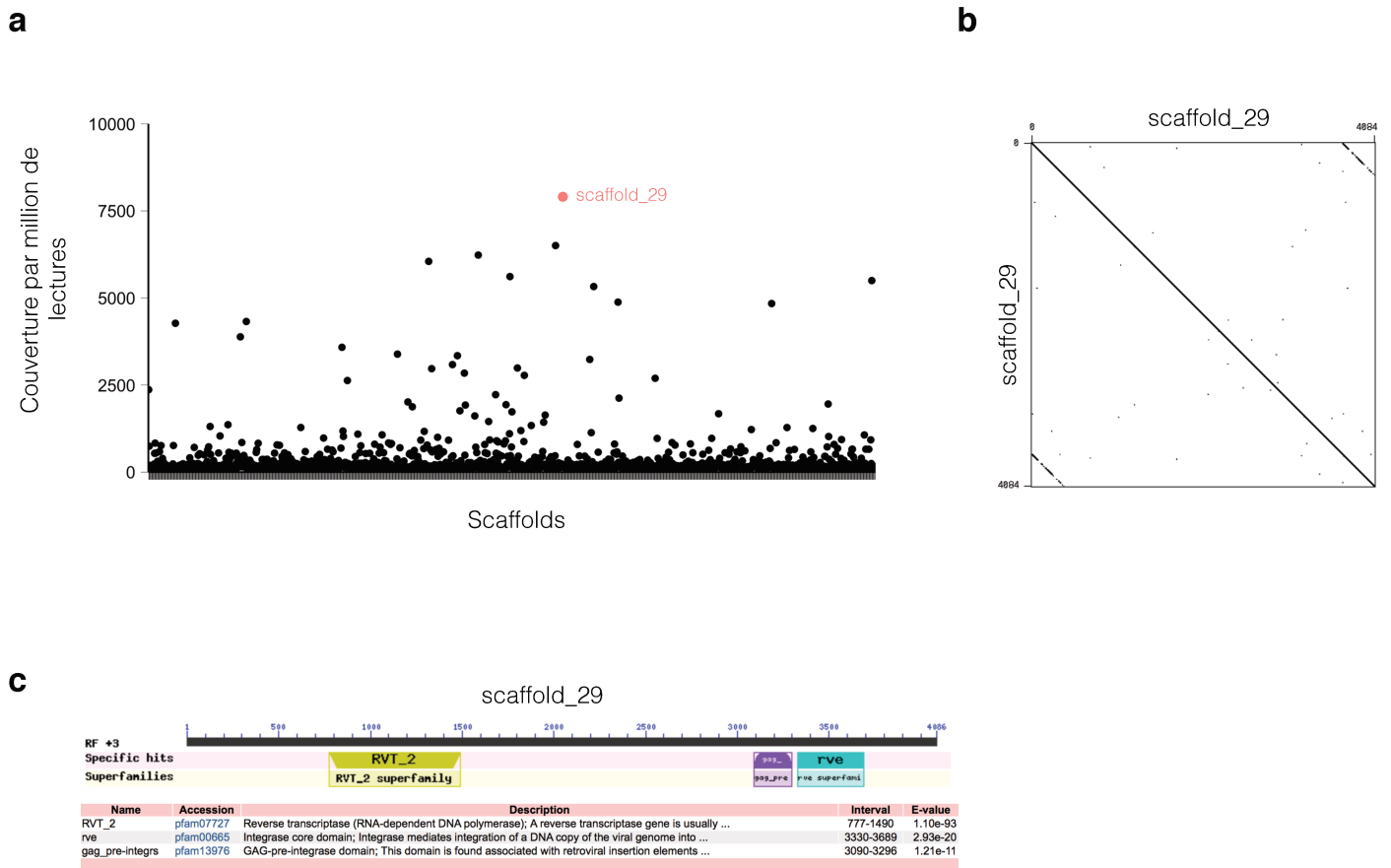


Figure 34. Détection d'ET actif dans le mobilome des bourgeons floraux de *Picea abies* par assemblage *de novo*. (a) Couverture de séquençage de chaque scaffold assemblé *de novo*. Chaque point noir correspond à un scaffold et le nombre de lectures normalisé par million de lectures qui lui est associé. Le scaffold_29 est le scaffold le plus couvert avec 7928 lectures par million de lectures. (b) Analyse par dot-plot du scaffold_29. Ce dot-plot montre que le scaffold_29 a la structure d'un RT à LTR avec ses deux extrémités répétées. (c) Analyse des domaines conservés du scaffold_29. Trois domaines caractéristiques d'un RT à LTR sont identifiés (transcriptase réverse, GAG, integrase) (NCBI).

filtres sont appliqués pour réduire les petits fragments. Dans un second temps, des bases de données publiques telles que NCBI ou Repeat Masker ont été utilisées pour valider la présence de signatures caractéristiques d'ET (domaines conservés, séquences répétées de type LTR ou TIR) dans les régions du génome qui présentent une forte couverture de séquençage.

Cette collaboration nous a notamment permis de montrer que l'identification d'ET actifs par la technique du mobilome-seq ne nécessitait ni de fichier d'annotation, ni même de base de données propre à l'espèce. De façon importante, une dérégulation massive des ET dans le mobilome n'a pas été observée dans les lignées mutantes *DDMI*.

3.3 Les très gros génomes ont-ils un mobilome plus actif ?

Nous avons illustré dans l'introduction la corrélation entre la taille d'un génome et son contenu en ET (Figure 1). Finalement, l'activité du mobilome est-elle aussi corrélée à la proportion en ET dans un génome ? Pour tenter de répondre à cette question nous avons réalisé une analyse du mobilome de différents tissus extraits à partir de l'épicéa (*Picea abies*) (Tableau 2 et Table S1 - Annexes).

L'épicéa est le premier Gymnosperme séquencé avec un génome de 20 Gb composé de plus de 70% d'ET. Aujourd'hui environ 12 Gb ont été séquencées et assemblées en environ 10 millions de scaffolds (Nystedt et al. 2013). Des analyses de méthylome ont montré que le niveau de méthylation de l'épicéa était très élevé en comparaison avec les autres plantes (Ausin et al. 2016). Nous avons séquencé le mobilome des aiguilles du pin et des bourgeons floraux pour lesquels le niveau de méthylation est similaire (Ausin et al. 2016). Le logiciel d'alignement que nous utilisons, Bowtie2 (Langmead et Salzberg 2012), ne permet pas d'aligner des lectures sur un génome de référence pour lequel il y a un aussi grand nombre de scaffolds. À partir de nos données de séquençage, nous avons développé une analyse sans génome de référence (Figure 31). Les lectures ont été assemblées en scaffolds, autrement dit chacun des scaffolds représente un cercle d'ADN présent dans le mobilome. Pour chaque scaffold assemblé, la couverture de séquençage est déterminée. Le scaffold_29, le plus couvert dans la banque du mobilome (7928 lectures par million de lectures) des bourgeons floraux correspond à un rétrotransposon à LTR (Figure 34a). En effet, l'alignement de la séquence du scaffold contre elle-même montre la structure caractéristique d'un rétrotransposon avec ses deux LTR à chaque extrémité du scaffold (Figure 34b) et par

recherche des domaines conservés, les domaines de la GAG, de l'intégrase et de la transcriptase réverse ont été identifiés (Figure 34c). De plus, par analyse BLAST sur le génome de *Picea abies* il apparaît que cet élément semble appartenir à une petite famille d'ET avec peu de copies dans le génome (<10). Afin de confirmer l'activité de cet élément dans les bourgeons floraux, une validation par Southern blot sera nécessaire.

Aujourd'hui nous nous sommes seulement intéressées aux scaffolds les plus couverts et au regard des résultats actuels, nous ne pouvons pas confirmer ou infirmer notre hypothèse sur la corrélation entre l'abondance des ADNec et la taille d'un génome. L'étude approfondie de ces données pourra apporter des pistes de réponses. Néanmoins, un des résultats majeurs de ces travaux en collaboration est que nous pouvons détecter l'activité somatique d'ET sans génome de référence à l'aide la technique du mobilome.

3.4 L'ARN polymérase II est-elle impliquée dans le contrôle des ET ?

Nous avons à plusieurs reprises souligné le rôle des mécanismes épigénétiques tels que la méthylation de l'ADN dans le contrôle des ET. Les acteurs moléculaires impliqués dans le maintien de la méthylation de l'ADN et dans la mise en place de la méthylation *de novo* ont été largement caractérisés chez *Arabidopsis*. Néanmoins, l'activité d'un rétrotransposon dépend également de son activité transcriptionnelle (voir Introduction page 9) et donc par conséquent de l'activité de l'ARN polymérase II (Pol II). Or, le rôle de Pol II dans la régulation des ET n'est pas clairement défini.

Afin de caractériser le rôle de Pol II dans le contrôle des ET, l'équipe d'Étienne Bucher (INRA d'Angers) a développé une stratégie visant à inhiber l'activité de Pol II à l'aide d'une drogue pharmaceutique, l' α -amanitine, dans différents fonds génétiques et dans différentes conditions, normales ou après un stress thermique chez *Arabidopsis*. L'étude montre qu'en réponse à un stress, la réduction de l'activité de Pol II a un effet drastique d'hypométhylation en CHH au niveau des ET et augmente la production d'ADNec. De plus, la combinaison entre l'inhibition de Pol II et l'inhibition de la méthylation de l'ADN (traitement à la zébularine) provoque un *burst* d'activité transpositionnelle du rétrotransposon *ONSEN* après un stress thermique chez *Arabidopsis thaliana*.

J'ai contribué à cette étude en analysant le mobilome de riz traité par le double traitement α -


amanitine et zébularine. J'ai montré que ce traitement induisait la production d'ADNecc du rétrotransposon *Houba* et j'ai contribué à valider ces données par PCR. En conclusion, la répression de la Pol II induit des *bursts* de rétrotransposition et ouvre des perspectives nouvelles dans le but de créer de la diversité génétique et épigénétique qui peut ensuite être utilisée pour l'agriculture. Les résultats obtenus ont fait l'objet d'une publication dans *Genome Biology*.

RESEARCH

Open Access



Inhibition of RNA polymerase II allows controlled mobilisation of retrotransposons for plant breeding

Michael Thieme¹, Sophie Lanciano^{2,3}, Sandrine Balzergue⁴, Nicolas Daccord⁴, Marie Mirouze^{2,3} and Etienne Bucher^{4*} 

Abstract

Background: Retrotransposons play a central role in plant evolution and could be a powerful endogenous source of genetic and epigenetic variability for crop breeding. To ensure genome integrity several silencing mechanisms have evolved to repress retrotransposon mobility. Even though retrotransposons fully depend on transcriptional activity of the host RNA polymerase II (Pol II) for their mobility, it was so far unclear whether Pol II is directly involved in repressing their activity.

Results: Here we show that plants defective in Pol II activity lose DNA methylation at repeat sequences and produce more extrachromosomal retrotransposon DNA upon stress in *Arabidopsis* and rice. We demonstrate that combined inhibition of both DNA methylation and Pol II activity leads to a strong stress-dependent mobilization of the heat responsive *ONSEN* retrotransposon in *Arabidopsis* seedlings. The progenies of these treated plants contain up to 75 new *ONSEN* insertions in their genome which are stably inherited over three generations of selfing. Repeated application of heat stress in progeny plants containing increased numbers of *ONSEN* copies does not result in increased activation of this transposon compared to control lines. Progenies with additional *ONSEN* copies show a broad panel of environment-dependent phenotypic diversity.

Conclusions: We demonstrate that Pol II acts at the root of transposon silencing. This is important because it suggests that Pol II can regulate the speed of plant evolution by fine-tuning the amplitude of transposon mobility. Our findings show that it is now possible to study induced transposon bursts in plants and unlock their use to induce epigenetic and genetic diversity for crop breeding.

Keywords: Epigenetics, DNA methylation, Genome integrity, Evolution, *Oryza sativa*, *Arabidopsis thaliana*

Background

Like retroviruses, long terminal repeat (LTR) retrotransposons (class I elements), which represent the most abundant class of transposable elements (TEs) in eukaryotes, transpose via a copy and paste mechanism. This process requires the conversion of a full length RNA polymerase II (Pol II) transcript into extrachromosomal complementary DNA (ecDNA) by reverse transcription [1]. In their life cycle LTR retrotransposons can produce extrachromosomal circular DNA (eccDNA), which is an

indicator for their ongoing activity [2]. In plants, TEs are increasingly seen as a source of genetic and epigenetic variability and thus important drivers of evolution [3–6]. However, plants have evolved several regulatory pathways to retain control over the activity of these potentially harmful mobile genetic elements. Cytosine methylation (^mC) plays a central role in TE silencing in plants [7]. In addition, plants have evolved two Pol II-related RNA polymerases, Pol IV and Pol V, that are essential to provide specific silencing signals leading to RNA-directed DNA methylation (RdDM) at TEs [8], thereby limiting their mobility [9–11]. More recently, various additional non-canonical Pol IV-independent RdDM pathways have been described [12]. Notably it was found that Pol II

* Correspondence: etienne.bucher@inra.fr

⁴IRHS, Université d'Angers, INRA, AGROCAMPUS-Ouest, SFR4207 QUASAV, Université Bretagne Loire, 49045 Angers, France

Full list of author information is available at the end of the article



© The Author(s). 2017 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated.

itself also plays an important role in RdDM [13, 14] by feeding template RNAs into downstream factors such as RNA-DEPENDENT RNA POLYMERASE 6 (RDR6), resulting in dicer-dependent or -independent initiation and establishment of TE-specific DNA methylation [15]. Beyond that, recent work suggests a new “non-canonical” branch of RdDM that specializes in targeting transcriptionally active full-length TEs [16]. This pathway functions independently of RDRs via Pol II transcripts that are directly processed by DCL3 into small interfering RNAs (siRNAs).

Results

Here, we wanted to investigate if Pol II could play a direct role in repressing TE mobility in plants. For this purpose we chose the well-characterized heat-responsive *copia*-like *ONSEN* retrotransposon [11] of *Arabidopsis* and took advantage of the hypomorphic *nrbp2-3* mutant allele that causes reduced NRPB2 (the second-largest component of Pol II) protein levels [14]. Using quantitative real-time PCR (qPCR), we determined that challenging *nrbp2-3* seedlings by heat stress (HS) led to a mild increase

in total *ONSEN* copy number (sum of ecDNA, eccDNA and new genomic insertions) relative to control stress (CS) and compared to the wild type (WT) (Fig. 1a). This result is supported by the observed dose-responsive increase in *ONSEN* copy number after HS and pharmacological inactivation of Pol II with α -amanitin (A), a potent Pol II inhibitor [17] that does not affect Pol IV or Pol V [18] (Fig. 1b). In order to test the interaction between Pol II-mediated repression of TE activation and DNA methylation, we grew WT and *nrbp2-3* plants on media supplemented with zebularine (Z), an inhibitor of DNA methyltransferases active in plants [19], and subjected them to HS. To ensure the viability of the *nrbp2-3* seedlings we choose a moderate amount of Z (10 μ M). The presence of Z in the medium during HS generally enhanced the production of *ONSEN* copies. Importantly, this induced increase in *ONSEN* copy number was more distinct in the *nrbp2-3* background (Fig. 1a). This indicated that both DNA methylation and Pol II transcriptional activity contribute to the repression of *ONSEN* ecDNA production. To complete their lifecycle, the reverse transcribed ecDNA of activated retrotransposons

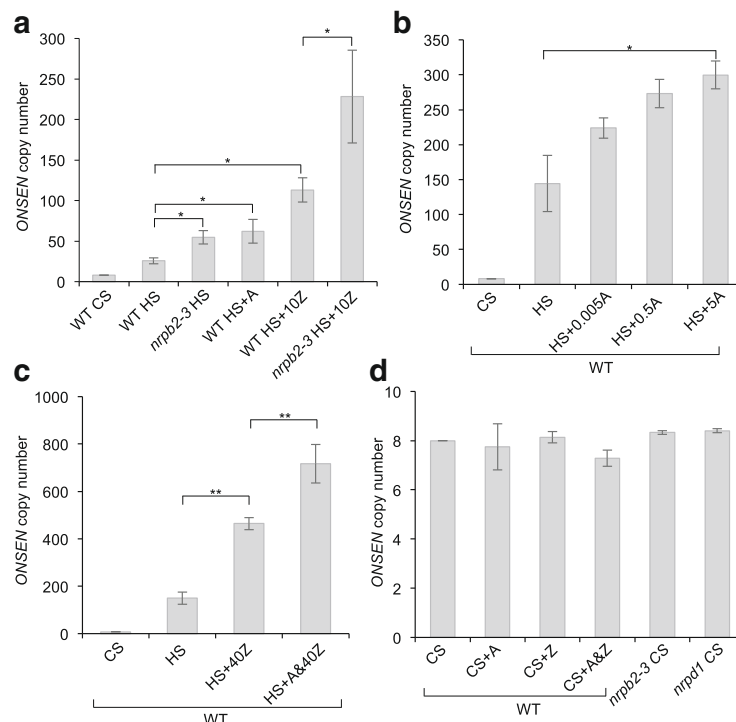
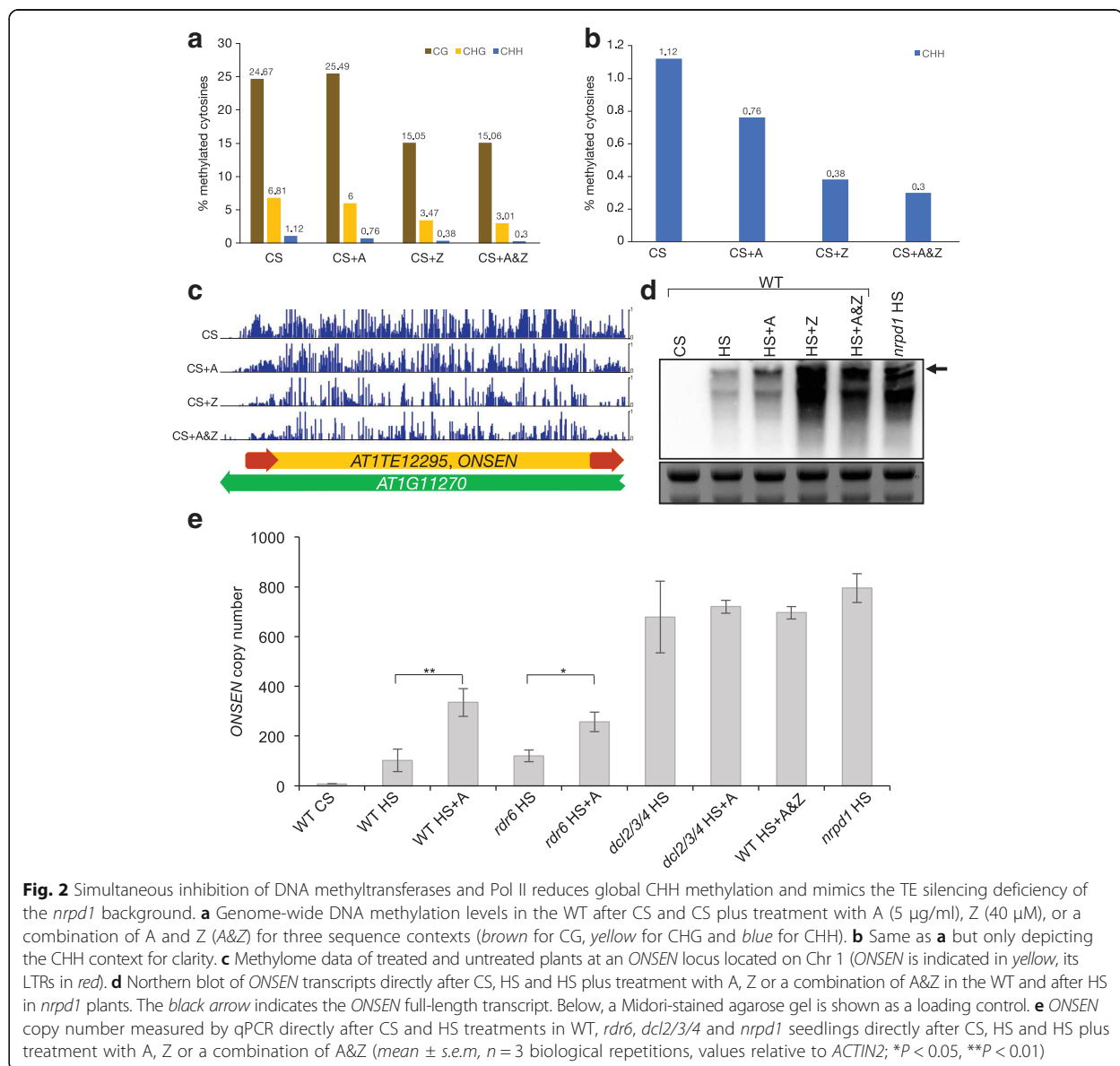


Fig. 1 Pol II represses the HS-dependent mobility of the *ONSEN* retrotransposon in *Arabidopsis*. *ONSEN* copy number in *Arabidopsis* seedlings measured by qPCR directly after CS and HS treatments. **a** In the WT and the *nrbp2-3* mutant and after HS plus treatments with α -amanitin (A; 5 μ g/ml) or zebularine (Z; 10 μ M) (mean \pm standard error of the mean (s.e.m.), $n = 6$ biological repetitions). **b** In the WT and after HS plus treatment with A at different concentrations (μ g/ml) as specified on the x-axis (mean \pm s.e.m., $n = 4$ biological repetitions). **c** In the WT and after HS plus treatment with Z (40 μ M) or a combination of A (5 μ g/ml) and Z (A&40Z) (mean \pm s.e.m., $n = 3$ biological repetitions). **d** In the WT after chemical treatment with A (5 μ g/ml), Z (40 μ M), a combination of A and Z (A&Z) or in the *nrbp2-3* and *nrpd1* backgrounds following CS (mean \pm s.e.m., $n = 3$ biological repetitions). All values are relative to *ACTIN2*. * $P < 0.05$, ** $P < 0.01$

has to integrate back into the genome [1]. Given that we observed a strong increase in *ONSEN* copy number after HS and treatment with moderate amounts of Z in the *nrbp2-3* background, we wanted to address the inheritance of additional *ONSEN* copies by the offspring. For this we compared the average *ONSEN* copy number of pooled S1 seedlings obtained from Z-treated and heat-stressed WT and *nrbp2-3* plants grown under controlled conditions on soil by qPCR. We observed a distinct increase in the overall *ONSEN* copy number exclusively in the *nrbp2-3* background (Additional file 1: Figure S1).

Because both DNA methylation and Pol II can be inhibited by the addition of specific drugs, we wanted to test if treating WT plants with both A and Z at the same

time could strongly activate and even mobilize *ONSEN* after a HS treatment. We grew WT seedlings on MS medium supplemented with Z (40 μ M) [19] individually or combined with A (5 μ g/ml, A&Z). Consistent with the strong activation of *ONSEN* in HS and Z-treated *nrbp2-3* seedlings, the combined treatment (A&Z) of the WT gave rise to a very high (Fig. 1c) HS-dependent (Fig. 1d) increase in *ONSEN* copy number, comparable to that in the *nrbp1* background (Fig. 2e). We noted that the overall amplitude of HS-dependent *ONSEN* activation could vary between different waves of stress applications in terms of copy number (Fig. 1a, b). Yet, the observed enhancing effect of Pol II and DNA methyltransferase inhibition with A and Z on *ONSEN* activation was consistent in



independent experiments (Figs. 1a–c and 2e). To detect activated TEs at the genome-wide level we took advantage of the production of eccDNA by active retrotransposons. eccDNA is a byproduct of the LTR retrotransposon life cycle [20]. Using mobilome sequencing, which comprises a specific amplification step of circular DNA followed by high-throughput sequencing to identify eccDNA derived from active LTR retrotransposons [2], we found that only *ONSEN* was activated by HS in combination with A&Z (Additional file 1: Figure S2). Confirming our qPCR data, more *ONSEN*-specific reads were detected in the presence of A and Z in the medium.

To better understand the mechanisms by which the drugs enhanced the activation of *ONSEN* after HS at the DNA level, we assessed how they influenced DNA methylation at the genome-wide level using whole-genome bisulfite sequencing (WGBS) after CS. Overall, we found that all drug treatments affected global DNA methylation levels. While the treatment with Z affected all sequence contexts, we observed that inhibition of Pol II primarily affected cytosine methylation in the CHG and CHH sequence contexts (where H is an A, T or G). The combined A&Z treatment had a slight additive demethylating effect in the CHG and CHH contexts compared to A or Z alone (Fig. 2a, b). DNA methylation levels at one *ONSEN* locus (*ATITE12295*) is depicted in Fig. 2c. Treatment with A led to a slight decrease in DNA methylation, which was more apparent in Z- and A&Z-treated plants. We then checked by northern blot whether the degree of reduction in DNA methylation would coincide with increased *ONSEN* transcript levels directly after HS. We found that treatment with Z alone resulted in the highest *ONSEN* transcript level after HS (Fig. 2d). Considering the data obtained on *ONSEN* ecDNA (Fig. 1c), we concluded that a substantial proportion of these Z-induced transcripts were not suitable templates for *ONSEN* ecDNA synthesis.

In *Drosophila*, it has been shown that Pol II-mediated antisense transcription results in the production of TE-derived siRNAs in a Dicer-2-dependent manner [21]. In support of this in *Arabidopsis*, a recent publication pointed out the importance of DCL3 in regulating *ONSEN* in the *ddm1* background [16]. To elucidate whether the effect of Pol II inhibition was also dicer-dependent, we grew both *rdm6* and *dcl2/3/4* triple mutant plants on A, applied HS and measured *ONSEN* ecDNA levels. Strikingly, we found that A still enhanced ecDNA accumulation in *rdm6* plants, whereas inhibition of Pol II had no additional effect in the *dcl2/3/4* triple mutant (Fig. 2e).

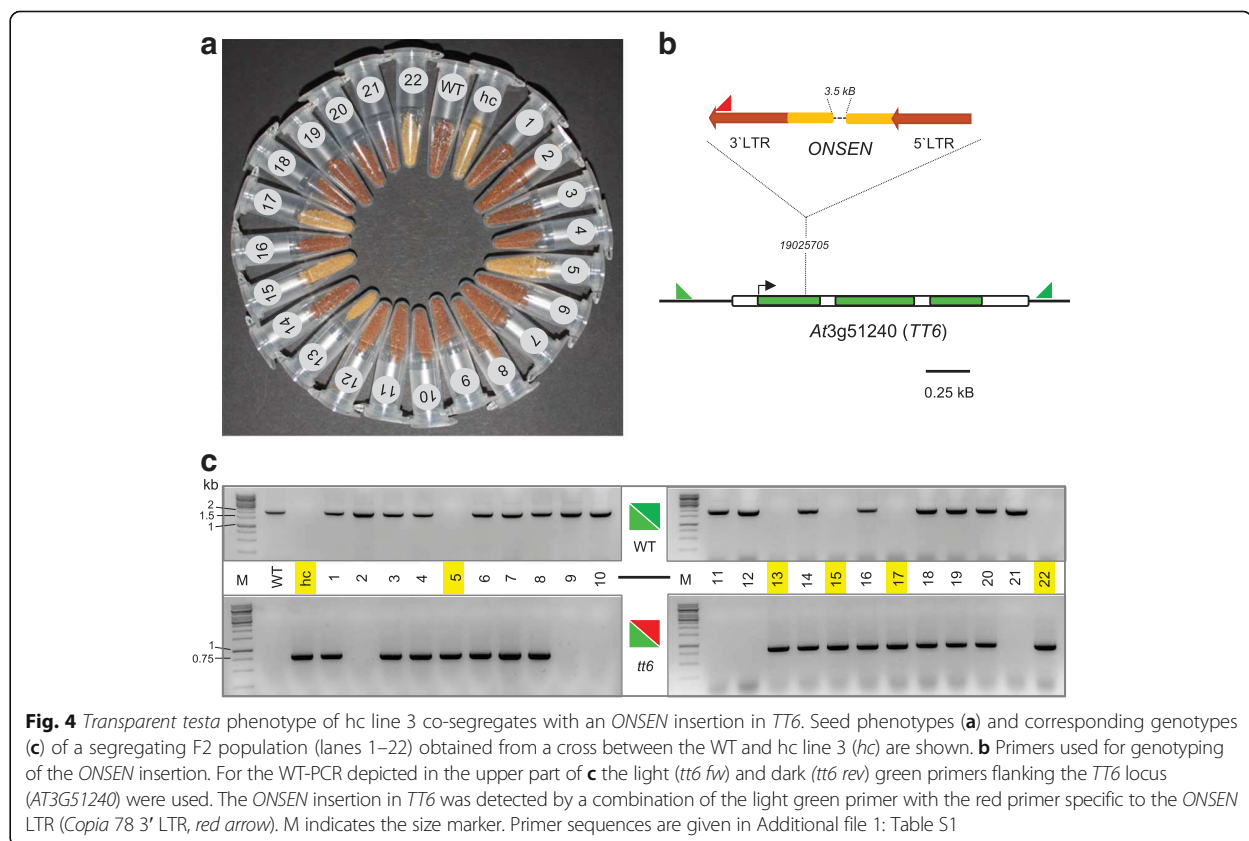
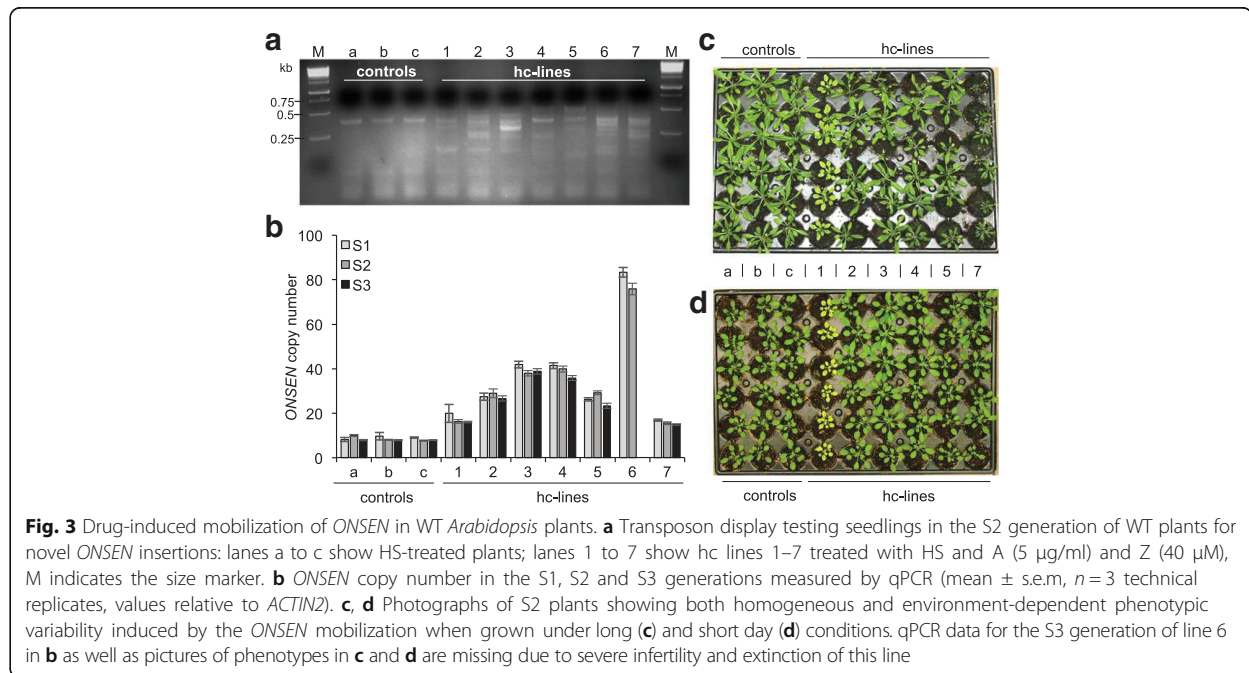
Induced mobilization of endogenous TEs in plants has so far been very inefficient, thus limiting their use in basic research and plant breeding [3]. In the case of *Arabidopsis*, transposition of *ONSEN* in HS-treated WT plants has not been observed [11, 22]. Because the A&Z

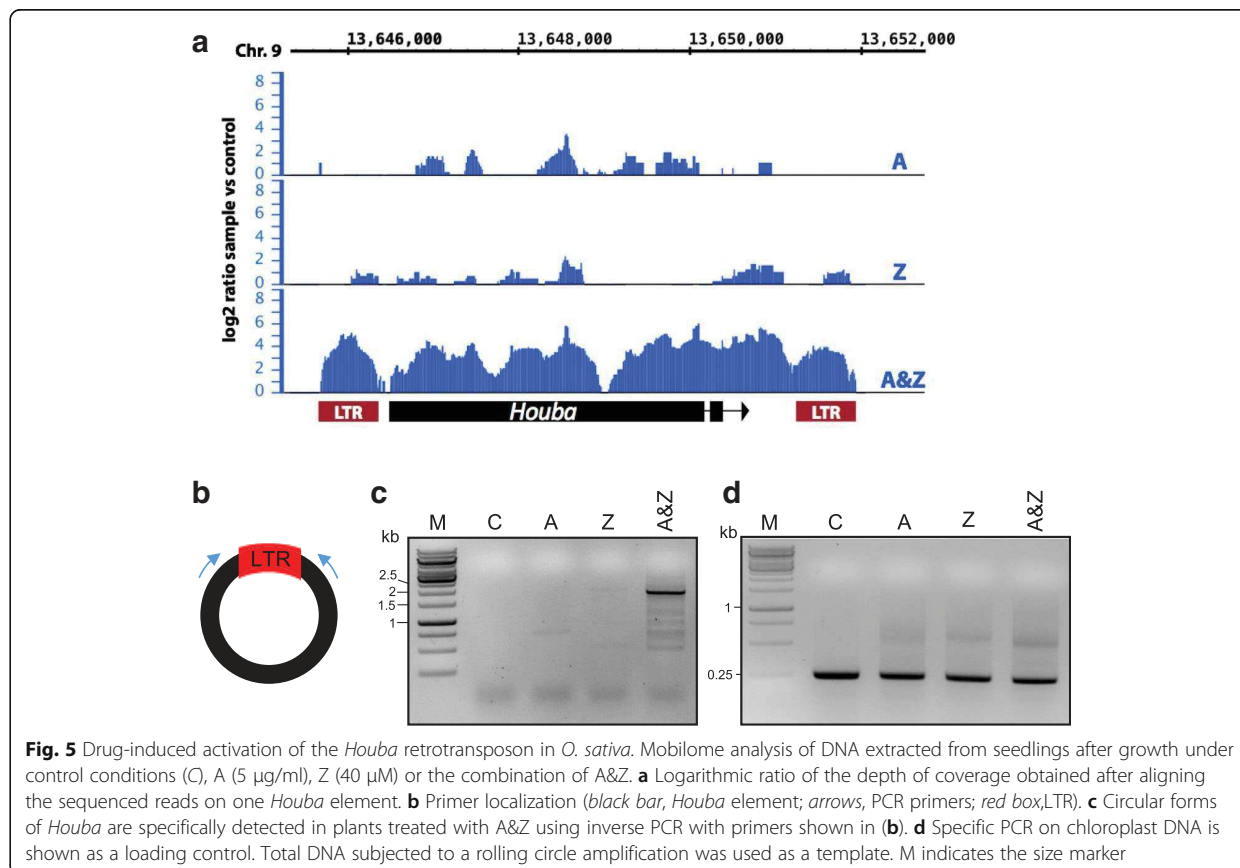
drug treatment resulted in high accumulation of *ONSEN* copy numbers—essentially mimicking plants defective in NRDP1 (Fig. 2e)—we wanted to test if the combined drug treatment could lead to efficient *ONSEN* mobilization in WT plants. First, we assessed by qPCR if, and at what frequencies, new *ONSEN* copies could be detected in the progeny of A&Z-treated and heat stressed plants. In fact, we found new *ONSEN* insertions in 29.4% of the tested S1 (selfed first generation) pools (n = 51), with pools having up to 52 insertions (Additional file 1: Figure S3). We then confirmed stable novel *ONSEN* insertions in a subset of independent individual high copy plants by transposon display (Fig. 3a), qPCR (Fig. 3b) and sequencing of 11 insertions in a selected high-copy line (hc line 3; Fig. 4; Additional file 1: Figure S4). Tracking *ONSEN* copy numbers over three generations of selfing indicated that the new insertions were stably inherited (Fig. 3b). Furthermore, the re-application of heat stress and drugs in the S3 generation of two hc lines did not lead to greater accumulation of *ONSEN* copies compared to control lines, but we instead observed stronger silencing in lines with more *ONSEN* copies (Additional file 1: Figure S5).

TE insertions can interrupt genes or alter their expression by recruiting epigenetic marks or by stress-dependent readout transcription from the 3' LTR into flanking regions [6]. To test this, we grew the S2 generation of the selected hc lines under long- and short-day conditions. Interestingly, we observed that many hc lines showed clear and homogenous phenotypes in response to the different growth conditions (plant size, chlorophyll content and flowering time; Fig. 3c, d).

To demonstrate that *ONSEN* insertions could directly influence such developmental phenotypes, we closely investigated hc line 3, which produced white seeds (Fig. 4a). Using a candidate gene approach, we found that an *ONSEN* insertion in *transparent testa 6* (*TT6*, *AT3G51240*; Fig. 4b) was responsible for the recessive white seed phenotype [23, 24]. This was confirmed by segregation analysis of the F2 generation of a cross between WT and hc line 3 (Fig. 4a) followed by genotyping (Fig. 4c).

Next, we wanted to test if Pol II plays a more general role in repressing TEs in plants. Due to its significantly different epigenetic and TE landscape compared to *Arabidopsis*, we wanted to test if we could mobilize TEs in rice (*Oryza sativa*) [25], a genetically well-characterized monocotyledonous crop. To capture drug-induced mobilized TEs, we characterized the active mobilome in *O. sativa* seedlings that were grown on MS medium supplemented with no drugs, A only, Z only or a combination of A and Z, using the same approach as we used for *Arabidopsis*. We identified *Houba*, a copia-like retrotransposon [26], as highly activated only when plants were treated with A&Z (Fig. 5a). Bona fide activity of





Houba was supported by the detection of eccDNA containing LTR–LTR junctions (Additional file 1: Figure S6). The activation of *Houba* was further confirmed by eccDNA-specific PCR on the *Houba* circles (Fig. 5b–d).

Discussion

In this study, we show the importance of Pol II in the repression of TE mobility in plants. By choosing the well-characterized heat inducible *ONSEN* retrotransposon, we were able to specifically address the role of Pol II in silencing transcriptionally active endogenous TEs in WT plants. Recent studies propose Pol II as the primary source for the production of TE-silencing signals that can then feed into the RNA silencing and DNA methylation pathways [15]. Our data strongly support these findings at two levels. First, we found that inhibition of Pol II activity reduced the degree of DNA methylation at *ONSEN*, demonstrating its distinct role in this process, and that Pol II also contributes to reinforcing silencing at the genome-wide level, primarily in the CHH but also in the CHG context. Second, our finding that DCL enzymes are sufficient to process the silencing signal produced by Pol II suggest that Pol II acts at very early steps in the TE silencing pathway by providing substrates

to these enzymes. The observation that inhibition of Pol II in the *rdr6* background still further enhanced *ONSEN* accumulation after HS supports the notion that Pol II plays a central role in the previously proposed expression-dependent RdDM pathway [16].

Using mobilome sequencing we confirmed previous findings [2] that this approach is a powerful diagnostic tool to detect mobile retrotransposons: we detected highest levels of eccDNA of *ONSEN* in HS and drug-treated *Arabidopsis* seedlings and found new insertions in successive generations of these plants. Using the same approach on rice we were able to detect production of *Houba* eccDNA after drug treatments, suggesting that the progeny will then contain novel *Houba* insertions. This is still to be confirmed and may be hampered by the already very high *Houba* copy number present in the genome [27].

Our findings may indicate that Pol II is primarily involved in silencing young, recently active retrotransposons and perhaps to a lesser extent other tightly silenced TEs. Indeed, there are indications of very recent natural transposition events for *ONSEN* [28] and *Houba* [29] in the *Arabidopsis* and rice genomes, respectively. For instance, the annual temperature range has and may still contribute to contrasting *ONSEN* mobilization events in different *Arabidopsis* accessions [28]. *Houba* is the most

abundant TE of the *copia* family in rice and has been active in the last 500,000 years [30].

Overall, our findings lead to the question of when plants lower their guard: under what conditions could Pol II be less effective in silencing TEs? Certain stresses that affect the cell cycle have been reported to lead to the inactivation of Pol II [31, 32]; this would provide a window of opportunity for TEs to be mobilized. Therefore, combined stresses that affect the cell cycle and activate TEs may lead to actual TE bursts under natural growth conditions. Interestingly, it has been reported that retrotransposon-derived short interspersed element (SINE) transcripts can inhibit Pol II activity [33]. This strongly suggests the presence of an ongoing arms race between retrotransposons and Pol II. Considering that almost all organisms analyzed so far have TEs [4] and RNA polymerases [34] and the reliance of TEs on host RNA polymerases, it may—from an evolutionary point of view—not come as a surprise that Pol II also has a function as an important regulator of retrotransposon activity. Strikingly, it has been shown in both *Saccharomyces cerevisiae* and *Drosophila melanogaster* that Pol II-dependent intra-element antisense transcription plays an important role in TE silencing [21, 35]. In addition, we observed a discrepancy in *ONSEN* transcript accumulation and measured ecDNA after HS in seedlings that were treated with zebularine only. This substantiates the notion that both the quantity and quality of transcripts affect regulation, reverse transcription and successful integration of retrotransposons. This is well in line with previous observations demonstrating that different TE-derived transcripts have distinct functions in the regulation of TE activity [36]. As a next step it will be of great interest to investigate if Pol II-dependent antisense transcription of TEs and subsequent dicer-dependent processing may be the key to solve “the chicken and the egg problem” of *de novo* silencing functional retrotransposons in eukaryotes.

Finally, our findings will allow future studies on the potential beneficial role TEs play in adaptation to stresses. Indeed, two recent studies point out the adaptive potential of retrotransposon and, more specifically, *ONSEN* copy number variation in natural accessions [28] and RdDM mutant backgrounds of *Arabidopsis* [37]. Upon mobilization, the heat-response elements in the LTRs of *ONSEN* [38] can create new gene regulatory networks responding to heat stress [11]. Therefore, it will now be of great interest to test if the *ONSEN* hc lines obtained in this study are better adapted to heat stress. This will allow us to test if retrotransposon-induced genetic and epigenetic changes more rapidly create beneficial alleles than would occur by random mutagenesis. Furthermore, the observation that HS did not lead to a stronger activation of *ONSEN* in hc lines

compared to WT plants suggests that genome stability is not compromised in these lines. This result can be explained by at least two possible mechanisms: (i) the occurrence of insertions of inverted duplications of *ONSEN*, such as has been observed for the *Mu killer* locus in maize [39]—such insertions will lead to the production of double-stranded RNA feeding into gene silencing and thereby limit the activity of that TE; and (ii) balancing of TE activity and integrated copy number as has been described for *EVADÉ* in *Arabidopsis* [40]. In this case, when a certain TE copy number threshold is reached robust transcriptional gene silencing takes over, thereby limiting TE mobility and ensuring genome stability. The stability of new TE insertions is an important aspect in light of the future use of TEs in crop breeding and trait stability.

Conclusions

TEs are important contributors to genome evolution. The ability to mobilize them in plants and possibly in other eukaryotes in a controlled manner with straightforward drug application, as shown here, opens the possibility to study their importance in inducing genetic and epigenetic changes resulting from external stimuli. Because the induced transposition of *ONSEN* can efficiently produce developmental changes in *Arabidopsis*, it will be very interesting to test if specific stress-induced TE activation can be used for directed crop breeding for better stress tolerance in the near future.

Methods

Plant material

All *Arabidopsis* mutants used in this study (*nprpb2-3* [14], *nprpd1-3* [41], *rdr6* [42], *dcl2/3/4* triple mutant [43]) are in the Col-0 background. For *O. sativa japonica*, the cultivar Nipponbare was used.

Growth conditions

Prior to germination, *Arabidopsis* seeds were stratified for 2 days at 4 °C. Before and during stress treatments plants were grown under controlled conditions in a Sanyo MLR-350 growth chamber on solid ½ MS medium (1% sucrose, 0.5% Phytigel (Sigma), pH 5.8) under long day conditions (16 h light) at 24 °C (day) and 22 °C (night) (*Arabidopsis*) and 12 h at 28 °C (day) and 27 °C (night) (*O. sativa*).

To analyze successive generations, seedlings were transferred to soil and grown under long day conditions (16 h light) at 24 °C (day) and 22 °C (night) (*Arabidopsis*) in a Sanyo MLR-350 growth chamber until seed maturity.

For phenotyping, *Arabidopsis* plants were grown under long day conditions (16 h light) at 24 °C (day) and 22 °C (night) and short day conditions (10 h light) at 21 °C (day) and 18 °C (night).

Stress and chemical treatments

Surface sterilized seeds of *Arabidopsis* and *O. sativa* were germinated and grown on solid ½ MS medium that was supplemented with sterile filtered zebularine (Sigma; stock, 5 mg/ml in DMSO), α -amanitin (Sigma; stock, 1 mg/ml in water) or a combination of both chemicals. Control stresses (6 °C for 24 h followed by control conditions for 24 h, CS) and heat stresses (6 °C for 24 h followed by 37 °C for 24 h, HS) of *Arabidopsis* seedlings were conducted as described previously [11].

DNA analysis

For qPCR and prior to digestions, total DNA from *Arabidopsis* plants was extracted with the DNeasy Plant Mini Kit (Qiagen) following the manufacturer's recommendations. For the qPCRs to measure the *ONSEN* copy number following HS and chemical treatments the aerial parts of at least ten *Arabidopsis* plants per replicate were pooled prior to DNA extraction. To track *ONSEN* copy numbers in the S1–3 generations of controls (only HS) and hc lines (HS + A&Z treatment) DNA from true leaves was extracted. For the estimation of the *ONSEN* transposition frequency, total DNA of pools consisting of at least eight seedlings of the progeny of HS + A&Z-treated plants was isolated. The DNA concentration was measured with a Qubit Fluorometer (Thermo Fisher Scientific). The copy numbers of *ONSEN* were determined with qPCRs on total DNA using a TaqMan master mix (Life Technologies) in a final volume of 10 μ l in the Light-Cycler 480 (Roche). *ACTIN2* (*AT3G18780*) was used to normalize DNA levels. Primer sequences are given in Additional file 1: Table S1.

For the mobilome-seq analysis total DNA from the pooled aerial parts of three 10-day-old *O. sativa* seedlings was extracted as previously reported [44]. Genomic DNA (5 μ g) for each sample was purified using a GeneClean kit (MPBio, USA) according to the manufacturer's instructions. ecDNA was isolated from the GeneClean product using PlasmidSafe DNase (Epicentre, USA) according to the manufacturer's instructions, except that the 37 °C incubation was performed for 17 h. DNA samples were precipitated by adding 0.1 volume of 3 M sodium acetate (pH 5.2), 2.5 volumes of ethanol and 1 μ l of glycogen (Fisher, USA) and incubating overnight at –20 °C. The precipitated circular DNA was amplified by random rolling circle amplification using the Illustra TempliPhi kit (GE Healthcare, USA) according to the manufacturer's instructions except that the incubation was performed for 65 h at 28 °C. The DNA concentration was determined using the DNA PicoGreen kit (Invitrogen, USA) using a LightCycler480 (Roche, USA). One nanogram of amplified ecDNA from each sample was used to prepare the libraries using the Nextera XT library kit (Illumina, USA) according to the manufacturer's instructions. DNA quality

and concentration were determined using a high sensitivity DNA Bioanalyzer chip (Agilent Technologies, USA). Samples were pooled and loaded onto a MiSeq platform (Illumina, USA) and 2 \times 250-nucleotide paired-end sequencing was performed. Quality control of FASTQ files was done using the FastQC tool (version 0.10.1). To remove any read originating from organelle circular genomes, reads were mapped against the mitochondria and chloroplast genomes using the program Bowtie2 version 2.2.2 71 with –sensitive local mapping. Unmapped reads were mapped against the reference genome IRGSP1.0 (<http://rgp.dna.affrc.go.jp/E/IRGSP/Build5/build5.html>) using the following parameters: –sensitive local, -k 1. DNA from both mitochondria and chloroplast genomes integrated in nuclear genomes was masked (1,697,400 bp). The TE-containing regions cover 194,224,800 bp in *O. sativa*. Finally, the bam alignment files were normalized and compared using deeptools [45] and visualized with the Integrative Genomics Viewer (IGV) software (<https://www.broadinstitute.org/igv/>). Data from the mobilome analysis were submitted to GEO (accession number GSE90484).

The presence of circular *Houba* copies was tested by an inverse PCR on 7 ng of the rolling-circle amplified template that was also used for sequencing. A PCR specific to chloroplast DNA served as a loading control. PCR products were separated on a 1% agarose gel that was stained with a Midori Green Nucleic Acid Staining Solution (Nippon Genetics Europe). Primer sequences are given in Additional file 1: Table S1.

Transposon display

The integration of additional copies of *ONSEN* into the genome of heat stressed and treated plants was ascertained by a simplified transposon display based on the GenomeWalker Universal kit (Clontech Laboratories), as previously described [11] with the following modifications: 300 ng of total DNA from adult plants in the S2 generation of heat stressed and A&Z-treated plants was extracted with a DNeasy Plant Mini Kit (QIAGEN) and digested with blunt cutter restriction enzyme *DraI* (NEB). After purification with a High Pure PCR Product Purification Kit (Roche) digested DNA was ligated to the annealed GenWalkAdapters 1&2. The PCR was performed with the adaptor-specific primer AP1 and the *ONSEN*-specific primer Copia78 3' LTR. The PCR products were separated on a 2% agarose gel that was stained with Midori Green. For primer sequence information, see Additional file 1: Table S1.

Cloning, sequencing and genotyping of new insertions

To identify the genomic region of new *ONSEN* insertions, the PCR product of the transposon display was purified using a High Pure PCR Product Purification Kit

(Roche), ligated into a pGEM-T vector (Promega) and transformed into *Escherichia coli*. After a blue white selection, positive clones were used for the insert amplification and sequencing (StarSEQ). The obtained sequences were analyzed with Geneious 8.2.1 and blasted against the *Arabidopsis* reference genome. The standard genotyping PCRs to prove novel *ONSEN* insertions were performed with combinations of the *ONSEN*-specific primer Copia78 3' LTR and primers listed in Additional file 1: Table S1.

RNA analysis and northern blotting

Total RNA from the aerial part of at least ten *Arabidopsis* seedlings was isolated using the TRI Reagent (Sigma) according to the manufacturer's recommendations. RNA concentration was measured (Qubit RNA HS Assay Kit, Thermo Fisher) and 15 µg of RNA was separated on a denaturing 1.5% agarose gel, blotted on a Hybond-N⁺ (GE Healthcare) membrane and hybridized with 25 ng of a gel-purified and P³²-labelled probe (Megaprime DNA Labelling System, GE Healthcare) specific to the full length *ONSEN* transcript (see Additional file 1: Table S1 for primer sequences). Northern blots were repeated in three independent experiments with the same results.

Whole-genome DNA methylation analysis

Whole-genome bisulfite sequencing library preparation and DNA conversion were performed as previously reported [46]. Bisulphite read mapping and methylation value extraction were done on the *Arabidopsis* TAIR10 genome sequence using BSMAP v2.89 [47]. Following mapping of the reads the fold coverages of the genome for CS, CS + A, CS + Z and CS + A&Z were 13.4, 13.2, 18.4 and 16.3, respectively. Data from the bisulphite sequencing analysis have been submitted to GEO (accession number GSE99396).

Statistics

Statistical analyses were performed with SigmaPlot (v. 11.0). Depending on the normality of the data, either an H-test or a one-way ANOVA was performed. The Student-Newman-Keuls method was used for multiple comparisons.

Additional file

Additional file 1: Table S1. Table of all primers used in this study. **Figure S1.** Increase in *ONSEN* copy numbers in S1 pools of heat-stressed and Z-treated *nripb2-3* plants. **Figure S2.** Detection of eccDNAs originating from *ONSEN* loci following heat stress and chemical treatments in *Arabidopsis*. **Figure S3.** Increase in *ONSEN* copy numbers in S1 pools of heat-stressed and A&Z-treated WT plants. **Figure S4.** Summary of confirmed novel *ONSEN* insertions in hc line 3. **Figure S5.** Stress-induced activation of *ONSEN* in the S3 generation after initial HS treatment. **Figure S6.** *Houba* forms LTR–LTR junction eccDNAs after combined A&Z treatment. (PDF 1660 kb)

Acknowledgements

We wish to thank Emilija Hristova for her support at the beginning of this project. We thank Christel Llauro for technical support on the production of the mobilomes and Todd Blevins for providing the *dcl2/3/4* triple mutant line. We thank the IMAC platform from the Structure Fédérative de Recherche 'Qualité et Santé du Végétal' (SFR QUASAV) for their technical support (Illumina sequencing).

Funding

This work was supported by grants provided by the European Commission (PITN-GA-2013-608422-IDP BRIDGES to MT and ERC grant 725701 BUNGEE to EB) and the region of Pays de la Loire (ConnecTalent EPICENTER project awarded to EB).

Availability of data and materials

The mobilome sequencing data and whole-genome DNA methylation analysis data are available in the GEO (accession numbers GSE90484 and GSE99396, respectively).

Authors' contributions

MT and EB conceived the study. MT, SL, SB and MM performed experiments. ND performed methylome analyses. MT and EB wrote the paper with contributions from SL and MM. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Competing interests

MT and EB declare that a patent application based on the presented discoveries has been submitted to the European Patent Office (PCT/EP2016/079276). EB is CEO of epibreed Ltd, a company that has an exclusive use license for this patent.

Author details

¹Botanical Institute, Zürich-Basel Plant Science Center, University of Basel, Hebelstrasse 1, 4056 Basel, Switzerland. ²Institut de Recherche pour le Développement, UMR232 DIADE Diversité Adaptation et Développement des Plantes, Université Montpellier 2, Montpellier, France. ³University of Perpignan, Laboratory of Plant Genome and Development, 58 Avenue Paul Alduy, 66860 Perpignan, France. ⁴IRHS, Université d'Angers, INRA, AGROCAMPUS-Ouest, SFR4207 QUASAV, Université Bretagne Loire, 49045 Angers, France.

Received: 13 February 2017 Accepted: 27 June 2017

Published online: 07 July 2017

References

- Schulman AH. Retrotransposon replication in plants. *Curr Opin Virol*. 2013;3:604–14.
- Lanciano S, Carpentier M-C, Llauro C, Jobet E, Robakowska-Hyzorek D, Lasserre E, et al. Sequencing the extrachromosomal circular mobilome reveals retrotransposon activity in plants. *PLoS Genet*. 2017;13:e1006630.
- Paszowski J. Controlled activation of retrotransposition for plant breeding. *Curr Opin Biotechnol*. 2015;32C:200–6.
- Huang CRL, Burns KH, Boeke JD. Active transposition in genomes. *Annu Rev Genet*. 2012;46:651–75.
- Belyayev A. Bursts of transposable elements as an evolutionary driving force. *J Evol Biol*. 2014;27:2573–84.
- Lisch D. How important are transposons for plant evolution? *Nat Rev Genet*. 2013;14:49–61.
- Miura A, Yonebayashi S, Watanabe K, Toyama T, Shimada H, Kakutani T. Mobilization of transposons by a mutation abolishing full DNA methylation in *Arabidopsis*. *Nature*. 2001;411:212–4.
- Matzke MA, Mosher RA. RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. *Nat Rev Genet*. 2014;15:394–408.
- Tsukahara S, Kobayashi A, Kawabe A, Mathieu O, Miura A, Kakutani T. Bursts of retrotransposition reproduced in *Arabidopsis*. *Nature*. 2009;461:423–U125.
- Mirouze M, Reinders J, Bucher E, Nishimura T, Schneeberger K, Ossowski S, et al. Selective epigenetic control of retrotransposition in *Arabidopsis*. *Nature*. 2009;461:427–30.

11. Ito H, Gaubert H, Bucher E, Mirouze M, Vaillant I, Paszkowski J. An siRNA pathway prevents transgenerational retrotransposition in plants subjected to stress. *Nature*. 2011;472:115–9.
12. Matzke MA, Kanno T, Matzke AJM. RNA-directed DNA methylation: the evolution of a complex epigenetic pathway in flowering plants. *Annu Rev Plant Biol*. 2015;66:243–67.
13. Gao Z, Liu H-L, Daxinger L, Pontes O, He X, Qian W, et al. An RNA polymerase II- and AGO4-associated protein acts in RNA-directed DNA methylation. *Nature*. 2010;465:106–9.
14. Zheng B, Wang Z, Li S, Yu B, Liu J-Y, Chen X. Intergenic transcription by RNA Polymerase II coordinates Pol IV and Pol V in siRNA-directed transcriptional gene silencing in Arabidopsis. *Genes Dev*. 2009;23:2850–60.
15. Cuerda-Gil D, Slotkin RK. Non-canonical RNA-directed DNA methylation. *Nature Plants*. 2016;2:16163.
16. Panda K, Ji L, Neumann DA, Daron J, Schmitz RJ, Slotkin RK. Full-length autonomous transposable elements are preferentially targeted by expression-dependent forms of RNA-directed DNA methylation. *Genome Biol*. 2016;17:1–19.
17. Lindell TJ, Weinberg F, Morris PW, Roeder RG, Rutter WJ. Specific inhibition of nuclear RNA polymerase II by alpha-Amanitin. *Science*. 1970;170:447–9.
18. Haag JR, Ream TS, Marasco M, Nicora CD, Norbeck AD, Pasa-Tolić L, et al. In vitro transcription activities of Pol IV, Pol V, and RDR2 reveal coupling of Pol IV and RDR2 for dsRNA synthesis in plant RNA silencing. *Mol Cell*. 2012;48:811–8.
19. Baubec T, Pecinka A, Rozhon W, Mittelsten SO. Effective, homogeneous and transient interference with cytosine methylation in plant genomic DNA by zebularine. *Plant J*. 2009;57:542–54.
20. Flavell AJ, Ish-Horowitz D. Extrachromosomal circular copies of the eukaryotic transposable element Copia in cultured *Drosophila* cells. *Nature*. 1981;292:591–5.
21. Russo J, Harrington AW, Steiniger M. Antisense transcription of retrotransposons in *Drosophila*: an origin of endogenous small interfering RNA precursors. *Genetics*. 2016;202:107–21.
22. Matsunaga W, Ohama N, Tanabe N, Masuta Y, Masuda S, Mitani N, et al. A small RNA mediated regulation of a stress-activated retrotransposon and the tissue specific transposition during the reproductive period in Arabidopsis. *Front Plant Sci*. 2015;6:48.
23. Rosso MG, Li Y, Strizhov N, Reiss B, Dekker K, Weisshaar B. An Arabidopsis thaliana T-DNA mutagenized population (GABI-Kat) for flanking sequence tag-based reverse genetics. *Plant Mol Biol*. 2003;53:247–59.
24. Appelhagen I, Thiedig K, Nordholt N, Schmidt N, Hupé G, Sagasser M, et al. Update on transparent testa mutants from Arabidopsis thaliana: characterisation of new alleles from an isogenic collection. *Planta*. 2014;240:955–70.
25. Kawahara Y, la Bastide de M, Hamilton JP, Kanamori H, Mccombie WR, Ouyang S. Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice*. 2013;6:4–10.
26. Panaud O, Vitte C, Hivert J, Muziak S, Talag J, Brar D, et al. Characterization of transposable elements in the genome of rice (*Oryza sativa* L.) using Representational Difference Analysis (RDA). *Mol Genet Genomics*. 2002;268:113–21.
27. Vitte C, Panaud O. LTR retrotransposons and flowering plant genome size: emergence of the increase/decrease model. *Cytogenet Genome Res*. 2005;110:91–107.
28. Quadrana L, Silveira AB, Mayhew GF, LeBlanc C, Martienssen RA, Jeddeloh JA, et al. The Arabidopsis thaliana mobilome and its impact at the species level. *elife*. 2016;5:e15716.
29. Vitte C, Panaud O, Quesneville H. LTR retrotransposons in rice (*Oryza sativa*, L.): recent burst amplifications followed by rapid DNA loss. *BMC Genomics*. 2007;8:218–15.
30. Wicker T, Keller B. Genome-wide comparative analysis of copia retrotransposons in Triticeae, rice, and Arabidopsis reveals conserved ancient evolutionary lineages and distinct dynamics of individual copia families. *Genome Res*. 2007;17:1072–81.
31. Oelgeschlager T. Regulation of RNA polymerase II activity by CTD phosphorylation and cell cycle control. *J Cell Physiol*. 2001;190:160–9.
32. Palancade B, Bensaude O. Investigating RNA polymerase II carboxyl-terminal domain (CTD) phosphorylation. *Eur J Biochem*. 2003;270:3859–70.
33. Pai DA, Kaplan CD, Kweon HK, Murakami K, Andrews PC, Engelke DR. RNAs nonspecifically inhibit RNA polymerase II by preventing binding to the DNA template. *RNA*. 2014;20:644–55.
34. Lazcano A, Fastag J, Gariglio P, Ramirez C, Oro J. On the early evolution of RNA-polymerase. *J Mol Evol*. 1988;27:365–76.
35. Berretta J, Pinskaya M, Morillon A. A cryptic unstable transcript mediates transcriptional trans-silencing of the Ty1 retrotransposon in *S. cerevisiae*. *Genes Dev*. 2008;22:615–26.
36. Chang W, Jääskeläinen M, Li S-P, Schulman AH. BARE retrotransposons are translated and replicated via distinct RNA pools. *PLoS One*. 2013;8:e72270–12.
37. Ito H, Kim J-M, Matsunaga W, Saze H, Matsui A, Endo TA, et al. A Stress-activated transposon in Arabidopsis induces transgenerational abscisic acid insensitivity. *Sci Rep*. 2016;6:23181.
38. Pietzenk B, Markus C, Gaubert H, Bagwan N, Merotto A, Bucher E, et al. Recurrent evolution of heat-responsiveness in Brassicaceae COPIA elements. *Genome Biol*. 2016;17:209.
39. Slotkin R, Freeling M, Lisch D. Heritable transposon silencing initiated by a naturally occurring transposon inverted duplication. *Nat Genet*. 2005;37:641–4.
40. Mari-Ordóñez A, Marchais A, Etcheverry M, Martin A. Reconstructing de novo silencing of an active plant retrotransposon. *Nat Genet*. 2013;45:1029–39.
41. Herr AJ, Jensen MB, Dalmay T, Baulcombe DC. RNA polymerase IV directs silencing of endogenous DNA. *Science*. 2005;308:118–20.
42. Peragine A, Yoshikawa M, Wu G, Albrecht H, Poethig R. SGS3 and SGS2/SDE1/RDR6 are required for juvenile development and the production of trans-acting siRNAs in Arabidopsis. *Genes Dev*. 2004;18:2368–79.
43. Blevins T, Pontes O, Pikaard CS, Meins F. Heterochromatic siRNAs and DDM1 independently silence aberrant 5S rDNA transcripts in Arabidopsis. *PLoS One*. 2009;4:e5932–2.
44. Mette MF, van der Winden J, Matzke MA, Matzke AJ. Production of aberrant promoter transcripts contributes to methylation and silencing of unlinked homologous promoters in trans. *EMBO J*. 1999;18:241–8.
45. Ramirez F, Dündar F, Diehl S, Grüning BA, Manke T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res*. 2014;42:W187–91.
46. Becker C, Hagmann J, Müller J, Koenig D, Stegle O, Borgwardt K, et al. Spontaneous epigenetic variation in the Arabidopsis thaliana methylome. *Nature*. 2011;480:245–9.
47. Xi Y, Li W. BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics*. 2009;10:232.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit



3.5 Quel est le rôle des ADNec viraux chez la drosophile ?

Dans la plupart des études précédemment citées, nous nous sommes uniquement intéressés à l'étude du mobilome dans le but d'identifier les ET actifs. Or le mobilome d'un tissu ou d'un organisme est constitué d'une population d'ADNec de différentes catégories (voir Introduction page 30).

L'équipe de Maria-Carla Saleh (Institut Pasteur) s'intéresse aux relations entre les virus et les insectes au niveau du système immunitaire des insectes. En effet, les insectes et notamment les moustiques, sont connus pour transmettre des virus, par exemple aux hommes, en jouant le rôle de vecteur. Les insectes ont ainsi développé des mécanismes immunitaires puissants leur permettant d'être asymptomatique à la suite d'une infection virale. Récemment, cette équipe (Goic *et al.* 2013; 2016) a montré que l'ARN viral était reverse-transcrit en ADN par des RT provenant de rétrotransposons. De plus cet ADNec viral est reconnu par la machinerie RNAi de l'insecte hôte induisant la production de petits ARN impliqués dans la réponse antivirale de l'hôte. Afin de tester la présence d'ADNec viral chez des mouches *Drosophila melanogaster* infectées par des virus à ARN, l'équipe de M.C Saleh s'est intéressée à la technique du mobilome-seq. J'ai participé à l'analyse des données du mobilome qui nous ont notamment permis de détecter la présence d'ADNec viral. De plus, nos collaborateurs ont montré que l'inoculation d'ADNec viral chez la drosophile induisait la production de petits ARN spécifiques des virus conférant une forte résistance aux virus. Les résultats de cette étude suggèrent que la synthèse d'ADNec viral pourrait participer à la réponse immunitaire chez *Drosophila*. Les résultats de cette étude ont fait l'objet d'une publication actuellement en révision (Poirier *et al.*, en révision).

3.6 Conclusion

De la drosophile (170 Mb) au pin (20 Gb), dans des conditions normales ou lors de stress abiotiques et biotiques, dans des fonds sauvages ou dans des fonds mutants épigénétiques, nos analyses bioinformatiques ont été développées dans le but d'obtenir une méthode adaptée aux génomes complexes et pour cribler un grand nombre de tissus, mutants ou conditions de stress. L'ensemble des collaborations développées pendant ma thèse nous a également permis de rendre plus polyvalente notre méthode, par l'utilisation de la technique CRISPR Cas9 ou par l'assemblage *de novo* des données de séquençage par exemple. Enfin nous avons identifié

de nouveaux ET candidats dont l'activité devra être analysée plus finement.

En parallèle à nos études, l'engouement pour les ADNec a conduit à la multiplication très récente de méthodes pour les analyser. Dernièrement, Shoura et *al* (2017) ont apporté une nouvelle alternative dans le séquençage des mobilomes, appelés circulomes dans leur étude. Ils ont élégamment modifié la purification des ADNec afin d'obtenir suffisamment d'ADNec pour supprimer l'étape d'amplification et séquencer directement grâce à l'association entre une stratégie biophysique (centrifugation par gradient de chlorure de césium) et une stratégie biochimique (multiplication des digestions avec exonucléase). Néanmoins les ADNec ainsi identifiés demandent à être validés par une autre technique.

Les travaux détaillés ici montrent que de nombreuses questions biologiques peuvent trouver leurs réponses dans l'étude du mobilome et il semblerait que la caractérisation de ce répertoire génomique sous-exploré n'en soit qu'à son commencement.

- DISCUSSION GÉNÉRALE -

Grâce aux avancées de la génomique et au vu du nombre exponentiel de génomes séquencés disponibles, il est clairement démontré que les ET prolifèrent dans les génomes. Or, peu d'exemples dans la littérature illustraient l'activité transpositionnelle des ET. Les méthodes de détection disponibles lors de mes débuts en thèse ne semblaient pas assez puissantes pour évaluer le niveau de transposition global des ET et nous n'avions aucune idée de la fréquence des mouvements des ET. Est-ce un évènement rare ? Est-ce-que la réactivation d'un ET est fréquente dans les génomes mais à cause du nombre important de copies d'ET dans les génomes, leur activité est difficilement détectable ? Pour répondre à ces questions, nous nous sommes intéressés à un répertoire génomique inexploré et sous-estimé, le mobilome. Identifier les ADNec correspondant à des ET dans un organisme ou un tissu donné s'est avéré être un moyen efficace et à faible coût afin d'évaluer le niveau transpositionnel d'un génome.

L'étude des mobilomes nous a beaucoup appris sur l'activité somatique des ET et le contrôle de ces éléments s'avère bien plus complexe qu'initialement supposé. Malgré le nombre important d'organismes, de tissus et de conditions testés pendant ma thèse, nous n'avons jamais observé une réactivation massive d'ET même dans des tissus où l'épigénome est totalement déstabilisé. En effet, le modèle qui illustre la transmission des marques épigénétiques d'une génération à une autre chez les plantes propose une reprogrammation de ces marques qui se traduit par une réactivation transitoire des ET. S'il est vrai que l'hypométhylation observée dans l'albumen de plusieurs variétés de riz induit la réactivation transcriptionnelle massive des ET, l'activité transpositionnelle de ces ET n'est pas observée. Nos données de mobilome-seq nous ont montré que le relâchement épigénétique observé à la fois dans les tissus entourant les gamètes et à la fois dans l'albumen après la fécondation ne semble pas suffisant pour permettre une réactivation massive des ET contrairement à ce que le modèle suggérait (Martínez et Köhler 2017).

Si les marques épigénétiques jouent un rôle central dans le maintien du *silencing* de ces éléments, la réactivation des ET semble néanmoins résulter de la combinaison de plusieurs facteurs qui peuvent être épigénétiques mais aussi environnementaux et développementaux. Selon la ou les conditions étudiées, l'analyse des mobilomes nous a également montré que la transposition des ET était spécifiquement ciblée à quelques familles d'ET suggérant que les mécanismes de régulation pouvaient être famille-dépendant. Cette observation avait également été faite lors d'analyses génomiques, notamment chez le riz, qui illustraient les traces de *bursts* de transposition spécifiques de certaines familles au cours de l'évolution

(Vitte et al. 2007). Quelles sont alors les spécificités de ces familles ? Comment l'activité des ET peut être régulée par différents mécanismes indépendants ?

La caractérisation de l'activité de *PopRice* s'est révélée être un intéressant modèle pour étudier ce type de mécanismes famille-dépendant. Effectivement, l'activité de *PopRice* semblait être expliquée à la fois par un facteur épigénétique, l'hypométhylation de l'albumen et à la fois par un facteur développemental, la présence de boîtes à glutélines dans le LTR. L'analyse de *PopRice* nous a conduit à comparer les différentes sous-familles qui constituent la famille *Osr4*. S'il est connu que seuls quelques éléments d'une même famille sont capables de s'activer dans une condition (par exemple la copie de *Tos17* sur le chromosome 7 dans les cals), ici nous décrivons l'« indépendance » d'une sous-famille. Et dans ce cas précis, nous n'avons pas seulement étudié un mécanisme famille-dépendant mais sous-famille dépendant ce qui suggère que les acteurs responsables de l'activité des ET peuvent être très caractéristiques de certains éléments. L'accumulation de boîtes à glutélines dans le LTR de *PopRice* est la seule différence notable entre les éléments *PopRice* et les autres éléments d'*Osr4*. Or, aujourd'hui aucun lien direct n'a été établi entre la présence de ces boîtes et la transposition spécifique de *PopRice*. Paradoxalement, tous les membres de la famille semblent être transcrits et déméthylés dans l'albumen de riz nous obligeant à nous tourner vers de nouvelles pistes pour identifier les mécanismes de régulation de *PopRice*.

L'activité conservée et contrastée de *PopRice* dans de nombreuses variétés de riz asiatiques nous a permis d'initier un nouveau type d'analyse GWAS basée sur la présence (ou l'absence) d'ADNec de *PopRice*. Deux gènes spécifiques de l'albumen semblent être des candidats intéressants pour la suite de l'analyse. L'utilisation des variations de l'activité d'un ET comme phénotype n'est possible qu'à l'aide de méthodes capables de détecter l'activité somatique d'un ET. En d'autres termes, dans des études futures la combinaison entre des analyses mobilomiques et des analyses GWAS pourrait s'avérer fructueuse pour la caractérisation des ET.

En revanche, si nous avons effectivement montré que la présence d'ADNec de *PopRice* était liée à la présence de néo-insertions de *PopRice* dans le génome de l'albumen, la validation par PCR des insertions somatiques détectées à partir des données de reséquençage s'est avérée difficile et même un reséquençage profond n'a pas permis d'évaluer précisément le niveau de transposition somatique. Les difficultés techniques liées à l'étude de l'activité somatique des ET restent un frein qui pourra être dépassé par les nouvelles techniques de séquençage de 3^{ème}

génération. Dans ce contexte le mobilome-seq permet d'identifier rapidement des ET candidats dont les néo-insertions peuvent ensuite être analysées par des techniques de séquençage ciblé ou capture.

Nos travaux rejoignent un ensemble d'avancées techniques récents qui offrent également de nouvelles perspectives sur l'étude des ET. Le rôle des ET dans la création de nouveaux réseaux de gènes par exemple a dans un premier temps été soupçonné par le biais d'analyses de ChIP-seq montrant que les ET étaient souvent liés par des facteurs de transcription et qu'ils présentaient des niveaux d'expression tissu ou condition spécifique. Dernièrement et à l'aide de la méthode d'*editing* CRISPR, l'équipe de Cédric Feschotte a démontré l'implication directe de ERV dans la mise en place des réponses immunitaires chez les mammifères (Chuong et *al.* 2016a). La possibilité d'éliminer spécifiquement certains rétrovirus ou ET par CRISPR-Cas9 est désormais un outil puissant pour valider l'impact des ET sur les génomes hôtes. De plus, les travaux de Leandro Quadrana (2016) menés dans l'équipe de Vincent Colot ont montré que l'utilisation du GWAS pouvait être une technique efficace pour l'étude des acteurs de la régulation des ET chez *Arabidopsis*. Dans leur étude, ils ont utilisé, pour chaque famille d'ET, le nombre d'insertions d'ET par génome comme phénotype afin de caractériser les régions du génome contrôlant la mobilité de ces familles. Les perspectives qui découleront de cette étude s'annoncent très séduisantes. Pour finir, le développement et la précision d'outils de séquençage comme le séquençage *single cell* avec la technologie Nanopore (voir Introduction page 27) devrait faciliter l'identification des variants structuraux dont les insertions d'ET (Debladis et *al.* 2017).

Les exemples (et les revues) sur l'importance des ET dans les génomes se multiplient et il semblerait que leur « légitimité » dans les génomes eucaryotes ne soit plus à démontrer. Qu'il s'agisse de leurs rôles dans la régulation des gènes, dans l'apparition de nouveautés biologiques (développement du placenta chez les mammifères par exemple), ou de caractères agronomiques (variété de raisin Chardonnay ou architecture du maïs par exemple), le potentiel évolutif des ET est définitivement admis dans la communauté scientifique. L'étude des ET dans des plantes à intérêt agronomique est aujourd'hui possible par la mise en place de techniques telles que le mobilome-seq, adaptées aux génomes complexes et qui permettront d'améliorer notre compréhension sur l'adaptation et l'évolution des organismes afin de les mettre à profit pour l'agriculture de demain.

Pour finir, durant ma thèse, je me suis principalement intéressée aux ADNec qui correspondent à des ET. Or, comme mentionné dans le chapitre 1 de l'introduction, le mobilome d'une cellule ou d'un tissu renferme une population d'ADNec très hétérogène. À titre d'exemple, notre collaboration avec l'équipe de M.C Saleh a mis en évidence le lien entre la présence d'ADNec viral et la mémoire immunitaire de l'hôte chez la drosophile. Des études récentes ont également illustré l'importance de ces ADNec dans l'évolution des cancers chez l'homme (Turner et *al.* 2017) ou dans l'implication de mécanismes de défense chez la paramécie (Allen et *al.* 2017). Le rôle de ces cercles suscite l'engouement des scientifiques (Pennisi 2017) et l'intérêt nouveau et grandissant pour les ADNec a conduit à la multiplication très récente de méthodes pour les analyser. Le séquençage du mobilome est aujourd'hui la technique la plus développée pour permettre une caractérisation globale de ces répertoires génomiques sous-explorés et ouvrent de nouvelles voies en génomiques dont les enjeux sont importants notamment dans le milieu médical.

Damon Lisch a écrit en décembre 2016 sur Twitter que « les généticiens avaient besoin d'accepter que les génomes des plantes étaient à la fois chaotiques et ordonnés ». Aujourd'hui, il est loin le temps où le dogme central de la biologie moléculaire régnait en maître. Il est indéniable que les futures analyses des mobilomes apporteront leur lot de réponses sur la compréhension de la dynamique des génomes. Cette nouvelle ère s'annonce passionnante.

- MATÉRIEL & MÉTHODES -

Matériel et Méthodes – Chapitre 3 – Partie 2

Régulations génétiques et épigénétiques de *PopRice*

Analyse transcriptionnelle. Les ARN totaux ont été isolés à partir de feuilles, fleurs et grains en utilisant la solution Tri-reagent (MRC) et en suivant les instructions du fabricant. Les ARN ont été traités avec une DNase du kit RQ1 (Promega) et 1,25 µg ont été réverse-transcrits en ADNc en utilisant le kit GoScript (Promega). Les analyses de PCR quantitative en temps réel (RT-qPCR) ont été réalisées avec 7 à 35 ng d'ADNc. Les analyses RT-qPCR ont été effectuées sur un LightCycler 480 (Roche) en utilisant le kit Takyon No Rox SYBR MasterMix dTTP Blue (Eurogentec) en suivant les instructions du fabricant. Les amorces ont été utilisées avec une concentration de 2 µM. Les conditions de RT-qPCR sont les suivantes : une étape de dénaturation à 95°C pendant 5 minutes, suivie de 40 cycles à 95°C pendant 10s, 60°C pendant 10s et une augmentation progressive de 0,04°C par seconde jusqu'à 95°C et une étape finale de refroidissement à 40°C pendant 30 secondes. Trois réplicats biologiques ont été analysés pour chaque tissu. Les niveaux d'expression relative de *PopRice* et d'*Osr4* comparés au gène eIF-5a (Xu et al. 2015) ont été calculés à partir de la formule suivante : $2^{-(\text{moyenne (CT PopRice)} - \text{CT gène référence}))}$. Les séquences des amorces utilisées sont référencées en Annexes (Tableau S2).

Analyse du niveau de méthylation. L'ADN génomique a été extrait à partir de feuilles, de grains et d'albumen de Nipponbare en utilisant la méthode CTAB (Clarke 2009). Pour chacun des échantillons, 500 ng ont été digérés par *HpaII* dans un volume total de 20 µl et la réaction a été incubée à 37°C sur la nuit. Le produit de la digestion a ensuite été dilué par un facteur 5 et 2 µl ont été utilisés pour réaliser la réaction de PCR. Deux réplicats biologiques ont été analysés pour chacun des échantillons.

Analyse *in silico* de données bisulfites. Les données de Zemach et al., (2010) ont été téléchargées sur la plateforme Gene Expression Omnibus (GEO) portant le numéro d'accèsion GSE22591. Les fichiers SRR059001.sra et SRR059005.sra ont été convertis en fichiers FASTQ à l'aide du module fastq-dump. Les index de séquençage Illumina ont été éliminés à l'aide du programme java Trimmomatic. Les lectures ont été alignées contre le génome de référence IRGSP1.0 à l'aide du programme Bismarck via l'algorithme de Bowtie2 et les paramètres suivants ont été utilisés : le premier ancrage d'une lecture se fait par fenêtres de 20 bp (-L 20) avec un seul *mismatch* autorisé (-N 1), la taille de l'insert entre les deux

lectures est de maximum 1000 bp (-X 1000) et les paramètres standards ont été utilisés pour les scores des alignements (--score_min L,0-0.2) autorisant 2 *mismatches* par lecture. Les duplications de PCR sont éliminées de l'analyse par le programme de Bismarck. Un fichier CX report est créé avec l'état de méthylation de chaque cytosine du génome et le nombre de lectures associées. La couverture et la détection de DMR (régions différenciellement méthylées) sont analysées à l'aide du programme DMRcaller du logiciel R.

Analyse comparative de *PopRice* au sein des céréales

Identification de la sous-famille *PopRice* dans Nipponbare, les 12 génomes *Oryza* et dans les génomes des céréales. Afin de déterminer l'histoire évolutive des éléments de *PopRice*, une séquence consensus de *PopRice* a été utilisée pour détecter par BLAST tous les rétrotransposons à LTR appartenant à la même famille dans le génome de référence IRGSP-1.0, dans les 12 génomes *Oryza* assemblés (The International *Oryza* Map Alignment Consortium, *en révision*) et dans les génomes des *Poaceae* (*Zea mays*, *Triticum aestivum*, *Sorghum bicolor*). Les paramètres du BLAST sont les suivants : HSP>4000 bp, minimum 70% d'identité, e-value < e-50. Toutes les séquences sélectionnées ont été alignées à l'aide du logiciel MAFFT et visualisées à l'aide du logiciel SEAVIEW (voir Annexes). Tous les éléments incomplets ont été éliminés et 47 éléments ont été sélectionnés pour la famille *Osr4* (*PopRice* inclus) pour Nipponbare et 231 éléments pour les 12 génomes. Les arbres phylogénétiques ont été construits à partir de la méthode PhyML et visualisés avec le logiciel FigTree (voir Annexes). L'arbre représentant les éléments de Nipponbare a été réalisé à partir des séquences entières des éléments alors que l'arbre des 12 génomes a été réalisé à partir de la séquence des LTR de chaque élément.

Analyses des données de mobilome-seq des variétés de riz africains et vietnamiens et des grains de céréales. La préparation des ADNec et des banques de séquençage Illumina a été réalisée en suivant le même protocole que celui décrit précédemment (Lanciano et al., 2017). Les séquences des variétés de riz ont été alignées contre le génome de référence IRGSP1.0 (International Rice Genome Sequencing Project version 5 ; voir Annexes) et les paramètres d'alignements sont les suivants : alignement local sensible (--sensitive local mapping) et seul le meilleur hit est conservé (-k 1). Pour chaque banque, un fichier *.bam* est obtenu et visualisé avec le logiciel *Integrative Genomics Viewer* (IGV) software (voir Annexes).

Southern blot. L'ADN génomique est extrait à l'aide de la méthode CTAB et les échantillons

sont déposés sur un gel d'agarose de 0,8% et transférés sur une membrane en nylon Hybond-N+ (GE, USA). Tous les Southern blot effectués durant ma thèse ont été réalisés à l'aide d'une sonde non-radioactive et le signal d'hybridation est détecté avec le système DIG (Roche, Suisse) en suivant les instructions du fabricant et comme décrit précédemment par Picault et *al.* (2009). Les bains de stringence ont été réalisés à 65°C dans une solution SSC à 0.5X. Les sondes ont été amplifiées par PCR à partir d'ADN génomique. Les séquences des sondes utilisées sont référencées en Annexes.

Phénotypage des variétés vietnamiennes par PCR. Pour chaque variété, l'ADN des grains germés a été extrait en utilisant la méthode CTAB. Les réactions PCR sont réalisées à partir de 2 µl d'ADN (environ 20 ng) dans un volume final de 15 µl. Toutes les paires d'amorces ont été dessinées à l'aide du programme Primer3 (voir Annexes) et la qualité a été testée avec OligoCalc (voir Annexes) et avec BLAST (voir Annexes). Les conditions de PCR utilisées sont les suivantes : une première étape de dénaturation à 95°C pendant 5 minutes suivie de 30 cycles à 95°C pendant 30 secondes, une étape d'hybridation pendant 30 secondes, une étape d'élongation à 72°C pendant 1 minute et 10 secondes et une étape finale d'extension à 72°C pendant 5 minutes. 8 µl du produit PCR ont été déposés sur un gel d'agarose à 1,5% et la migration a été effectuée à 135 mV pendant 30 minutes. L'ADN est rendu visible par le GelRed dye (Biotium). Les images des gels ont été obtenues avec le système d'imagerie UGenius gel (Syngene). Les séquences des amorces utilisées sont référencées en Annexes.

Analyse GWAS des variétés de riz vietnamiens. La matrice de marqueurs SNP a été générée à partir d'une étude GBS comme décrit précédemment par Phung et *al.* (2014). Brièvement, les marqueurs qui ont plus de 20% de données manquantes et qui ont une fréquence allélique rare (<5%) sont éliminés à l'aide du programme Beagle v3.3.2 (Browning et Browning 2007). 21623 marqueurs SNP ont été retenus pour la suite de l'analyse. Afin de déterminer la structure de la population une analyse à composante principale (ACP) est effectuée à l'aide du programme Eigensoft (Price et *al.* 2006). Les cinq premières composantes principales (PC) expliquent la plus grande partie de la variance et représentent l'effet fixé de la covariance. Afin d'associer le phénotype (présence ou absence de *PopRice*) avec le génotype, un modèle linéaire mixte est réalisé en prenant en compte la structure de la population (5 PC, effet fixé) et les liens de parenté (matrice kinship IBS, effet aléatoire). Le modèle est exécuté à l'aide du logiciel EMMAX (Kang et *al.* 2010). Une matrice kinship IBS est une matrice d'apparementement entre chaque individu 2 à 2. Elle a été obtenue à l'aide du

programme EMMAX-kin en suivant les paramètres par défaut. Les PC et la matrice kinship sont utilisés pour contrôler les faux-positifs qui peuvent devenir importants dans les populations structurées. Les résultats sont illustrés avec deux types de représentation graphique un QQ-plot et un Manhattan plot réalisés sous R à l'aide du package qqman (Turner 2014).

Activité transpositionnelle de *PopRice*

Reséquençage de l'albumen de Nipponbare. Les grains matures de Nipponbare ont été disséqués à l'aide d'une épingle sous la loupe binoculaire. L'ADN de l'albumen a été extrait par la méthode au CTAB (Clarke 2009). Les banques et le séquençage ont été réalisés par Novogene Co. (USA) et des lectures pairées de 250 nt ont été obtenues. La qualité des fichiers FASTQ a été évaluée à l'aide de l'outil FastQC (version 0.10.1 ; voir Annexes). Dans le but de déterminer la couverture moyenne de séquençage, les lectures sont alignées sur 72 gènes uniques avec Bowtie2 (Langmead et Salzberg 2012) et la couverture de chaque gène est calculée avec le module coverageBED du programme BEDTools. La moyenne et l'écart-type de la profondeur de séquençage sont ensuite calculées. Au préalable, les 72 gènes uniques ont été sélectionnés par l'alignement des séquences CDS contre le génome de référence IRGSP1-0 par BLAST. Seules, les séquences avec des hits uniques ont été retenus pour la suite de l'analyse.

Analyse des 3000 génomes et détection des néo-insertions de *PopRice*. Afin de détecter les insertions de *PopRice* dans les 3000 génomes ou les néo-insertions de *PopRice* dans l'albumen de Nipponbare, un pipeline d'analyse a été développé par Marie-Christine Carpentier (thèse sous la direction d'Olivier Panaud, UPVD). Les lectures sont alignées contre l'ET de référence (ici la séquence de *PopRice_11*) en utilisant le programme d'alignement Bowtie2 (Langmead et Salzberg 2012) avec un seul hit autorisé (-k 1). Les paires de lectures alignées / non-alignées sont filtrées à partir du FLAG et les lectures non-alignées sont alignées par BLAST contre le génome de référence IRGSP1.0 dans le but d'identifier les points d'insertions de *PopRice*. Le nombre de lectures qui soutiennent l'insertion est calculé à l'aide du module coverageBED du programme BEDTools. Afin de valider *in silico* l'insertion nouvellement identifiée, des dot-plots sont réalisés en prenant la séquence de la région d'insertion (10 kb), la séquence de l'ET (ici *PopRice_11*) et la séquence

des lectures qui soutiennent l'insertion. Pour qu'une néo-insertion soit validée, la première lecture d'une paire doit s'aligner sur la région d'insertion (qui ne présente aucune homologie de séquence avec l'ET) alors que la seconde lecture doit s'aligner sur l'ET.

Détection des lectures coupées. Afin de détecter les lectures qui chevauchent l'ET et les régions flanquantes de la néo-insertion, une analyse des lectures dites « coupées » a été effectuée. Les lectures sont alignées contre le génome de référence en utilisant le programme *segemehl* (Coil et al. 2015) avec les paramètres suivants : -S (alignement des lectures coupées) -A 95 (précision de 95%) -U 24 (score minimum de 24) -Z 25 (longueur minimum de 25) -W 95 (95% de la lecture). Les lectures « coupées » sont filtrées par rapport au FLAG à l'aide du module *view* du programme SAMtools et sont réunies dans un fichier *.bam*. Ainsi, seules les lectures qui n'alignent pas correctement dans leur paire (-f 14) et qui ont des alignements multiples (-F 256) sont considérées comme des lectures « coupées ». Le module *coverageBED* du programme BEDTools est utilisé pour déterminer la profondeur de séquençage de ces lectures coupées. Les fichiers *.bam* sont visualisés avec IGV.

Activité de *PopRice* en condition de stress

Culture de cals et traitement ABA. La culture de cals a été générée à partir d'embryon de grain de la variété Nipponbare. La méthode utilisée pour générer les cals a été décrite par Sallaud et al. (2003). Durant 5 semaines les grains ont été inoculés sur un milieu d'induction de cals. Chaque cal est ensuite formé de plusieurs unités, génétiquement identiques. Les unités sont séparées et cultivées sur un milieu de maintenance pendant 4 semaines et sont ensuite transférées pendant une semaine sur un milieu de maturation. Les cals alors âgés de 10 semaines ont été imbibés d'ABA à une concentration de 5 μ M pendant une semaine. Une expérience contrôle (sans ABA) a également été réalisée. L'ADN issu des deux conditions (avec ABA et sans ABA) a ensuite été extrait à l'aide de la méthode CTAB. L'analyse par PCR a été réalisée dans les mêmes conditions décrites précédemment. Les séquences des amorces utilisées ont été référencées en Annexes.

Matériel et Méthodes – Chapitre 3 – Partie 3

1. Stratégies bioinformatiques du mobilome-seq

Alignement des lectures contre un génome de référence. Pour les organismes pour lesquels nous disposons d'un génome de référence, les lectures de séquençage sont alignées sur le génome de référence. Les paramètres utilisés pour l'alignement sont identiques à ceux précédemment décrits (Lanciano et al., 2017) autrement dit : alignement local et sensible et pas d'alignement multiple (-k 1), seul le meilleur hit est gardé pour chacune des paires de lectures. Pour chaque banque, un fichier d'alignement *.bam* correspondant aux régions génomiques enrichies en lectures est visualisé avec le logiciel IGV. Le module coverageBED du programme BEDTools et un fichier d'annotation des ET du génome sont utilisés pour déterminer la couverture de séquençage de chaque ET. Les données de couverture sont ensuite normalisées par le nombre total de lectures exprimé en rpm (lectures par million de lectures). Seuls les ET qui sont couverts sur 90% de leur longueur totale sont considérés comme des candidats intéressants.

Création d'une base de données d'ET ou utilisation de base de données publiques. Pour les organismes pour lesquels nous ne disposons pas de base de données d'ET, deux stratégies ont été mises en place. La première consiste à utiliser une base de données existante pour une espèce proche. Cette base de données est alignée contre le génome de référence de l'espèce étudiée par BLAST et les hits supérieurs à 100 bp et avec une e-value $> 1e^{-50}$ sont conservés. Un fichier *.gff* d'annotation est obtenu et utilisé pour l'analyse de la couverture de séquençage des ET. La seconde stratégie consiste à utiliser des bases de données publiques telles que NCBI ou Repeat Masker pour identifier toutes les régions du génome qui présentent une forte couverture de séquençage.

Assemblage *de novo* des données de mobilome Pour les organismes pour lesquels nous ne disposons pas de génome de référence, les lectures des données de séquençage du mobilome sont assemblées *de novo* par le programme A5-miseq (Coil et al. 2015) spécifiquement développé pour assembler des données Mi-Seq. Pour chacune des banques, un fichier *.fasta* et un fichier *.bam* sont obtenus et le module *idxstats* provenant du programme SAMtools est utilisé pour déterminer le nombre de lectures correspondant à chacun des scaffold assemblé. Les données de couverture sont normalisées par le nombre total de lectures utilisés pour l'assemblage *de novo* et sont exprimées en lectures par million. Les scaffolds sont annotés en

utilisant une analyse par BLAST contre les génomes des organelles et contre une base de données d'ET. Pour les organismes pour lesquels nous ne disposons pas de base de données, les scaffolds correspondant à des ET sont identifiés par une recherche des domaines conservés et sont annotés l'aide du logiciel repeatmasker.

2. CRISPR-mobilome-seq

Extraction d'ADN. L'ADN génomique a été extrait à partir de feuilles de maïs *Zea mays* B73 par la méthode du CTAB (Clarke 2009). L'ADNcc a été isolé à partir de 15 µg d'ADN fraîchement extrait à l'aide de la DNase PlasmidSafe (Epicentre) en suivant les instructions du fabricant dans un volume total de 50 µl. La durée de l'incubation à 37°C a été prolongée de 17 heures. 20 µl de la réaction ont été utilisés pour la digestion avec l'enzyme Cas9.

Transcription des ARN guides. Deux oligonucléotides d'ADN d'environ 120bp ont été synthétisés (Eurofins Genomics) et utilisés comme matrice pour la transcription *in vitro* des ARN guides. Chacun des oligonucléotides synthétisés contient la séquence du promoteur T7 RNA polymérase suivi de la séquence de la cible, ici le plasmide mitochondrial, (environ 20nt) et se termine par une région d'environ 80nt d'une région conservée. Les séquences cibles ont été sélectionnées à partir de la séquence du plasmide (NC_001400.1) à l'aide du logiciel Benchling (voir Annexes). La région conservée de l'ARN guide permet l'interaction avec la protéine Cas9. 1 pM de chaque oligonucléotide synthétisé a été utilisé pour transcrire *in vitro* les deux ARN guides à l'aide du kit Quick T7 High Yield RNA synthesis Kit (New England BioLabs) en suivant les instructions du fabricant. La qualité des ARN transcrits a été évaluée sur gel.

Digestion de l'ADN par CRISPR Cas9. La digestion des plasmides mitochondriaux a été réalisée à partir du kit *in vitro* « digestion with Cas9 nuclease, *S. pyogenes* » (New England Biolabs). Les réactions ont été effectuées avec 6 ou 9 µl de chaque ARN guide et 2 ou 3 µl d'enzyme Cas9 pour 1,45 µg d'ADN. Les réactions ont été incubées à 37°C pendant 4h. Afin de neutraliser la réaction, une digestion avec une protéinase K a été réalisée à 65°C pendant 15 minutes. Une seconde réaction de PlasmidSafe a été réalisée dans les mêmes conditions que la première. L'ADN a ensuite été précipité par l'ajout de 0,1 volume d'acétate de sodium (3M), 2,5 volumes d'éthanol et 1 µl de glycogène. L'ADN circulaire précipité a ensuite été amplifié par une amplification aléatoire circulaire en utilisant le kit Illustra TempliPhi (GE

Healthcare). Pour cela, l'ADN circulaire précipité a été resuspendu directement dans le tampon Illustra TempliPhi et la réaction a été réalisée en suivant les instructions du kit à l'exception du temps d'incubation qui a été allongé à 65h à 28°C.

Validation par PCR. Les réactions de PCR ont été réalisées à partir de 2 µl de chaque produit de la réaction TempliPhi dilué au 100e dans un volume final de 15 µl en utilisant la polymérase GoTaq (Promega). Les amorces ont été dessinées à l'aide du logiciel Primer3 (voir Annexes) et la qualité a été évaluée à l'aide du logiciel OligoCalc (voir Annexes) et alignées sur la séquence du plasmide. Les conditions de PCR utilisées sont les suivantes : une première étape de dénaturation à 95°C pendant 5 minutes suivie de 30 cycles à 95°C pendant 30 secondes, une étape d'hybridation pendant 30 secondes, une étape d'élongation à 72°C pendant 1 minute et 10 secondes et une étape finale d'extension à 72°C pendant 5 minutes. 8 µl du produit PCR ont été déposés sur un gel d'agarose à 1,5% et la migration a été effectuée à 135 mV pendant 30 minutes. L'ADN est rendu visible par le GelRed dye (Biotium). Les images des gels ont été obtenues avec le système d'imagerie UGenius gel (Syngene).

- BIBLIOGRAPHIE -

- Allen SE, Hug I, Pabian S, Rzeszutek I, Hoehener C, Nowacki M. 2017. Circular concatemers of ultra-short DNA segments produce regulatory RNAs. *Cell* **168**: 990–999.
- Anderson SN, Johnson CS, Jones DS, Conrad LJ, Gou X, Russell SD, Sundaresan V. 2013. Transcriptomes of isolated *Oryza sativa* gametes characterized by deep sequencing: evidence for distinct sex-dependent chromatin and epigenetic states before fertilization. *Plant J* **76**: 729–741.
- Ausin I, Feng S, Yu C, Liu W, Kuo HY, Jacobsen EL, Zhai J, Gallego-Bartolome J, Wang L, Egertsdotter U, et al. 2016. DNA methylome of the 20-gigabase Norway spruce genome. *Proceedings of the National Academy of Sciences* **113**: E8106–E8113.
- Avery OT, MacLeod CM, McCarty MD. 1944. Studies on the chemical nature of the substance inducing transformation of pneumococcal types. *Journal of Experimental Medicine* **29**: 137–158.
- Baillie JK, Barnett MW, Upton KR, Gerhardt DJ, Richmond TA, De Sapio F, Brennan PM, Rizzu P, Smith S, Fell M, et al. 2011. Somatic retrotransposition alters the genetic landscape of the human brain. *Nature* **479**: 534–537.
- Baucom RS, Estill JC, Chaparro C, Upshaw N, Jogi A, Deragon J-M, Westerman RP, SanMiguel PJ, Bennetzen JL. 2009. Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. *PLoS Genet* **5**: e1000732.
- Baudry C, Malinsky S, Restituito M, Kapusta A, Rosa S, Meyer E, Bétermier M. 2009. PiggyMac, a domesticated piggyBac transposase involved in programmed genome rearrangements in the ciliate *Paramecium tetraurelia*. *Genes Dev* **23**: 2478–2483.
- Becker C, Hagemann J, Müller J, Koenig D, Stegle O, Borgwardt K, Weigel D. 2011. Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome. *Nature* **480**: 245–249.
- Bennetzen JL, Kellogg EA. 1997. Do plants have a one-way ticket to genomic obesity? *The Plant Cell* **9**: 1509–1514.
- Bhattacharyya N, Roy P. 1986. Extrachromosomal DNA from a dicot plant *Vigna radiata*. *FEBS Letters* **208**: 386–390.
- Britten RJ, Kohne DE. 1968. Repeated sequences in DNA. *Science* **161**: 529–540.
- Browning SR, Browning BL. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* **81**: 1084–1097.
- Burns KH. 2017. Transposable elements in cancer. *Nature Reviews Cancer* **17**: 415–424.
- Bushman FD. 2003. Targeting survival: integration site selection by retroviruses and LTR-retrotransposons. *Cell* **115**: 135–138.
- Butelli E, Licciardello C, Zhang Y, Liu J, Mackay S, Bailey P, Reforgiato-Recupero G, Martin C. 2012. Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. *The Plant Cell* **24**: 1242–1255.
- Cavrak VV, Lettner N, Jamge S, Kosarewicz A, Bayer LM, Mittelsten Scheid O. 2014. How a

- retrotransposon exploits the plant's heat stress response for its activation. *PLoS Genet* **10**: e1004115.
- Chargaff E. 1950. Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. *Cell Mol Life Sci.* **6**: 201–204
- Cheng C, Tarutani Y, Miyao A, Ito T, Yamazaki M, Sakai H, Fukai E, Hirochika H. 2015a. Loss of function mutations in the rice chromomethylase OsCMT3a cause a burst of transposition. *Plant J* **83**: 1069–1081.
- Cho J, Paszkowski J. 2017. Regulation of rice root development by a retrotransposon acting as a microRNA sponge. *eLife* 2017;6:e30038.
- Chuong EB, Elde NC, Feschotte C. 2016a. Regulatory activities of transposable elements: from conflicts to benefits. *Nature Reviews Genetics* **18**: 71–86
- Chuong EB, Elde NC, Feschotte C. 2016b. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* **351**: 1083–1087.
- Clarke JD. 2009. Cetyltrimethyl ammonium bromide (CTAB) DNA miniprep for plant DNA isolation. *Cold Spring Harb Protoc* **2009**: pdb.prot5177.
- Cohen S, Houben A, Segal D. 2008. Extrachromosomal circular DNA derived from tandemly repeated genomic sequences in plants. *Plant J* **53**: 1027–1034.
- Cohen S, Lavi S. 1996. Induction of circles of heterogeneous sizes in carcinogen-treated cells: two-dimensional gel analysis of circular DNA molecules. *Molecular and Cellular Biology* **16**: 2002–2014.
- Cohen S, Mechali M. 2001. A novel cell-free system reveals a mechanism of circular DNA formation from tandem repeats. *Nucleic Acids Res* **29**: 2542–2548.
- Cohen S, Méchali M. 2002. Formation of extrachromosomal circles from telomeric DNA in *Xenopus laevis*. *EMBO Rep* **3**: 1168–1174.
- Cohen S, Regev A, Lavi S. 1997. Small polydispersed circular DNA (spcDNA) in human cells: association with genomic instability. *Oncogene* **14**: 977–985.
- Cohen S, Segal D. 2009. Extrachromosomal circular DNA in eukaryotes: possible involvement in the plasticity of tandem repeats. *Cytogenet Genome Res* **124**: 327–338.
- Cohen Z, Bacharach E, Lavi S. 2006. Mouse major satellite DNA is prone to eccDNA formation via DNA Ligase IV-dependent pathway. *Oncogene* **25**: 4515–4524.
- Coil D, Jospin G, Darling AE. 2015. A5-miseq: an updated pipeline to assemble microbial genomes from Illumina MiSeq data. *Bioinformatics* **31**: 587–589.
- Copetti D, Zhang J, El-Baidouri M, Gao D, Wang J, Barghini E, Cossu RM, Angelova A, Maldonado L CE, Roffler S, et al. 2015. RiTE database: a resource database for genus-wide rice genomics and evolutionary biology. *BMC Genomics* **16**: 538.
- Crick F. 1970. Central dogma of molecular biology. *Nature* **227**: 561–563.

- Davidson EH, Britten RJ. 1979. Regulation of gene expression: possible role of repetitive sequences. *Science* **204**: 1052–1059
- Debladis E, Llauro C, Carpentier M-C, Mirouze M, Panaud O. 2017. Detection of active transposable elements in *Arabidopsis thaliana* using Oxford Nanopore Sequencing technology. *BMC Genomics* **18**: 537.
- Denner J. 2016. Expression and function of endogenous retroviruses in the placenta. *APMIS* **124**: 31–43.
- Dillon LW, Kumar P, Shibata Y, Wang Y-H, Willcox S, Griffith JD, Pommier Y, Takeda S, Dutta A. 2015. Production of extrachromosomal microDNAs is linked to mismatch repair pathways and transcriptional activity. *Cell Reports* **11**: 1749–1759.
- Dimitri P. 1997. Constitutive heterochromatin and transposable elements in *Drosophila melanogaster*. *Genetica* **100**: 85–93.
- El Baidouri M, Panaud O. 2012. Genome-wide analysis of transposition using next generation sequencing technologies. *Topics in Current Genetics* **24**: 59–70.
- El Baidouri M, Carpentier MC, Cooke R, Gao D, Lasserre E, Llauro C, Mirouze M, Picault N, Jackson SA, Panaud O. 2014. Widespread and frequent horizontal transfers of transposable elements in plants. *Genome Research* **24**: 831–838.
- Evers T, Millar S. 2002. Cereal grain structure and development: some implications for quality. *Journal of Cereal Science* **36**: 261–284
- Ewing AD. 2015. Transposable element detection from whole genome sequence data. *Mobile DNA* **6**: 24.
- Faye B, Arnaud F, Peyretailade E, Brassat E, Dastugue B, Vaudry C. 2008. Functional characteristics of a highly specific integrase encoded by an LTR-retrotransposon. *PLoS ONE* **3**(9): e3185
- Feschotte C. 2008. Transposable elements and the evolution of regulatory networks. *Nature Reviews Genetics* **9**: 397–405.
- Feschotte C, Pritham EJ. 2007. DNA transposons and the evolution of eukaryotic genomes. *Annu Rev Genet* **41**: 331–368.
- Finnegan DJ. 1989. Eukaryotic transposable elements and genome evolution. *Trends Genet* **5**: 103–107.
- Flavell AJ, Ish-Horowicz D. 1981. Extrachromosomal circular copies of the eukaryotic transposable element *copia* in cultured *Drosophila* cells. *Nature* **292**: 591–595.
- Flavell AJ, Ish-Horowicz D. 1983. The origin of extrachromosomal circular *copia* elements. *Cell* **34**: 415–419.
- Fultz D, Choudury SG, Slotkin RK. 2015. Silencing of active transposable elements in plants. *Current Opinion in Plant Biology* **27**: 67–76.
- Gao X, Hou Y, Ebina H, Levin HL, Voytas DF. 2008. Chromodomains direct integration of

- retrotransposons to heterochromatin. *Genome Research* **18**: 359–369.
- Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, et al. 2002. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**: 498–511.
- Gaubatz JW. 1990. Extrachromosomal circular DNAs and genomic sequence plasticity in eukaryotic cells. *Mutat Res* **237**: 271–292.
- Ghose T. 2004. Oswald Avery: the professor, DNA, and the Nobel Prize that eluded him. *Canadian Bulletin of Medical History* **21**:135-44.
- Gogvadze E, Buzdin A. 2009. Retroelements and their impact on genome evolution and functioning. *Cell Mol Life Sci* **66**: 3727–3742.
- Goic B, Stapleford KA, Frangeul L, Doucet AJ, Gausson V, Blanc H, Schemmel-Jofre N, Cristofari G, Lambrechts L, Vignuzzi M, et al. 2016. Virus-derived DNA drives mosquito vector tolerance to arboviral infection. *Nature Communications* **7**: 12410.
- Goic B, Vodovar N, Mondotte JA, Monot C, Frangeul L, Blanc H, Gausson V, Vera-Otarola J, Cristofari G, Saleh M-C. 2013. RNA-mediated interference and reverse transcription control the persistence of RNA viruses in the insect model *Drosophila*. *Nat Immunol* **14**: 396–403.
- Grandbastien M-A. 2015. LTR retrotransposons, handy hitchhikers of plant regulation and stress response. *Biochim Biophys Acta* **1849**: 403–416.
- Grant-Downton R, Kourmpetli S, Hafidh S, Khatab H, Le Trionnaire G, Dickinson H, Twell D. 2013. Artificial microRNAs reveal cell-specific differences in small RNA activity in pollen. *Curr Biol* **23**: R599–601.
- Gu W, Crawford ED, O'Donovan BD, Wilson MR, Chow ED, Retallack H, DeRisi JL. 2016. Depletion of Abundant Sequences by Hybridization (DASH): using Cas9 to remove unwanted high-abundance species in sequencing libraries and molecular counting applications. *Genome Biology* **17**: 41.
- Hahn PJ. 1993. Molecular biology of double-minute chromosomes. *Bioessays* **15**: 477–484.
- Hayashi K, Yoshida H. 2009. Refunctionalization of the ancient rice blast disease resistance gene Pit by the recruitment of a retrotransposon as a promoter. *Plant J* **57**: 413–425.
- Hirochika H. 1997. Retrotransposons of rice: their regulation and use for genome analysis. *Plant Mol Biol* **35**: 231–240.
- Hirochika H, Otsuki H. 1995. Extrachromosomal circular forms of the tobacco retrotransposon *Tto1*. *Gene* **165**: 229–232.
- Hirochika H, Sugimoto K, Otsuki Y, Tsugawa H, Kanda M. 1996. Retrotransposons of rice involved in mutations induced by tissue culture. *Proceedings of the National Academy of Sciences* **93**: 7783–7788.
- Hirsch CD, Springer NM. 2017. Transposable element influences on gene expression in plants. *Biochim Biophys Acta* **1860**: 157–165.

- Hsieh T-F, Ibarra CA, Silva P, Zemach A, Eshed-Williams L, Fischer RL, Zilberman D. 2009. Genome-wide demethylation of Arabidopsis endosperm. *Science* **324**: 1451–1454.
- Ibarra CA, Feng X, Schoft VK, Hsieh TF, Uzawa R, Rodrigues JA, Zemach A, Chumak N, Machlicova, Nishimura T, Rojas D, Fischer RL, Tamaru H, Zilberman D. 2012. Active DNA demethylation in plant companion cells reinforces transposon methylation in gametes. *Science* **337**: 1357–1360.
- Ingouff M, Rademacher S, Holec S, Šoljić L, Xin N, Readshaw A, Foo SH, Lahouze B, Sprunck S, Berger F. 2010. Zygotic resetting of the HISTONE 3 variant repertoire participates in epigenetic reprogramming in Arabidopsis. *Curr Biol* **20**: 2137–2143.
- Ingouff M, Selles B, Michaud C, Vu TM, Berger F, Schorn AJ, Autran D, Van Durme M, Nowack MK, Martienssen RA, et al. 2017. Live-cell analysis of DNA methylation during sexual reproduction in Arabidopsis reveals context and sex-specific dynamics controlled by noncanonical RdDM. *Genes Dev* **31**: 72–83.
- Ito H, Gaubert H, Bucher E, Mirouze M, Vaillant I, Paszkowski J. 2011. An siRNA pathway prevents transgenerational retrotransposition in plants subjected to stress. *Nature* **472**: 115–119.
- Jahn CL, Klobutcher LA. 2002. Genome remodeling in ciliated protozoa. *Annu Rev Microbiol* **56**: 489–520.
- Jeon JS, Lee S, Jung KH, Jun SH, Jeong DH, Lee J, Kim C, Jang S, Yang K, Nam J, et al. 2000. T-DNA insertional mutagenesis for functional genomics in rice. *Plant J* **22**: 561–570.
- Jiang N, Bao Z, Zhang X, Hirochika H, Eddy SR, McCouch SR, Wessler SR. 2003. An active DNA transposon family in rice. *Nature* **421**: 163–167.
- Jiang N, Ferguson AA, Slotkin RK. 2011. Pack-Mutator-like transposable elements (Pack-MULEs) induce directional modification of genes through biased insertion and DNA acquisition. *Proceedings of the National Academy of Sciences* **108**: 1537–1542.
- Jiang N, Feschotte C, Zhang X, Wessler SR. 2004. Using rice to understand the origin and amplification of miniature inverted repeat transposable elements (MITEs). *Current Opinion in Plant Biology* **7**: 115–119.
- Jordan IK, Rogozin IB, Glazko GV, Koonin EV. 2003. Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet* **19**: 68–72.
- Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S-Y, Freimer NB, Sabatti C, Eskin E. 2010. Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* **42**: 348–354.
- Kawabe A, Nasuda S. 2005. Structure and genomic organization of centromeric repeats in Arabidopsis species. *Molecular Genetics and Genomics*. **72**: 593-602
- Kawashima T, Lorković ZJ, Nishihama R, Ishizaki K, Axelsson E, Yelagandula R, Kohchi T, Berger F. 2015. Diversification of histone H2A variants during plant evolution. *Trends in Plant Science* **20**: 419–425.
- Kilzer JM, Stracker T, Beitzel B, Meek K, Weitzman M, Bushman FD. 2003. Roles of host cell

- factors in circularization of retroviral dna. *Virology* **314**: 460–467.
- Komatsu M, Chujo A, Nagato Y, Shimamoto K. 2003. FRIZZY PANICLE is required to prevent the formation of axillary meristems and to establish floral meristem identity in rice spikelets. *Development* **130**: 3841–3850.
- Kumar P, Dillon LW, Shibata Y, Jazaeri AA, Jones DR, Dutta A. 2017. Normal and cancerous tissues release extrachromosomal circular DNA (eccDNA) into the circulation. *Mol Cancer Res*. **15**:1197-1205.
- Kunarso G, Chia N-Y, Jeyakani J, Hwang C, Lu X, Chan Y-S, Ng HH, Bourque G. 2010. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat Genet* **42**: 631–634.
- Kuttler F, Mai S. 2007. Formation of non-random extrachromosomal elements during development, differentiation and oncogenesis. *Semin Cancer Biol* **17**: 56–64.
- Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.
- Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**: 357–359.
- Lerat E, Capy P. 1999. Retrotransposons and retroviruses: analysis of the envelope gene. *Mol Biol Evol* **16**: 1198–1207.
- Li H, Freeling M, Lisch D. 2010. Epigenetic reprogramming during vegetative phase change in maize. *Proceedings of the National Academy of Sciences* **107**: 22184–22189.
- Li L, Olvera JM, Yoder KE, Mitchell RS. 2001. Role of the non-homologous DNA end joining pathway in the early steps of retroviral infection. *The EMBO Journal* **20**: 3272–3281.
- Lippman Z, Gendrel A-V, Black M, Vaughn MW, Dedhia N, McCombie WR, Lavine K, Mittal V, May B, Kasschau KD, et al. 2004. Role of transposable elements in heterochromatin and epigenetic control. *Nature* **430**: 471–476.
- Lisch D. 2013. How important are transposons for plant evolution? *Nature Reviews Genetics* **14**: 49–61.
- Liu S, Yeh C-T, Ji T, Ying K, Wu H, Tang HM, Fu Y, Nettleton D, Schnable PS. 2009. Mu transposon insertion sites and meiotic recombination events co-localize with epigenetic marks for open chromatin across the maize genome. *PLoS Genet* **5**: e1000733.
- Liu SJ, Xu HH, Wang WQ, Li N, Wang WP. 2015. A proteomic analysis of rice seed germination as affected by high temperature and ABA treatment. *Physiologia Plantarum*. **154**:142-61.
- Liu Y, Fang J, Xu F, Chu J, Yan C, Schläppi MR, Wang Y, Chu C. 2014. Expression patterns of ABA and GA metabolism genes and hormone levels during rice seed development and imbibition: a comparison of dormant and non-dormant rice cultivars. *Journal of Genetics and Genomics* **41**: 327–338.

- Lopato S, Borisjuk N, Langridge P, Hrmova M. 2014. Endosperm transfer cell-specific genes and proteins: structure, function and applications in biotechnology. *Front Plant Sci* **5**: 64.
- Lu C, Chen J, Zhang Y, Hu Q, Su W, Kuang H. 2012. Miniature inverted-repeat transposable elements (MITEs) have been accumulated through amplification bursts and play important roles in gene expression and species diversity in *Oryza sativa*. *Mol Biol Evol* **29**: 1005–1017.
- Ludwig SR, Pohlman RF, Vieira J, Smith AG, Messing J. 1985. The nucleotide sequence of a mitochondrial replicon from maize. *Gene* **38**: 131–138.
- Makarevitch I, Waters AJ, West PT, Stitzer M, Hirsch CN, Ross-Ibarra J, Springer NM. 2015. Transposable elements contribute to activation of maize genes in response to abiotic stress. *PLoS Genet* **11**: e1004915.
- Martínez G, Köhler C. 2017. Role of small RNAs in epigenetic reprogramming during plant sexual reproduction. *Current Opinion in Plant Biology* **36**: 22–28.
- Martínez G, Panda K, Köhler C, Slotkin RK. 2016. Silencing in sperm cells is directed by RNA movement from the surrounding nurse cell. *Nat Plants* **2**: 16030.
- McCarthy EM, Liu J, Lizhi G, McDonald JF. 2002. Long terminal repeat retrotransposons of *Oryza sativa*. *Genome Biology* **3**: research0053.
- McClintock B. 1984. The significance of responses of the genome to challenge. *Science* **226**: 792–801.
- Menéndez-Arias L, Sebastián-Martín A, Álvarez M. 2017. Viral reverse transcriptases. *Virus Res* **234**: 153–176.
- Mhiri C, Morel JB, Vernhettes S, Casacuberta JM, Lucas H, Grandbastien MA. 1997. The promoter of the tobacco *Tnt1* retrotransposon is induced by wounding and by abiotic stress. *Plant Mol Biol* **33**: 257–266.
- Mills RE, Bennett EA, Iskow RC, Devine SE. 2007. Which transposable elements are active in the human genome? *Trends Genet* **23**: 183–191.
- Mirouze M, Vitte C. 2014. Transposable elements, a treasure trove to decipher epigenetic variation: insights from Arabidopsis and crop epigenomes. *J Exp Bot* **65**: 2801–2812.
- Moon S, Jung K-H, Lee D-E, Jiang W-Z, Koh HJ, Heu M-H, Lee DS, Suh HS, An G. 2006. Identification of active transposon dTok, a member of the hAT family, in rice. *Plant and Cell Physiology* **47**: 1473–1483.
- Morton T, Petricka J, Corcoran DL, Li S, Winter CM, Carda A, Benfey PN, Ohler U, Megraw M. 2014. Paired-end analysis of transcription start sites in Arabidopsis reveals plant-specific promoter signatures. *The Plant Cell* **26**: 2746–2760.
- Mourier T. 2016. Potential movement of transposable elements through DNA circularization. *Curr Genet* **62**: 697–700.
- Muñoz-López M, García-Pérez JL. 2010. DNA transposons: nature and applications in genomics. *Current Genomics* **11**: 115–128.

- Møller HD, Larsen CE, Parsons L, Hansen AJ, Regenberg B, Mourier T. 2015a. Formation of extrachromosomal circular DNA from Long Terminal Repeats of retrotransposons in *Saccharomyces cerevisiae*. *G3* **6**: 453–462.
- Møller HD, Parsons L, Jørgensen TS, Botstein D, Regenberg B. 2015b. Extrachromosomal circular DNA is common in yeast. *Proceedings of the National Academy of Sciences* **112**: E3114–22.
- Nadeau NJ, Pardo-Diaz C, Whibley A, Supple MA, Saenko SV, Wallbank RWR, Wu GC, Maroja L, Ferguson L, Hanly JJ, et al. 2016. The gene cortex controls mimicry and crypsis in butterflies and moths. *Nature* **534**: 106–110.
- Naito K, Cho E, Yang G, Campbell MA, Yano K, Okumoto Y, Tanisaka T, Wessler SR. 2006. Dramatic amplification of a rice transposable element during recent domestication. *Proceedings of the National Academy of Sciences* **103**: 17620–17625.
- Naito K, Zhang F, Tsukiyama T, Saito H, Hancock CN, Richardson AO, Okumoto Y, Tanisaka T, Wessler SR. 2009. Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. *Nature* **461**: 1130–1134.
- Nie D-M, Ouyang Y-D, Wang X, Zhou W, Hu C-G, Yao J. 2013. Genome-wide analysis of endosperm-specific genes in rice. *Gene* **530**: 236–247.
- Nozawa K, Kawagishi Y, Kawabe A, Sato M, Masuta Y. 2017. Epigenetic regulation of a heat-activated retrotransposon in cruciferous vegetables. *Epigenomes* **1**: 7.
- Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin Y-C, Scofield DG, Vezzi F, Delhomme N, Giacomello S, Alexeyenko A, et al. 2013. The Norway spruce genome sequence and conifer genome evolution. *Nature* **497**: 579–584.
- Ohyanagi H, Ebata T, Huang X, Gong H, Fujita M, Mochizuki T, Toyoda A, Fujiyama A, Kaminuma E, Nakamura Y, et al. 2016. OryzaGenome: Genome Diversity Database of Wild Oryza Species. *Plant and Cell Physiology* **57**: e1(1–7)
- Orgel LE, Crick FH. 1980. Selfish DNA: the ultimate parasite. *Nature* **284**: 604–607.
- Panaud O. 2016. Horizontal transfers of transposable elements in eukaryotes: The flying genes. *C R Biol* **339**: 296–299.
- Park K, Kim MY, Vickers M, Park JS, Hyun Y, Okamoto T, Zilberman D, Fischer RL, Feng X, Choi Y, Scholten S. 2016. DNA demethylation is initiated in the central cells of Arabidopsis and rice. *Proceedings of the National Academy of Sciences* **113**: 15138–15143.
- Pennisi E. 2017. Circular DNA throws biologists for a loop. *Science* **356**: 996.
- Pereira V. 2004. Insertion bias and purifying selection of retrotransposons in the *Arabidopsis thaliana* genome. *Genome Biology* **5**: R79.
- Philippe C, Vargas-Landin DB, Doucet AJ, van Essen D, Vera-Otarola J, Kuciak M, Corbin A, Nigumann P, Cristofari G. 2016. Activation of individual L1 retrotransposon instances is restricted to cell-type dependent permissive loci. *eLife* **5**: e13926.
- Phung NTP, Mai CD, Hoang GT, Truong HTM, Lavarenne J, Gonin M, Nguyen KL, Ha TT, Do VN,

- Gantet P, et al. 2016. Genome-wide association mapping for root traits in a panel of rice accessions from Vietnam. *BMC Plant Biol* **16**: 64.
- Phung NTP, Mai CD, Mournet P, Frouin J, Droc G, Ta NK, Jouannic S, Lê LT, Do VN, Gantet P, et al. 2014. Characterization of a panel of Vietnamese rice varieties using DArT and SNP markers for association mapping purposes. *BMC Plant Biol* **14**: 371.
- Picault N, Chaparro C, Piégu B, Stenger W, Formey D, Llauro C, Descombin J, Sabot F, Lasserre E, Meynard D, et al. 2009. Identification of an active LTR retrotransposon in rice. *Plant J* **58**: 754–765.
- Piégu B, Bire S, Arensburger P, Bigot Y. 2015. A survey of transposable element classification systems - A call for a fundamental update to meet the challenge of their diversity and complexity. *Mol Phylogenet Evol* **86**: 90–109.
- Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**: 904–909.
- Qu LQ, Xing YP, Liu WX, Xu XP, Song YR. 2008. Expression pattern and activity of six glutelin gene promoters in transgenic rice. *J Exp Bot* **59**: 2417–2424.
- Quadrana L, Bortolini Silveira A, Mayhew GF, LeBlanc C, Martienssen RA, Jeddloh JA, Colot V. 2016. The *Arabidopsis thaliana* mobilome and its impact at the species level. *eLife* **5**: e15716.
- Raghavan V. 2003. Some reflections on double fertilization, from its discovery to the present. *New Phytologist* **159**: 565–583.
- Richardson SR, Morell S, Faulkner GJ. 2014. L1 retrotransposons and somatic mosaicism in the brain. *Annu Rev Genet* **48**: 1–27.
- Sallaud C, Meynard D, van Boxtel J, Gay C, Bès M, Brizard JP, Larmande P, Ortega D, Raynal M, Portefaix M, et al. 2003. Highly efficient production and characterization of T-DNA plants for rice (*Oryza sativa* L.) functional genomics. *Theor Appl Genet* **106**: 1396–1408.
- SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL. 1998. The paleontology of intergene retrotransposons of maize. *Nature Genetics* **20**: 43–45.
- Sarafianos SG, Das K, Tantillo C, Clark AD. 2001. Crystal structure of HIV-1 reverse transcriptase in complex with a polypurine tract RNA: DNA. *The EMBO Journal* **20**: 1449–1461.
- Schnable PS, Ware D, Fulton RS, Stein JC, Wei F et al. 2009. The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**: 1112–1115.
- Schoft VK, Chumak N, Choi Y. 2011. Function of the DEMETER DNA glycosylase in the *Arabidopsis thaliana* male gametophyte. *Proceedings of the National Academy of Sciences* **108**: 8042–8047.
- Schulman AH. 2013. Retrotransposon replication in plants. *Current Opinion in Virology* **3**: 604–614.
- Scott EC, Devine SE. 2017. The role of somatic L1 retrotransposition in human cancers. *Viruses* **9** : 131.

- Shahmuradov IA, Umarov RK, Solovyev VV. 2017. TSSPlant: a new tool for prediction of plant Pol II promoters. *Nucleic Acids Res* **45**: e65.
- Shan Q, Wang Y, Li J, Zhang Y, Chen K, Liang Z, Zhang K, Liu J, Xi JJ, Qiu J-L, et al. 2013. Targeted genome modification of crop plants using a CRISPR-Cas system. *Nature Biotechnology* **31**: 686–688.
- Shibata Y, Kumar P, Layer R, Willcox S, Gagan JR, Griffith JD, Dutta A. 2012. Extrachromosomal microDNAs and chromosomal microdeletions in normal tissues. *Science* **336**: 82–86.
- Shoura MJ, Gabdank I, Hansen L, Merker J, Gotlib J, Levene SD, Fire AZ. 2017. Intricate and cell-type-specific populations of endogenous circular DNA (eccDNA) in *Caenorhabditis elegans* and *Homo sapiens*. G3. 300141.
- Siefert JL. 2009. Defining the mobilome. *Methods Mol Biol* **532**: 13–27.
- Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. 2017. Detecting DNA cytosine methylation using nanopore sequencing. *Nature Methods* **14**: 407–410.
- Slotkin RK, Martienssen R. 2007. Transposable elements and the epigenetic regulation of the genome. *Nature Reviews Genetics* **8**: 272–285.
- Slotkin RK, Vaughn M, Borges F, Tanurdzić M, Becker JD, Feijó JA, Martienssen RA. 2009. Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. *Cell* **136**: 1451–1454.
- Song X, Cao X. 2017. Transposon-mediated epigenetic regulation contributes to phenotypic diversity and environmental adaptation in rice. *Current Opinion in Plant Biology* **36**: 111–118.
- Stroud H, Ding B, Simon SA, Feng S, Bellizzi M, Pellegrini M, Wang G-L, Meyers BC, Jacobsen SE. 2013. Plants regenerated from tissue culture contain stable epigenome changes in rice. *eLife* **2**: e00354.
- Sultana T, Zamborlini A, Cristofari G, Lesage P. 2017. Integration site selection by retroviruses and transposable elements in eukaryotes. *Nature Reviews Genetics* **18**: 292–308.
- Sundaresan V, Freeling M. 2007. An extrachromosomal form of the Mu transposons of maize. *Proceedings of the National Academy of Sciences* **84**: 1–5.
- The 3,000 rice genomes project. 2014. *GigaScience* **3**: 1–6.
- Ton-Hoang B, Pasternak C, Siguier P, Guynet C, Hickman AB, Dyda F, Sommer S, Chandler M. 2010. Single-stranded DNA transposition is coupled to host replication. *Cell* **142**: 398–408.
- Tsukahara S, Kawabe A, Kobayashi A, Ito T, Aizu T, Shin-i T, Toyoda A, Fujiyama A, Tarutani Y, Kakutani T. 2012. Centromere-targeted de novo integrations of an LTR retrotransposon of *Arabidopsis lyrata*. *Genes Dev* **26**: 705–713.
- Turner KM, Deshpande V, Beyter D, Koga T, Rusert J, Lee C, Li B, Arden K, Ren B, Nathanson DA, et al. 2017. Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. *Nature* **543**: 122–125.

- Turner SD. 2014. qqman: an R package for visualizing GWAS results using QQ and manhattan plots. *BioRxiv*. doi.org/10.1101/005165
- Van den Broeck D, Maes T, Sauer M, Zethof J, De Keukeleire P, D'hauw M, Van Montagu M, Gerats T. 1998. Transposon Display identifies individual transposable elements in high copy number lines. *Plant J* **13**: 121–129.
- Van't Hof AE, Campagne P, Rigden DJ, Yung CJ, Lingley J, Quail MA, Hall N, Darby AC, Saccheri IJ. 2016. The industrial melanism mutation in British peppered moths is a transposable element. *Nature* **534**: 102–105.
- Vaughan DA, Lu BR, Tomooka N. 2008. The evolving story of rice evolution. *Plant science* **174**: 394–408
- Vitte C, Panaud O, Quesneville H. 2007. LTR retrotransposons in rice (*Oryza sativa*, L.): recent burst amplifications followed by rapid DNA loss. *BMC Genomics* **8**: 218.
- Waddington CH. 1940. Organisers and genes. *Cambridge biological studies*.
- Waddington CH. 1942. The epigenotype. *International Journal of Epidemiology* **41**: 10–13.
- Watson JD, Crick FH. 1953. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature* **171**:737–738
- Wei F-J, Droc G, Guiderdoni E, Hsing Y-IC. 2013. International consortium of rice mutagenesis: resources and beyond. *Rice* **6**: 39.
- Wei L, Cao X. 2016. The effect of transposable elements on phenotypic variation: insights from plants to humans. *Science China Life Sciences* **59**: 24–37.
- Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, et al. 2007. A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics* **8**: 973–982.
- Worden AZ, Lee JH, Mock T, Rouzé P. 2009. Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes *Micromonas*. *Science* **324**: 268–272.
- Wright SI, Agrawal N, Bureau TE. 2003. Effects of recombination rate and gene density on transposable element distributions in *Arabidopsis thaliana*. *Genome Research* **13**: 1897–1903.
- Xie M, Hong C, Zhang B, Lowdon RF, Xing X, Li D, Zhou X, Lee HJ, Maire CL, Ligon KL, et al. 2013. DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. *Nat Genet* **45**: 836–841.
- Xing M-Q, Zhang Y-J, Zhou S-R, Hu W-Y, Wu X-T, Ye Y-J, Wu X-X, Xiao Y-P, Li X, Xue H-W. 2015. Global analysis reveals the crucial roles of DNA methylation during rice seed development. *Plant Physiology* **168**: 1417–1432.
- Xu H, Bao JD, Dai JS, Li Y, Zhu Y. 2015. Genome-wide identification of new reference genes for qRT-PCR normalization under high temperature stress in rice endosperm. *PLoS ONE* **10**: e0142015.

- Xue L-J, Zhang JJ, Xue H-W. 2012. Genome-wide analysis of the complex transcriptional networks of rice developing seeds. *PLoS ONE* **7**: e31081.
- Yang L, Bennetzen JL. 2009. Structure-based discovery and description of plant and animal Helitrons. *Proceedings of the National Academy of Sciences* **106**: 12832–12837.
- Yoshihara T, Washida H, Takaiwa F. 1996. A 45-bp proximal region containing AACAA and GCN4 motif is sufficient to confer endosperm-specific expression of the rice storage protein glutelin gene, *GluA-3*. *FEBS Lett* **383**: 213–218.
- Zemach A, Kim MY, Silva P, Rodrigues JA, Dotson B, Brooks MD, Zilberman D. 2010. Local DNA hypomethylation activates genes in rice endosperm. *Proceedings of the National Academy of Sciences* **107**: 18729–18734.
- Zhang Q, Arbuckle J, Wessler SR. 2000. Recent, extensive, and preferential insertion of members of the miniature inverted-repeat transposable element family *Heartbreaker* into genic regions of maize. *Proceedings of the National Academy of Sciences* **97**: 1160–1165.
- Zhou S-R, Yin L-L, Xue H-W. 2013. Functional genomics based understanding of rice endosperm development. *Current Opinion in Plant Biology* **16**: 236–246.

- ANNEXES-








Espèce	Génotype	Conditions / Tissus	Résultats	Collaborateurs
 <i>Arabidopsis thaliana</i> (135 Mb)	WT	zébularine + α -amanitin	ONSEN (Thieme et al., 2017)	Etienne Bucher (INRA Angers)
	mutants épigénétiques <i>atr56, morc</i>	feuille	pas de candidat intéressant	Steve Jacobsen (UCLA, USA)
 <i>Drosophila melanogaster</i> (175 Mb)	WT	virus à ARN	ADNecc provenant des virus	Carla Saleh (Institut Pasteur)
	mutant épigénétique (pivi mutant)	tissu ovarien	ZAM	Severine Chambeyron (IGH)
 <i>Oryza sativa ssp japonica & ssp indica</i> (400 Mb)	WT	virus à ARN (RYMV)	pas de candidat intéressant	Laurence Albar (IRD)
	mutant RNAi DDM1	feuille	Détection d'un SINE dans le WT mais pas de candidat intéressant dans le mutant	Yoshiki Habu (NIAS, Japon)
	WT	zébularine + α -amanitin	Houba (Thieme et al., 2017)	Etienne Bucher (INRA Angers)
 <i>Zea Mays</i> (3 Gb)	WT	grain	à refaire avec CRISPR digestion	Jonn Laurie (SLCU, Cambridge)
 <i>Triticum aestivum</i> cv. chinese spring (17 Gb)	WT	stress thermique	pas de candidat intéressant	Jérôme Salse (INRA)
 <i>Populus tremula x alba</i> (500 Mb)	WT	stress hydrique	pas de candidat intéressant	Stéphane Maury (Univ. Orléans)
	mutant RNAi DDM1	feuille	Détection de 2 transposons à ADN	Stéphane Maury (Univ. Orléans)
	mutant RNAi DDM1	stress hydrique	Détection de 2 transposons à ADN	Stéphane Maury (Univ. Orléans)
 <i>Picea abies</i> (20 Gb)	WT	aiguille de pin, culture <i>in vitro</i> , bourgeons floraux	RT à LTR dans les bourgeons floraux, 1 RT à LTR dans les aiguilles de pin	Steve Jacobsen (UCLA, USA)

Tableau S1. Ensemble des collaborations pour lesquelles j'ai participé à l'analyse.

Table S1. Coordonnées des séquences uniques.

gene ID	chromosome	start	end
Os01t0101150	chr01	57658	60090
Os01t0969000	chr01	42736008	42737917
Os01t0970700	chr01	42826168	42828686
Os01t0974701	chr01	43071516	43072507
Os01t0976300	chr01	43141847	43142667
Os02t0100250	chr02	12563	13520
Os02t0100600	chr02	33032	33966
Os02t0100700	chr02	34262	40938
Os02t0462200	chr02	15496218	15498439
Os02t0462300	chr02	15502622	15503795
Os02t0462401	chr02	15505302	15507245
Os02t0462900	chr02	15524149	15527804
Os02t0465600	chr02	15659380	15660788
Os02t0467200	chr02	15725827	15726670
Os02t0467500	chr02	15729134	15732855
Os03t0103700	chr03	262997	264249
Os03t0106000	chr03	356871	359717
Os03t0107800	chr03	471470	472258
Os03t0108600	chr03	505922	509002
Os04t0103351	chr04	227030	228665
Os04t0105050	chr04	327298	328772
Os04t0105300	chr04	354021	355335
Os04t0105400	chr04	357603	360450
Os04t0105450	chr04	359245	360369
Os04t0105700	chr04	370795	372419
Os04t0107700	chr04	471138	472885
Os05t0101300	chr05	95909	96901
Os05t0102300	chr05	145073	147497
Os05t0102400	chr05	148472	149341
Os05t0102500	chr05	149964	152034
Os05t0102800	chr05	160715	165048
Os05t0104800	chr05	260101	262268
Os05t0105000	chr05	267986	274638
Os05t0105550	chr05	301690	303708
Os05t0597200	chr05	29759285	29761362
Os06t0102750	chr06	190346	191129
Os06t0103500	chr06	227847	232614
Os06t0104000	chr06	276729	282165
Os06t0104050	chr06	277318	281892
Os06t0104100	chr06	282348	284538
Os07t0100500	chr07	36140	42057
Os07t0102000	chr07	115444	117742
Os07t0103200	chr07	196515	199802

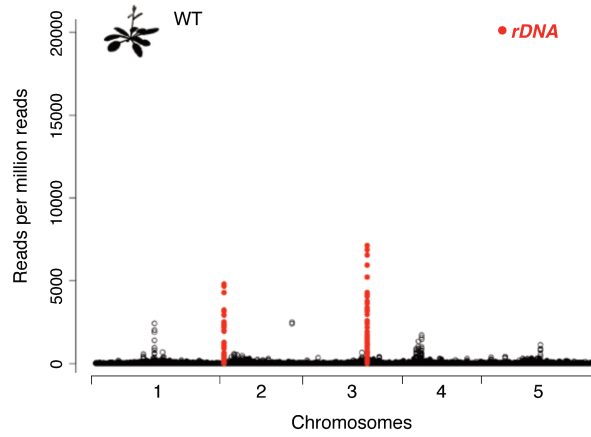
gene ID	chromosome	start	end
Os07t0105300	chr07	302259	303435
Os07t0107900	chr07	428798	431565
Os07t0105900	chr07	333276	334779
Os08t0100200	chr08	18289	24063
Os08t0100300	chr08	24352	25840
Os08t0100500	chr08	36145	45629
Os08t0100700	chr08	54055	58284
Os08t0101100	chr08	71348	72715
Os08t0104100	chr08	222484	225233
Os08t0105000	chr08	272854	276978
Os09t0104300	chr09	490959	493966
Os09t0121075	chr09	1655097	1656212
Os09t0122000	chr09	1726956	1728653
Os09t0123100	chr09	1780100	1786097
Os09t0123200	chr09	1790628	1799469
Os10t0102400	chr10	186845	188331
Os10t0105850	chr10	432396	433522
Os10t0105900	chr10	435646	436611
Os10t0116000	chr10	1034806	1037595
Os10t0117800	chr10	1	3657
Os11t0424400	chr11	13377937	13379488
Os11t0427500	chr11	13632428	13637209
Os11t0429100	chr11	13744266	13749580
Os11t0433600	chr11	14028186	14029896
Os11t0437600	chr11	14267738	14268964
Os11t0442900	chr11	14605584	14607289
Os12t0276100	chr12	10208672	10211101
Os12t0277000	chr12	10254122	10257442
Os12t0641600	chr12	27521758	27523074
Os12t0640550	chr12	27461619	27464059
Os12t0639100	chr12	27407374	27410751
Os12t0639000	chr12	27402825	27403416
Os12t0638900	chr12	27398742	27400640
Os12t0638800	chr12	27394294	27398419

Nom amorce	Utilisation	Séquence foward (5'-3')	Séquence réverse (5'-3')	Amplicon (bp)	Tm (°C)
1F_1R	Digestion CRISPR	TTTAACGGTACGGTAAGGACAAGT	CTTCTTGTCGCCCATCTCTTTAAC	1417	58
2F_2R	Digestion CRISPR	GAGGGCATTCAATTCTATGTGCAA	CCCTCATAAATCCCTCCAATCCAT	1420	58
Hpa1_F/R	Digestion <i>Hpa</i> I	CGTCTCAACACTCTCTTGCAGAAAAT	CCACGCTGATGATAGAATTCCTCAAC	889	60
eEF1a_F/R	PCR gène ménage (eEF1 α)	GATCTGGTAAGGAGCTGGAGAAGG	CCGTGCACAAAACACTACTTGAA	313	58
PRint_F/R	PCR <i>PopRice</i> interne	CTTCCTCCAAGTAGCTTCGGATGA	CTCGTTAGCTGGCAGTCAATCAAG	285	60
PRecc_F/R	PCR ADNecc <i>PopRice</i>	ACAAACTGCTGTCCTAACTGTCCT	GCAGCTATAAATATGTATCCAATCCT	258	55
qPCR1_F/R	RT-qPCR spécifique <i>PopRice</i>	CGACACCAGCAAGAGCACGAG	GGATGAAATGGTACCTCACTCGGA	293	60
qPCR2_F/R	RT-qPCR <i>Osr4</i>	TTGCTCGACTGCTTAGTGAT	GGACTTGCTATCCACCCTGA	71	60
<i>PopRice_sonde1</i>	Southern blot	GTTATTTCTGCTGCTCGTCGAC	GTCGCCGAGAAGATCCTCCATC	1022	57
<i>PopRice_sonde2</i>	Southern blot	GACTTGAAGGAGGAGGTCTACGTG	GGATGAAATGGTACCTCACTCGGA	1000	60
I8_F/R	Validation insertion 8	CCCATGGGCTTTAATTTCTGTGATT	ATGCTTTATTAGTTGGATGCCC	272	58
I11_F/R	Validation insertion 11	ACAGGAAAGTGTGGCTCTGATAC	ATACTAAATTGAGCGCCCAAGTCT	250	58

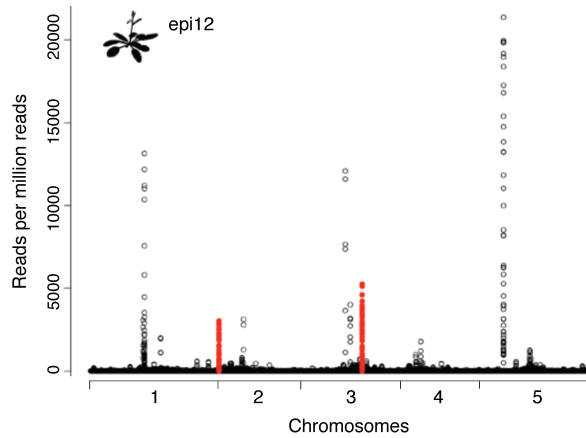
Tableau S2. Séquences et caractéristiques des amorces et des sondes utilisées.

Lanciano et al. SUPPLEMENTARY FIGURES

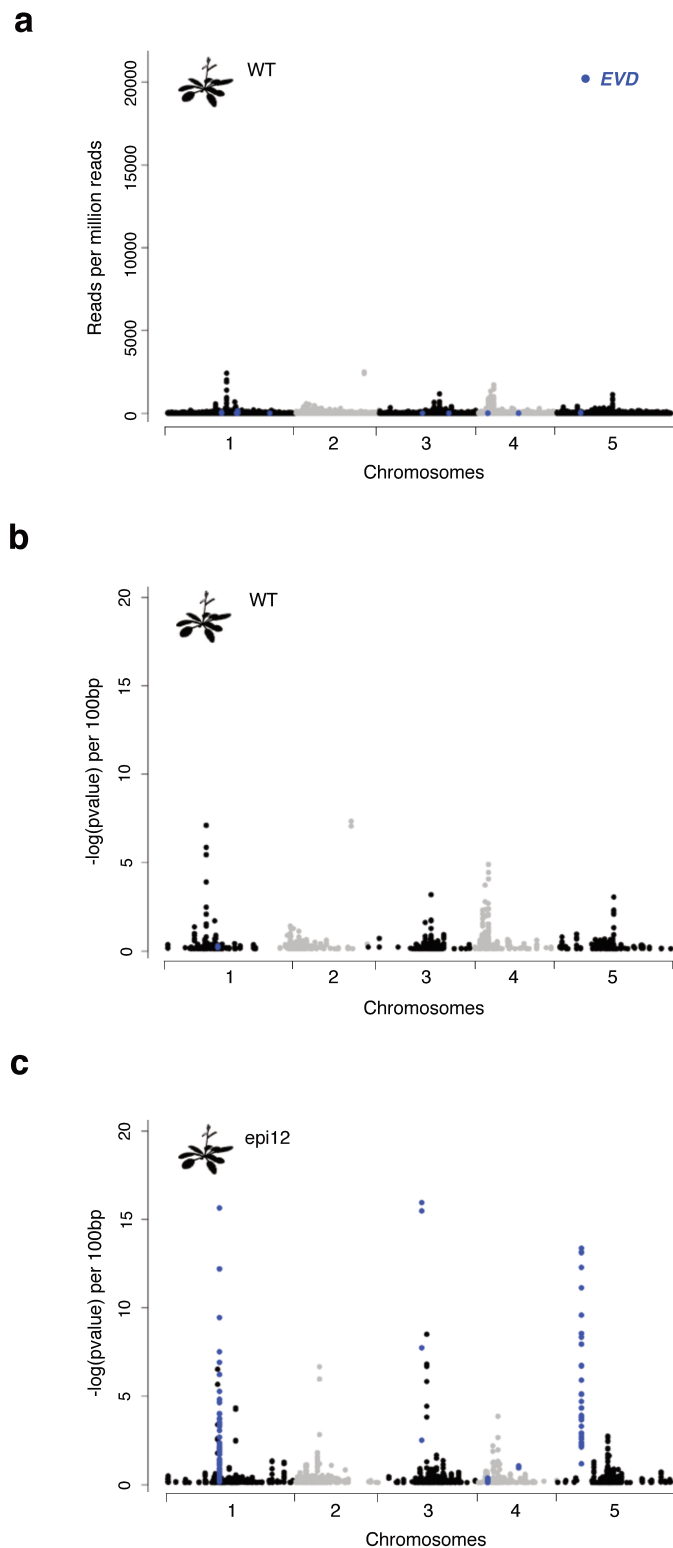
a



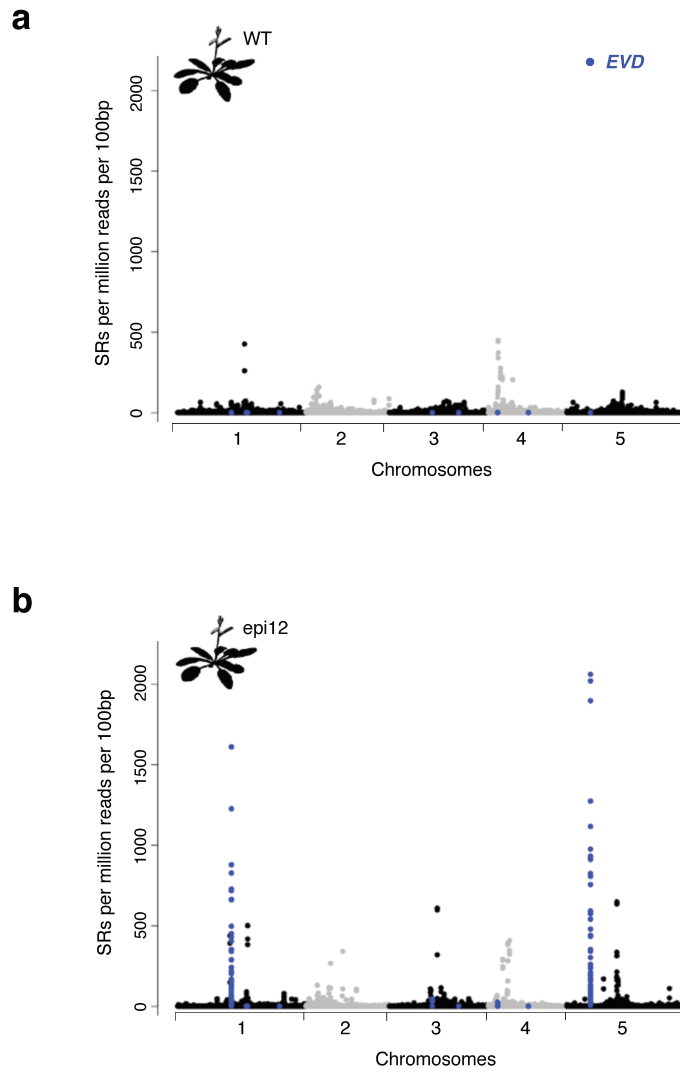
b



Supplementary Figure 2. Detection of eccDNAs originating from ribosomal DNA repeats in *A. thaliana*. (a) Abundance of reads mapping at TE-annotated and rDNA loci in the *A. thaliana* WT mobilome-seq library. Each dot represents the normalized coverage per million mapped reads per all TE-containing (*black circles*) or rDNA containing (*red dots*) 100bp windows obtained after aligning the sequenced reads on the reference genome. (b) Abundance of reads mapping at TE-annotated and rDNA loci in the *A. thaliana* epi12 mobilome-seq library. Legend as in (a).



Supplementary Figure 3. Analysis of *A. thaliana* mobilome-seq libraries. (a) Abundance of reads mapping at TE-annotated loci in the *A. thaliana* WT mobilome-seq library. (b) Statistical analysis of the WT mobilome-seq library presented in (a). (c) Statistical analysis of the epi12 mobilome-seq library presented in Figure 2a. Legend as in Figure 2a.



Supplementary Figure 4. Split read (SR) analysis of *A. thaliana* mobilome-seq libraries. (a) Abundance of SRs mapping at TE-annotated loci in the *A. thaliana* WT mobilome-seq library. (b) Abundance of SRs mapping at TE-annotated loci in the *A. thaliana* epi12 mobilome-seq library. Legend as in Figure 2a.

```

READ:      1 TCTCTTTGTGTCTTTAAAAGCATTGTAACACACAAAGTTACTATCTAATTCATCA 60
eccDNA EVD: 247 TCTCTTTGTGTCTTTAAAAGCATTGTAACACACAAAGTTACTATCTAATTCATCA 188

READ:      61 ATATGATTTGGTCTCATATCTCTCACATACAACTCTCTGTCTTCATAACTCTCTTA 120
eccDNA EVD: 187 ATATGATTTGGTCTCATATCTCTCACATACAACTCTCTGTCTTCATAACTCTCTTA 128

READ:      121 ATCTTTAATACAATCCGCATATCTTTCA---TGATCAAGACTCAAATAAGAAAGCCT 177
eccDNA EVD: 127 ATCTTTAATACAATCCGCATATCTTTCAAGATTGATCAAGACTCAAATAAGAAAGCCT 68

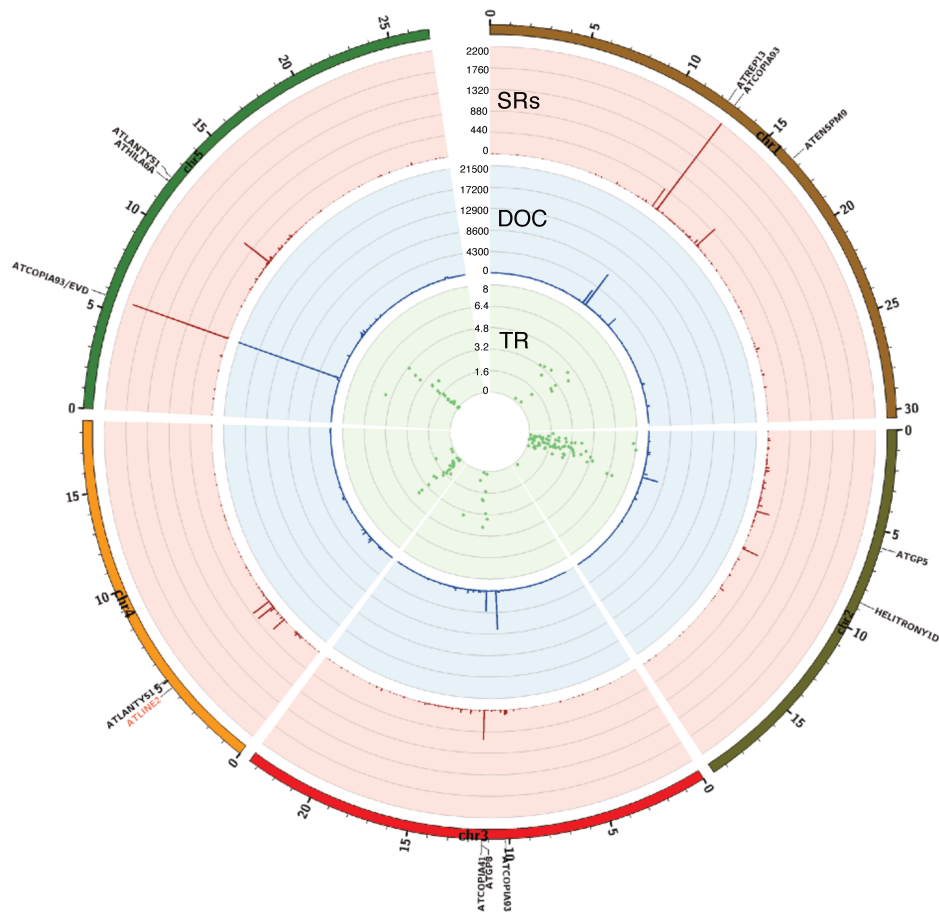
READ:      178 AGTATTGGATATGTAATAAGAGAGTGGGCCGAACATATGAGAAGCTATGAAGAGCTT 237
eccDNA EVD: 67 AGTATTGGATATGTAATAAGAGAGTGGGCCGAACATATGAGAAGCTATGAAGAGCTT 8

READ:      238 CTAGAAG 244
eccDNA EVD: 7 CTAGAAG 1

```

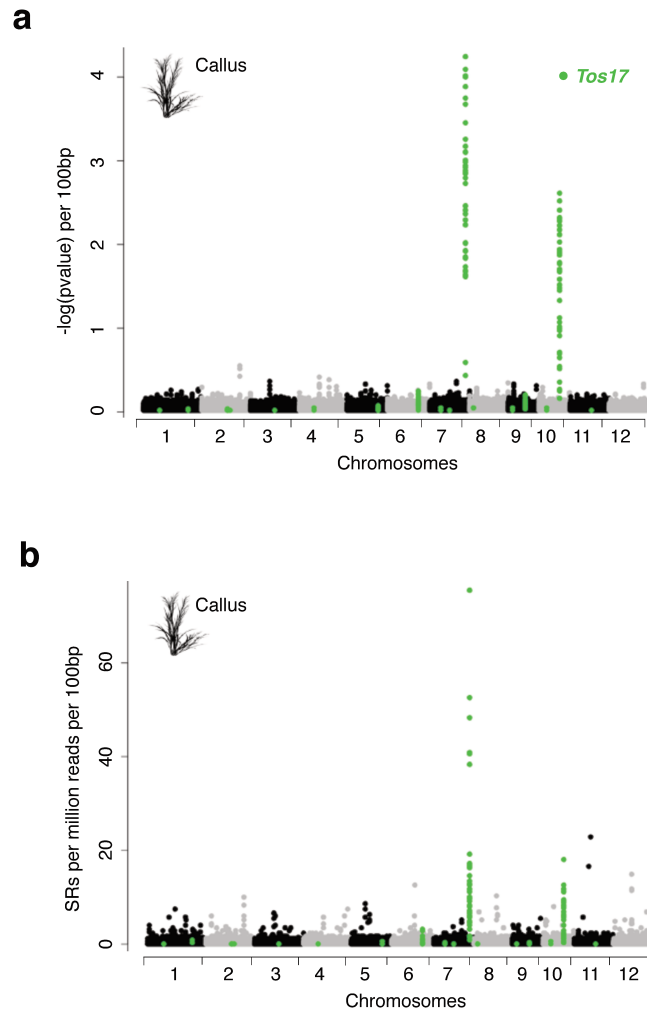
3' LTR
 5' LTR

Supplementary Figure 5. EVD form eccDNAs in *A.thaliana* epi12 line. Example of a SR identified in the epi12 mobilome-seq library spanning the junction of the 2LTR-circle corresponding to *EVD* aligned with an artificial junction corresponding to the 3' part of the 3'LTR (*red box*) fused to the 5' part of the 5'LTR (*yellow box*).



Supplementary Figure 6. Comparison between mobilome and transcriptome data.

A Circos plot of mobilome-seq and transcriptome data from *A. thaliana* epi12. The tracks, from outermost to innermost, show TEs coverage per million reads per 100bp for mobilome-seq split reads (SRs, *red track*), TEs coverage per million reads per 100bp for mobilome-seq depth of coverage for all aligned reads (DOC, *blue track*) represented as histograms and distribution of transcriptome coverage at TEs as a scatter plot (TR, *green track*). Transcriptome data are presented as the \log_2 of fold change in epi12 versus WT (Mirouze et al, 2012). The tracks are scaled separately to show modification fluctuations. The chromosomes sizes are indicated in megabase pairs. Names of TEs corresponding to the highest mobilome-seq peaks are indicated.



Supplementary Figure 7. Analysis of the *O. sativa* WT callus mobilome-seq libraries. (a) Statistical analysis of the mobilome-seq library presented in Figure 2c. Legend as in Figure 2c. (b) Abundance of SR mapping at TE-annotated loci in the WT callus mobilome-seq library.

b

```
Read:          40 TCCACCTTGAGTTTGAAGGGGGTGTAAATATATATACAAGCTAATGTACTGTATAGTT 99
eccDNA Tos17: 251 TCCACCTTGAGTGTGAAGGGGGTGTAAATATATATACAAGCTAATGTACTGGTAGTT 192

Read:          100 GGCCCATGTCCAGCCATCGGATGTCCAGCCATTGGATCTTGTATCTTGTATATACTTC 159
eccDNA Tos17: 191 GGCCCATGTCCAGCCATCGGATGTCCAGTCCATTGGATCTTGTATCTTGTATATACTTC 132

Read:          160 TCTATTGCTAATACTATTGTTAGGTTGCAAGTTAGTTAAGATGTTAAATATATATACAAG 219
eccDNA Tos17: 131 TCTATTGCTAATACTATTGTTAGGTTGCAAGTTAGTTAAGATGTTAAATATATATACAAG 72

Read:          220 CTAATGTACTGTATAGTTGGCCCATGTCCAGCCATCGGATGTCCAGCCATTGGATCTT 279
eccDNA Tos17: 71 CTAATGTACTGTATAGTTGGCCCATGTCCAGCCATCGGATGTCCAGTCCATTGGATCTT 12

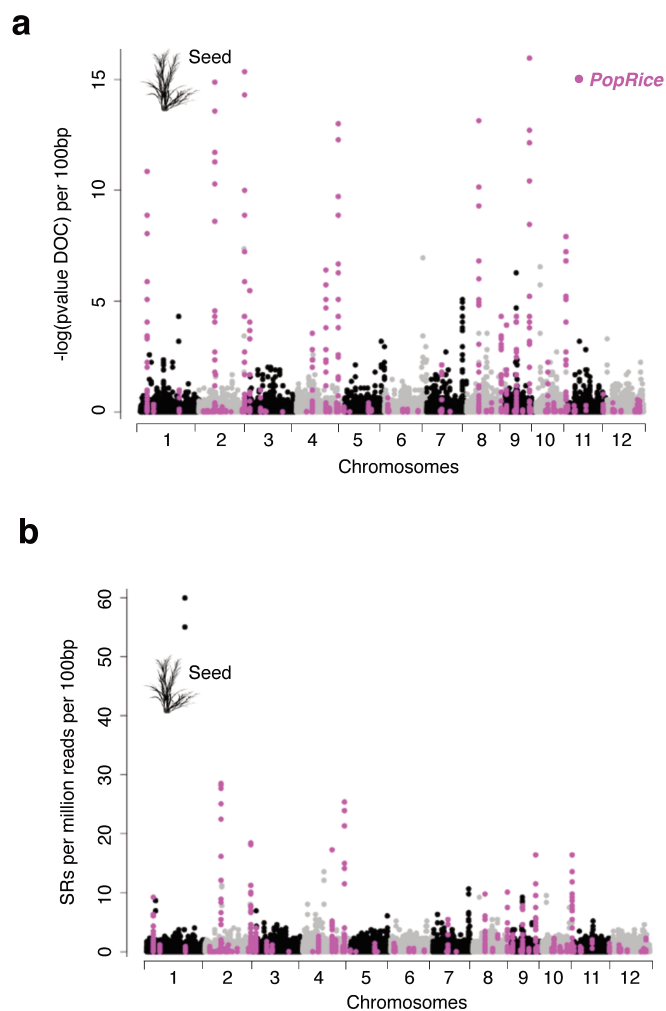
Read:          280 GTATCTTGAT 290
eccDNA Tos17: 11 GTATCTTGAT 1
```

c

```
Read:          1 GGCCAACAGTAAAATAAGTCTAAGACTACATCGCTATACACATATCTAACATGTTGAATTG 60
eccDNA Lullaby: 201 GGCCAACAGTAAAATAAGTCTAAGACTACATCGCTATACACATATCTAACATGTTGAATTG 142

Read:          61 TGCTAAAATCAGTACGGTGTATTCAGGAGAGCACTAGCATATACAATGGTATATGGGCC 120
eccDNA Lullaby: 141 TGCTAAAATCAGTACGGTGTATTCAGGAGAGCACTAGCATATACAATGGTATATGGGCC 82
```

Supplementary Figure 8. *Tos17* and *Lullaby* form eccDNAs in rice calli. (a) Circular forms of *Tos17* are specifically detected in calli using inverse PCR with primers localization depicted on the right (*black bar*: *Tos17* element, *arrows*: PCR primers, *grey boxes*: LTRs). Upper gel: PCR amplification of *Tos17* circles, middle gel: control PCR for *Tos17* detection, lower gel: PCR using eEF1 α primers as loading control. (b) Example of split read spanning the perfect junction of the 2LTR-circle corresponding to *Tos17*. Legend as in Supplementary Figure 5. (c) Example of split read spanning the perfect junction of the 2LTR-circle corresponding to *Lullaby*.



Supplementary Figure 9. Analysis of the *O. sativa* seed mobilome-seq libraries. (a) Statistical analysis of the mobilome-seq library presented in Figure 3a. Legend as in Figure 3A. **(b)** Abundance of split reads mapping at TE-annotated loci in the *O. sativa* WT seed mobilome-seq library presented in Figure 3a.

a

```
Read:          2 CAACAACTGCTCTAGAGCATCTCAACAACTGCTGTCCTAACTGCTCTAGAGCATCAA 61
eccDNA PopRice: 300 CAACAACTGCTCTAGAGCATCTCAACAACTGCTGTCCTAACTGCTCTAGAGCATCAA 241

Read:          62 ACTGAAAGTAGAGATGCTGTCTTAGAAAGCTGAAAAGCAGGAAAAAGGCACCACAAGACC 121
eccDNA PopRice: 240 ACTGAAAGTAGAGATGCTGTCTTAGAAAGCTGAAAAGCAGGAAAAAGGCACCACAAGACC 181

Read:          122 AAGACCAACCAGCCAAGACTTATTCCTACA TGTAGAATTTCAATGGTTACTCTCATTGA 181
eccDNA PopRice: 180 AAGACCAACCAGCCAAGACTTATTCCTACA TGTAGAATTTCAATGGTTACTCTCATTGA 121

Read:          182 GAAGAGGATGGCACTCGAGATGGGGCAATTTCTGGTTTCTTCATTCACTAAACACAAA 241
eccDNA PopRice: 120 GAAGAGGATGGCACTCGAGATGGGGCAATTTCTGGTTTCTTCATTCACTAAACACAAA 61

Read:          242 AGCCATGTC 250
eccDNA PopRice: 60 AGCCATGTC 52
```

b

```
Read:          1 TCCAAACTGAAAGTAGAGATGCTGCTTAGAAAGCTGAAAAGCAGGAAAAAGGCACCACA 60
eccDNA PopRice: 245 TCCAAACTGAAAGTAGAGATGCTGCTTAGAAAGCTGAAAAGCAGGAAAAAGGCACCACA 186

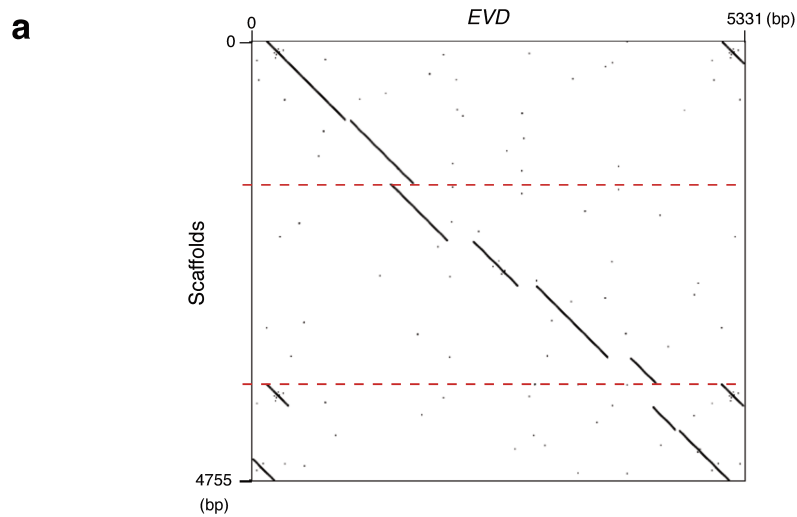
Read:          61 AGACCAAGACCAACCAGCCAAGACTTATTCCTACA AGCCTACAGGAAAGTGTGGCTCTGA 120
eccDNA PopRice: 185 AGACCAAGACCAACCAGCCAAGACTTATTCCTACA ----- 151

Read:          121 TACCAGATGTTAGAATTTCAATGGTTACTCTCATTGAGAAGAGGATGGCACTCGAGATGG 180
eccDNA PopRice: 150 -----TGTAGAATTTCAATGGTTACTCTCATTGAGAAGAGGATGGCACTCGAGATGG 98

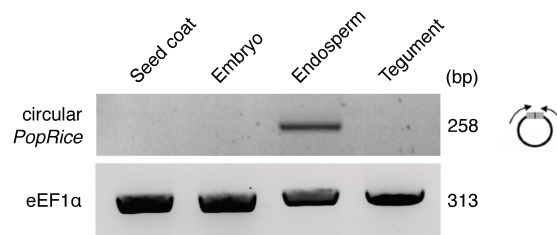
Read:          181 GGCAATTTCTGGTTTCTTCATTCACTAAACACAAAAGCCATGTCCTCCCAAAGAGGAT 240
eccDNA PopRice: 97 GGCAATTTCTGGTTTCTTCATTCACTAAACACAAAAGCCATGTCCTCCCAAAGAGGAT 38

Read:          241 TGGATACATAT 251
eccDNA PopRice: 37 TGGATACATAT 27
```

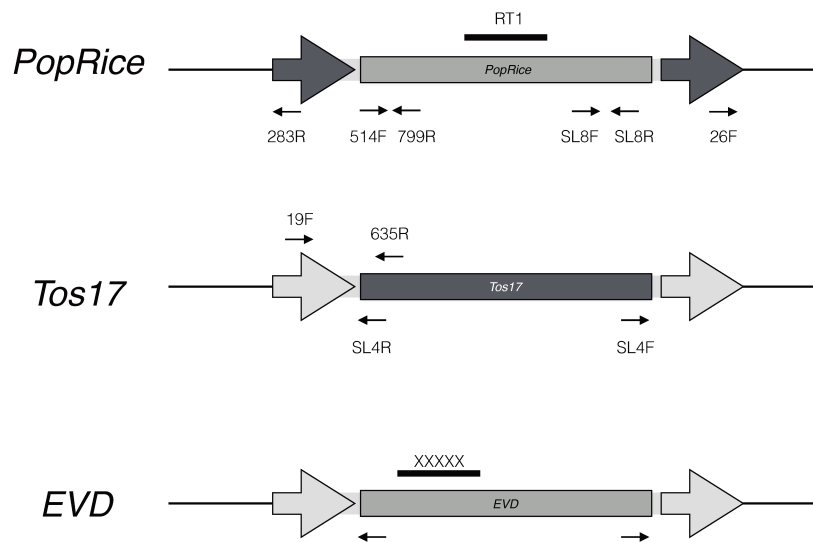
Supplementary Figure 10. *PopRice* form eccDNAs in rice seeds. Legend as in Supplementary Figure 5a. (a) Example of split read spanning the perfect junction of the 2LTR-circle corresponding to *PopRice*. (b) Example of split read spanning the unperfect junction of the 2LTR-circle corresponding to *PopRice*. A Primer finding Site (PBS) sequence is reprinted in blue. The PBS is normally found after the 5'LTR in a linear *PopRice*.



Supplementary Figure 11. *De novo* assembly of eccDNAs. (a) Example of scaffolds obtained after *de novo* assembly of epi12 mobilome-seq library and corresponding to *EVD*. The presence of many scaffolds (and not only one) suggests that *EVD* forms a complex population of circles. (b) Example of the scaffold #29 obtained after *de novo* assembly of callus mobilome-seq library and corresponding to *Tos17*.



Supplementary Figure 12. Detection of *PopRice* eccDNAs by PCR. Circular forms of *PopRice* are specifically detected in the dissected rice endosperm using inverse PCR. Legend as in Fig. 4a. PCR using eEF1α primers is used as a control.



Supplementary Figure 13. Positions of primers used in this study. Sequences of primers are shown in supplementary table 4.

Supplementary Table 1. Characteristics of the *A. thaliana* mobilome-seq libraries.

Analysis \ Library	WT		epi12	
	#1	#2	#1	#2
Library size (reads)	74,086	170,910	415,282	587,918
Reads mapping against organelles	18,288	39,808	161,924	151,128
Reads mapping against genome	31,373	69,160	132,390	208,653
Mean coverage per 100bp (rpm)	2.301	2.606	1.897	2.372
Total number of scaffolds (<i>de novo</i> assembly)	536	1548	863	2637
Number of <i>de novo</i> assembled scaffolds significantly covered ($p < 0.05$)	28	76	62	124
Mean coverage for split reads (segemehl mapping) per 100bp (rpm)	0.250	0.285	0.121	0.209

Supplementary Table 2. Characteristics of the *O. sativa* mobilome-seq libraries.

Analysis \ Library	Callus			Leaf		Seed	
	#1	#2	#3	#1	#2	#1	#2
Library size (reads)	1,941,638	4,012,812	1,715,560	2,948,984	2,754,020	2,183,780	1,313,588
Reads mapping against organelles	1,010,444	2,560,746	872,218	360,288	455,302	352,580	169,550
Reads mapping against genome	483,741	829,725	469,544	1,508,727	1,516,212	1,443,600	978,964
Mean coverage per 100bp (rpm)	0.707	0.662	0.692	0.648	0.614	0.599	0.675
Total number of scaffolds (<i>de novo</i> assembly)	1057	3377	849	10425	6558	7479	6506
Number of <i>de novo</i> assembled scaffolds significantly covered ($p < 0.05$)	85	200	62	325	237	423	316
Mean coverage for split reads (segemehl mapping) per 100bp (rpm)	0.030	0.043	0.037	0.076	0.13	0.058	0.040

Supplementary Table 3. Localization of *PopRice* and *osr4* elements in the *O. sativa ssp. japonica* cv. Nipponbare reference genome.

Name	Chromosome	Start	End
<i>PopRice_1</i>	chr01	4 776 239	4 781 940
<i>osr4_10</i>	chr01	9 619 047	9 624 779
<i>osr4_1</i>	chr02	3 433 922	3 439 683
<i>osr4_30</i>	chr02	4 810 186	4 815 148
<i>osr4_28</i>	chr02	6 298 856	6 304 425
<i>osr4_23</i>	chr02	10 230 034	10 235 735
<i>PopRice_2</i>	chr02	11 897 051	11 902 751
<i>osr4_11</i>	chr02	17 177 630	17 183 380
<i>osr4_29</i>	chr02	20 017 590	20 022 447
<i>osr4_24</i>	chr02	32 073 359	32 079 087
<i>PopRice_3</i>	chr02	34 010 137	34 015 809
<i>osr4_12</i>	chr03	1 805 958	1 811 710
<i>PopRice_4</i>	chr03	1 910 824	1 916 518
<i>osr4_2</i>	chr03	9 807 146	9 812 816
<i>osr4_22</i>	chr04	11 784 416	11 790 165
<i>PopRice_5</i>	chr04	21 858 331	21 863 990
<i>osr4_19</i>	chr04	24 624 991	24 630 730
<i>PopRice_6</i>	chr04	29 764 342	29 769 992
<i>PopRice_7</i>	chr04	31 205 979	31 211 679
<i>osr4_3</i>	chr05	18 149 395	18 155 117
<i>osr4_8</i>	chr05	18 173 196	18 178 845
<i>PopRice_8</i>	chr06	2 568 077	2 573 729
<i>PopRice_9</i>	chr06	3 207 031	3 212 695
<i>osr4_18</i>	chr06	13 707 057	13 712 806
<i>osr4_21</i>	chr06	24 968 651	24 974 392
<i>osr4_27</i>	chr07	10 183 108	10 188 828
<i>PopRice_10</i>	chr07	11 227 404	11 233 065
<i>PopRice_11</i>	chr08	9 051 840	9 057 543
<i>osr4_25</i>	chr08	14 745 774	14 751 442
<i>osr4_5</i>	chr08	25 668 348	25 674 136
<i>PopRice_12</i>	chr09	1 229 148	1 234 789
<i>osr4_9</i>	chr09	6 996 940	7 001 178
<i>PopRice_13</i>	chr09	8 572 145	8 577 847
<i>osr4_6</i>	chr09	18 114 586	18 120 338
<i>osr4_13</i>	chr09	19 370 956	19 376 699
<i>PopRice_14</i>	chr10	13 174 222	13 179 845
<i>osr4_20</i>	chr10	18 476 622	18 482 493
<i>PopRice_15</i>	chr10	22 300 481	22 306 180
<i>osr4_14</i>	chr10	22 402 727	22 408 469
<i>PopRice_16</i>	chr11	4 997 937	5 003 672
<i>PopRice_17</i>	chr11	26 689 013	26 694 658
<i>osr4_17</i>	chr12	4 851 245	4 856 938
<i>osr4_16</i>	chr12	5 895 106	5 900 860
<i>osr4_4</i>	chr12	11 047 463	11 053 165
<i>osr4_26</i>	chr12	21 273 154	21 278 677
<i>osr4_7</i>	chr12	23 149 361	23 155 098
<i>osr4_15</i>	chr12	24 767 171	24 772 919

Supplementary Table 4. List of primers used in this study.

Used	Primers name	Forward sequence (5'-3')	Reverse sequence (5'-3')	Amplicon (bp)	T _m (°C)
<i>EVD</i> probe (Southern blot)		TATTGATCAAGACTCAAATAAGAAAGG	TGAAAGAATATGCGGAATTGTATTTAA	406	55
Rice eEF1 α	2460F/2749R	GATCTGGTAAGGAGCTGGAGAAGG	CCGTGCACAAAACACTACCACTTGAA	313	58
<i>PopRice</i>	514F/799R	CTTCTCCAAGTAGCTTCGGATGA	CTCGTTAGCTGGCAGTCAATCAAG	285	60
circular <i>PopRice</i>	26F/283R	ACAAACTGCTGTCTAACTGTCTCT	GCAGCTATAAATATGTATCCAATCCT	258	55
<i>PopRice</i> RT-PCR	SL8F/SL8R	CGACACCAGCAAGAGCACGAG	GGATGAAATGGTACCTCACTCGGA	293	60
<i>PopRice</i> probe (Southern blot)	RT1F/RT1R	GTTATTTCTGCTGCTCGTCGAC	GTCGCCGAGAAGATCCTCCATC	1022	57
<i>Tos17</i>	19F/635R	CTAATGTACTGTATAGTTGGCCC	ACCAAAGATCACTAGCAACTC	581	60
circular <i>Tos17</i>	SL4F/SL4R	AACTCGAGAGCATCATCGGTTACA	CATTAGCTGTATGAACGGTGGCAC	1570-1428	62

Supplementary Table 5. Table with significant coverage values for 100 bp windows. For each library, the number of mapped reads per million per 100bp window is indicated with P-value < 10⁻³ (CHR: chromosome, BP: base pair start coordinate of the 100bp window).

CHR	BP	TE family	Reads per million reads	P-value	Mobilome library
1	13955100	ATLANTYS1	2024	1.37E-06	<i>A. thaliana</i> WT
1	13955200	ATLANTYS1	2415	7.85E-08	<i>A. thaliana</i> WT
1	13955300	ATLANTYS1	1894	3.55E-06	<i>A. thaliana</i> WT
1	13955400	ATLANTYS1	1403	1.25E-04	<i>A. thaliana</i> WT
2	16045500	ATDNAI27T9A	2400	8.73E-08	<i>A. thaliana</i> WT
2	16045600	ATDNAI27T9A	2487	4.61E-08	<i>A. thaliana</i> WT
3	14251300	ATLANTYS3	1171	6.49E-04	<i>A. thaliana</i> WT
4	2915300	ATLANTYS1	1345	1.89E-04	<i>A. thaliana</i> WT
4	3625500	ATLANTYS1	1576	3.57E-05	<i>A. thaliana</i> WT
4	3625600	ATLANTYS1	1721	1.25E-05	<i>A. thaliana</i> WT
4	3625700	ATLANTYS1	1460	8.22E-05	<i>A. thaliana</i> WT
5	13182100	ATMU1	1128	8.83E-04	<i>A. thaliana</i> WT
1	12362700	ATREP13	1548	4.09E-04	<i>A. thaliana</i> epi12
1	12363000	ATREP13	2641	2.15E-06	<i>A. thaliana</i> epi12
1	12363100	ATREP13	3053	2.99E-07	<i>A. thaliana</i> epi12
1	12754300	ATCOPIA93	1850	9.56E-05	<i>A. thaliana</i> epi12
1	12754400	ATCOPIA93	3532	3.03E-08	<i>A. thaliana</i> epi12
1	12754500	ATCOPIA93	3245	1.20E-07	<i>A. thaliana</i> epi12
1	12754600	ATCOPIA93	2152	2.24E-05	<i>A. thaliana</i> epi12
1	12754700	ATCOPIA93	1558	3.91E-04	<i>A. thaliana</i> epi12
1	12754900	ATCOPIA93	2909	5.95E-07	<i>A. thaliana</i> epi12
1	12755000	ATCOPIA93	7553	2.22E-16	<i>A. thaliana</i> epi12
1	12755100	ATCOPIA93	12188	0.00E+00	<i>A. thaliana</i> epi12
1	12755200	ATCOPIA93	13146	0.00E+00	<i>A. thaliana</i> epi12
1	12755300	ATCOPIA93	10366	0.00E+00	<i>A. thaliana</i> epi12
1	12755400	ATCOPIA93	4467	3.51E-10	<i>A. thaliana</i> epi12
1	12758400	ATCOPIA93	1701	1.95E-04	<i>A. thaliana</i> epi12
1	12758500	ATCOPIA93	2233	1.51E-05	<i>A. thaliana</i> epi12
1	12758600	ATCOPIA93	2454	5.26E-06	<i>A. thaliana</i> epi12
1	12758700	ATCOPIA93	1845	9.78E-05	<i>A. thaliana</i> epi12
1	12758800	ATCOPIA93	1716	1.82E-04	<i>A. thaliana</i> epi12
1	12758900	ATCOPIA93	1577	3.56E-04	<i>A. thaliana</i> epi12
1	12759000	ATCOPIA93	1505	5.04E-04	<i>A. thaliana</i> epi12
1	12759600	ATCOPIA93	1606	3.10E-04	<i>A. thaliana</i> epi12
1	12759700	ATCOPIA93	1395	8.58E-04	<i>A. thaliana</i> epi12
1	12760000	ATCOPIA93	5799	6.15E-13	<i>A. thaliana</i> epi12
1	12760100	ATCOPIA93	11196	0.00E+00	<i>A. thaliana</i> epi12
1	12760200	ATCOPIA93	11018	0.00E+00	<i>A. thaliana</i> epi12
1	12760300	ATCOPIA93	5799	6.15E-13	<i>A. thaliana</i> epi12
1	16583000	ATENSPM9	2003	4.57E-05	<i>A. thaliana</i> epi12
1	16583100	ATENSPM9	1965	5.50E-05	<i>A. thaliana</i> epi12
2	5829200	ATGP5	2785	1.08E-06	<i>A. thaliana</i> epi12
2	5829300	ATGP5	3125	2.12E-07	<i>A. thaliana</i> epi12
3	10181900	ATCOPIA93	3642	1.79E-08	<i>A. thaliana</i> epi12
3	10182000	ATCOPIA93	7659	1.11E-16	<i>A. thaliana</i> epi12
3	10182100	ATCOPIA93	11603	0.00E+00	<i>A. thaliana</i> epi12
3	10182200	ATCOPIA93	12077	0.00E+00	<i>A. thaliana</i> epi12
3	10182300	ATCOPIA93	7376	3.33E-16	<i>A. thaliana</i> epi12
3	11357300	ATCOPIA41	1759	1.48E-04	<i>A. thaliana</i> epi12

CHR	BP	TE family	Reads per million reads	P-value	Mobilome library
3	11357400	ATCOPIA41	3130	2.07E-07	<i>A. thaliana</i> epi12
3	11357500	ATCOPIA41	3192	1.54E-07	<i>A. thaliana</i> epi12
3	11357600	ATCOPIA41	4007	3.15E-09	<i>A. thaliana</i> epi12
3	11357700	ATCOPIA41	2722	1.45E-06	<i>A. thaliana</i> epi12
3	11357800	ATCOPIA41	2051	3.63E-05	<i>A. thaliana</i> epi12
4	4594400	ATLINE2	1778	1.35E-04	<i>A. thaliana</i> epi12
4	4594500	ATLINE2	1778	1.35E-04	<i>A. thaliana</i> epi12
5	5630000	ATCOPIA93	6240	7.55E-14	<i>A. thaliana</i> epi12
5	5630100	ATCOPIA93	13232	0.00E+00	<i>A. thaliana</i> epi12
5	5630200	ATCOPIA93	17263	0.00E+00	<i>A. thaliana</i> epi12
5	5630300	ATCOPIA93	19827	0.00E+00	<i>A. thaliana</i> epi12
5	5630400	ATCOPIA93	19966	0.00E+00	<i>A. thaliana</i> epi12
5	5630500	ATCOPIA93	19875	0.00E+00	<i>A. thaliana</i> epi12
5	5630600	ATCOPIA93	15384	0.00E+00	<i>A. thaliana</i> epi12
5	5630700	ATCOPIA93	13213	0.00E+00	<i>A. thaliana</i> epi12
5	5630800	ATCOPIA93	9988	0.00E+00	<i>A. thaliana</i> epi12
5	5630900	ATCOPIA93	8171	0.00E+00	<i>A. thaliana</i> epi12
5	5631000	ATCOPIA93	8200	0.00E+00	<i>A. thaliana</i> epi12
5	5631100	ATCOPIA93	6259	6.88E-14	<i>A. thaliana</i> epi12
5	5631200	ATCOPIA93	5281	7.23E-12	<i>A. thaliana</i> epi12
5	5631300	ATCOPIA93	4031	2.81E-09	<i>A. thaliana</i> epi12
5	5631400	ATCOPIA93	3930	4.53E-09	<i>A. thaliana</i> epi12
5	5631500	ATCOPIA93	3738	1.13E-08	<i>A. thaliana</i> epi12
5	5631600	ATCOPIA93	3139	1.98E-07	<i>A. thaliana</i> epi12
5	5631700	ATCOPIA93	2756	1.24E-06	<i>A. thaliana</i> epi12
5	5631800	ATCOPIA93	1749	1.55E-04	<i>A. thaliana</i> epi12
5	5631900	ATCOPIA93	1510	4.92E-04	<i>A. thaliana</i> epi12
5	5632900	ATCOPIA93	1802	1.20E-04	<i>A. thaliana</i> epi12
5	5633000	ATCOPIA93	1677	2.19E-04	<i>A. thaliana</i> epi12
5	5633100	ATCOPIA93	1740	1.62E-04	<i>A. thaliana</i> epi12
5	5633200	ATCOPIA93	1735	1.66E-04	<i>A. thaliana</i> epi12
5	5633300	ATCOPIA93	2181	1.95E-05	<i>A. thaliana</i> epi12
5	5633400	ATCOPIA93	2382	7.43E-06	<i>A. thaliana</i> epi12
5	5633500	ATCOPIA93	2368	7.96E-06	<i>A. thaliana</i> epi12
5	5633600	ATCOPIA93	2003	4.57E-05	<i>A. thaliana</i> epi12
5	5633700	ATCOPIA93	1682	2.14E-04	<i>A. thaliana</i> epi12
5	5633800	ATCOPIA93	1764	1.45E-04	<i>A. thaliana</i> epi12
5	5633900	ATCOPIA93	3158	1.81E-07	<i>A. thaliana</i> epi12
5	5634000	ATCOPIA93	4529	2.61E-10	<i>A. thaliana</i> epi12
5	5634100	ATCOPIA93	5837	5.12E-13	<i>A. thaliana</i> epi12
5	5634200	ATCOPIA93	8526	0.00E+00	<i>A. thaliana</i> epi12
5	5634300	ATCOPIA93	11833	0.00E+00	<i>A. thaliana</i> epi12
5	5634400	ATCOPIA93	14747	0.00E+00	<i>A. thaliana</i> epi12
5	5634500	ATCOPIA93	19175	0.00E+00	<i>A. thaliana</i> epi12
5	5634600	ATCOPIA93	18389	0.00E+00	<i>A. thaliana</i> epi12
5	5634700	ATCOPIA93	21361	0.00E+00	<i>A. thaliana</i> epi12
5	5634800	ATCOPIA93	18965	0.00E+00	<i>A. thaliana</i> epi12
5	5634900	ATCOPIA93	16798	0.00E+00	<i>A. thaliana</i> epi12
5	5635000	ATCOPIA93	13836	0.00E+00	<i>A. thaliana</i> epi12

CHR	BP	TE family	Reads per million reads	P-value	Mobilome library
5	5635100	ATCOPIA93	11037	0.00E+00	<i>A. thaliana</i> epi12
5	5635200	ATCOPIA93	6365	4.17E-14	<i>A. thaliana</i> epi12
7	26695100	LTR_fam158_tos17_expop1+	480	7.84E-04	<i>O. sativa</i> callus
7	26695200	LTR_fam158_tos17_expop1+	610	1.31E-04	<i>O. sativa</i> callus
7	26695300	LTR_fam158_tos17_expop1+	575	2.12E-04	<i>O. sativa</i> callus
7	26695400	LTR_fam158_tos17_expop1+	587	1.79E-04	<i>O. sativa</i> callus
7	26697900	LTR_fam158_tos17_expop1+	505	5.53E-04	<i>O. sativa</i> callus
7	26698000	LTR_fam158_tos17_expop1+	632	9.74E-05	<i>O. sativa</i> callus
7	26698100	LTR_fam158_tos17_expop1+	670	5.75E-05	<i>O. sativa</i> callus
7	26698200	LTR_fam158_tos17_expop1+	629	1.01E-04	<i>O. sativa</i> callus
7	26698300	LTR_fam158_tos17_expop1+	478	7.97E-04	<i>O. sativa</i> callus
7	26698500	LTR_fam158_tos17_expop1+	491	6.75E-04	<i>O. sativa</i> callus
7	26698600	LTR_fam158_tos17_expop1+	538	3.54E-04	<i>O. sativa</i> callus
7	26698700	LTR_fam158_tos17_expop1+	645	8.12E-05	<i>O. sativa</i> callus
7	26698800	LTR_fam158_tos17_expop1+	463	9.90E-04	<i>O. sativa</i> callus
1	4776300	LTR_fam51_osr4_poprice_expop1+	46	8.97E-05	<i>O. sativa</i> seed
1	4776400	LTR_fam51_osr4_poprice_expop1+	77	9.05E-09	<i>O. sativa</i> seed
1	4776500	LTR_fam51_osr4_poprice_expop1+	97	1.42E-11	<i>O. sativa</i> seed
1	4776600	LTR_fam51_osr4_poprice_expop1+	83	1.33E-09	<i>O. sativa</i> seed
1	4776700	LTR_fam51_osr4_poprice_expop1+	40	5.01E-04	<i>O. sativa</i> seed
1	4781700	LTR_fam51_osr4_poprice_expop1+	60	1.35E-06	<i>O. sativa</i> seed
1	4781800	LTR_fam51_osr4_poprice_expop1+	54	8.38E-06	<i>O. sativa</i> seed
1	4781900	LTR_fam51_osr4_poprice_expop1+	41	3.78E-04	<i>O. sativa</i> seed
1	28292600	DTM_MULE_japo_Os0086	39	6.64E-04	<i>O. sativa</i> seed
1	28292700	DTM_MULE_japo_Os0086	260	0.00E+00	<i>O. sativa</i> seed
1	28292800	DTM_MULE_japo_Os0086	266	0.00E+00	<i>O. sativa</i> seed
1	28292900	DTM_MULE_japo_Os0086	48	5.00E-05	<i>O. sativa</i> seed
2	11897000	LTR_fam51_osr4_poprice_expop1+	46	8.97E-05	<i>O. sativa</i> seed
2	11897100	LTR_fam51_osr4_poprice_expop1+	93	5.22E-11	<i>O. sativa</i> seed
2	11897200	LTR_fam51_osr4_poprice_expop1+	116	2.66E-14	<i>O. sativa</i> seed
2	11897300	LTR_fam51_osr4_poprice_expop1+	103	1.97E-12	<i>O. sativa</i> seed
2	11897400	LTR_fam51_osr4_poprice_expop1+	81	2.52E-09	<i>O. sativa</i> seed
2	11897500	LTR_fam51_osr4_poprice_expop1+	48	5.00E-05	<i>O. sativa</i> seed
2	11902200	LTR_fam51_osr4_poprice_expop1+	46	8.97E-05	<i>O. sativa</i> seed
2	11902300	LTR_fam51_osr4_poprice_expop1+	126	1.33E-15	<i>O. sativa</i> seed
2	11902400	LTR_fam51_osr4_poprice_expop1+	156	0.00E+00	<i>O. sativa</i> seed
2	11902500	LTR_fam51_osr4_poprice_expop1+	162	0.00E+00	<i>O. sativa</i> seed
2	11902600	LTR_fam51_osr4_poprice_expop1+	100	5.29E-12	<i>O. sativa</i> seed
2	11902700	LTR_fam51_osr4_poprice_expop1+	50	2.77E-05	<i>O. sativa</i> seed
2	33615400	DTM_clust113	72	4.41E-08	<i>O. sativa</i> seed
2	33615500	DTM_clust113	48	5.00E-05	<i>O. sativa</i> seed
2	33658200	rn_118-68_expop1+	41	3.78E-04	<i>O. sativa</i> seed
2	34010100	LTR_fam51_osr4_poprice_expop1+	70	6.04E-08	<i>O. sativa</i> seed
2	34010200	LTR_fam51_osr4_poprice_expop1+	122	5.00E-15	<i>O. sativa</i> seed
2	34010300	LTR_fam51_osr4_poprice_expop1+	159	0.00E+00	<i>O. sativa</i> seed
2	34010400	LTR_fam51_osr4_poprice_expop1+	139	0.00E+00	<i>O. sativa</i> seed
2	34010500	LTR_fam51_osr4_poprice_expop1+	83	1.33E-09	<i>O. sativa</i> seed
2	34010600	LTR_fam51_osr4_poprice_expop1+	48	5.00E-05	<i>O. sativa</i> seed
2	34015300	LTR_fam51_osr4_poprice_expop1+	60	1.35E-06	<i>O. sativa</i> seed

CHR	BP	TE family	Reads per million reads	P-value	Mobilome library
2	34015400	LTR_fam51_osr4_poprice_expop1+	129	4.44E-16	<i>O. sativa</i> seed
2	34015500	LTR_fam51_osr4_poprice_expop1+	157	0.00E+00	<i>O. sativa</i> seed
2	34015600	LTR_fam51_osr4_poprice_expop1+	160	0.00E+00	<i>O. sativa</i> seed
2	34015700	LTR_fam51_osr4_poprice_expop1+	91	1.00E-10	<i>O. sativa</i> seed
3	1910900	LTR_fam51_osr4_poprice_expop1+	43	2.14E-04	<i>O. sativa</i> seed
3	1911000	LTR_fam51_osr4_poprice_expop1+	40	5.01E-04	<i>O. sativa</i> seed
3	1916200	LTR_fam51_osr4_poprice_expop1+	46	8.97E-05	<i>O. sativa</i> seed
3	1916300	LTR_fam51_osr4_poprice_expop1+	57	3.38E-06	<i>O. sativa</i> seed
3	1916400	LTR_fam51_osr4_poprice_expop1+	40	5.01E-04	<i>O. sativa</i> seed
4	11804800	LTR_fam51_osr4_poprice_expop1+	42	2.84E-04	<i>O. sativa</i> seed
4	21858400	LTR_fam51_osr4_poprice_expop1+	59	1.84E-06	<i>O. sativa</i> seed
4	21858500	LTR_fam51_osr4_poprice_expop1+	64	3.93E-07	<i>O. sativa</i> seed
4	21863500	LTR_fam51_osr4_poprice_expop1+	44	1.60E-04	<i>O. sativa</i> seed
4	21863600	LTR_fam51_osr4_poprice_expop1+	51	2.06E-05	<i>O. sativa</i> seed
4	21863700	LTR_fam51_osr4_poprice_expop1+	59	1.84E-06	<i>O. sativa</i> seed
4	21863800	LTR_fam51_osr4_poprice_expop1+	54	8.38E-06	<i>O. sativa</i> seed
4	31206000	LTR_fam51_osr4_poprice_expop1+	63	5.36E-07	<i>O. sativa</i> seed
4	31206100	LTR_fam51_osr4_poprice_expop1+	112	1.01E-13	<i>O. sativa</i> seed
4	31206200	LTR_fam51_osr4_poprice_expop1+	112	1.01E-13	<i>O. sativa</i> seed
4	31206300	LTR_fam51_osr4_poprice_expop1+	83	1.33E-09	<i>O. sativa</i> seed
4	31210900	LTR_fam51_osr4_poprice_expop1+	44	1.60E-04	<i>O. sativa</i> seed
4	31211000	LTR_fam51_osr4_poprice_expop1+	48	5.00E-05	<i>O. sativa</i> seed
4	31211100	LTR_fam51_osr4_poprice_expop1+	48	5.00E-05	<i>O. sativa</i> seed
4	31211200	LTR_fam51_osr4_poprice_expop1+	66	2.11E-07	<i>O. sativa</i> seed
4	31211300	LTR_fam51_osr4_poprice_expop1+	107	5.28E-13	<i>O. sativa</i> seed
4	31211400	LTR_fam51_osr4_poprice_expop1+	151	0.00E+00	<i>O. sativa</i> seed
4	31211500	LTR_fam51_osr4_poprice_expop1+	89	1.92E-10	<i>O. sativa</i> seed
4	31211600	LTR_fam51_osr4_poprice_expop1+	54	8.38E-06	<i>O. sativa</i> seed
5	27400400	DTM_MULE_japo_Os0314.DTM_clust416	39	6.64E-04	<i>O. sativa</i> seed
6	28571500	GAIJIN_DNA_transposon_Oryza_sativa	41	3.78E-04	<i>O. sativa</i> seed
6	28571600	DTX-incomp-chim_Osati-B-R2026-Map10.GAIJIN_DNA_transposon_Oryza_sativa.RSU_clust154	68	1.13E-07	<i>O. sativa</i> seed
7	26694800	LTR_fam158_tos17_expop1+	54	8.38E-06	<i>O. sativa</i> seed
7	26694900	LTR_fam158_tos17_expop1+	53	1.13E-05	<i>O. sativa</i> seed
7	26695000	LTR_fam158_tos17_expop1+	51	2.06E-05	<i>O. sativa</i> seed
7	26695100	LTR_fam158_tos17_expop1+	44	1.60E-04	<i>O. sativa</i> seed
7	26695200	LTR_fam158_tos17_expop1+	53	1.13E-05	<i>O. sativa</i> seed
7	26695300	LTR_fam158_tos17_expop1+	48	5.00E-05	<i>O. sativa</i> seed
7	26695400	LTR_fam158_tos17_expop1+	46	8.97E-05	<i>O. sativa</i> seed
7	26698700	LTR_fam158_tos17_expop1+	43	2.14E-04	<i>O. sativa</i> seed
8	9051800	DHX-incomp_Osati-B-R11497-Map7_reversed.LTR_fam51_osr4_poprice_expop1+	42	2.84E-04	<i>O. sativa</i> seed
8	9051900	LTR_fam51_osr4_poprice_expop1+	92	7.23E-11	<i>O. sativa</i> seed
8	9052000	LTR_fam51_osr4_poprice_expop1+	113	7.22E-14	<i>O. sativa</i> seed
8	9052100	LTR_fam51_osr4_poprice_expop1+	86	5.06E-10	<i>O. sativa</i> seed
8	9052200	LTR_fam51_osr4_poprice_expop1+	54	8.38E-06	<i>O. sativa</i> seed
8	9057000	LTR_fam51_osr4_poprice_expop1+	38	8.77E-04	<i>O. sativa</i> seed
8	9057100	LTR_fam51_osr4_poprice_expop1+	52	1.53E-05	<i>O. sativa</i> seed
8	9057200	LTR_fam51_osr4_poprice_expop1+	61	9.94E-07	<i>O. sativa</i> seed
8	9057300	LTR_fam51_osr4_poprice_expop1+	67	1.55E-07	<i>O. sativa</i> seed
8	9057400	LTR_fam51_osr4_poprice_expop1+	53	1.13E-05	<i>O. sativa</i> seed

CHR	BP	TE family	Reads per million reads	P-value	Mobilome library
8	15220800	LTR_fam86_exp0p1+	42	2.84E-04	<i>O.sativa</i> seed
8	25668500	LTR_fam51_osr4_poprice_exp0p1+	48	5.00E-05	<i>O.sativa</i> seed
8	25668600	LTR_fam51_osr4_poprice_exp0p1+	38	8.77E-04	<i>O.sativa</i> seed
8	25668700	LTR_fam51_osr4_poprice_exp0p1+	41	3.78E-04	<i>O.sativa</i> seed
8	25668800	LTR_fam51_osr4_poprice_exp0p1+	40	5.01E-04	<i>O.sativa</i> seed
9	1229300	LTR_fam51_osr4_poprice_exp0p1+	45	1.20E-04	<i>O.sativa</i> seed
9	8362200	LTR_fam29_rire10_exp0p1+	46	8.97E-05	<i>O.sativa</i> seed
9	8362300	LTR_fam29_rire10_exp0p1+	63	5.36E-07	<i>O.sativa</i> seed
9	8362400	LTR_fam29_rire10_exp0p1+	51	2.06E-05	<i>O.sativa</i> seed
9	8572300	LTR_fam51_osr4_poprice_exp0p1+	46	8.97E-05	<i>O.sativa</i> seed
9	8577500	LTR_fam51_osr4_poprice_exp0p1+	48	5.00E-05	<i>O.sativa</i> seed
9	18114700	LTR_fam51_osr4_poprice_exp0p1+	38	8.77E-04	<i>O.sativa</i> seed
9	18114800	LTR_fam51_osr4_poprice_exp0p1+	46	8.97E-05	<i>O.sativa</i> seed
9	18114900	LTR_fam51_osr4_poprice_exp0p1+	44	1.60E-04	<i>O.sativa</i> seed
9	18119400	LTR_fam51_osr4_poprice_exp0p1+	39	6.64E-04	<i>O.sativa</i> seed
9	18119500	LTR_fam51_osr4_poprice_exp0p1+	55	6.20E-06	<i>O.sativa</i> seed
9	18119600	LTR_fam51_osr4_poprice_exp0p1+	94	3.77E-11	<i>O.sativa</i> seed
9	18119700	LTR_fam51_osr4_poprice_exp0p1+	110	1.95E-13	<i>O.sativa</i> seed
9	18119800	LTR_fam51_osr4_poprice_exp0p1+	140	0.00E+00	<i>O.sativa</i> seed
9	18119900	LTR_fam51_osr4_poprice_exp0p1+	133	1.11E-16	<i>O.sativa</i> seed
9	18120000	LTR_fam51_osr4_poprice_exp0p1+	106	7.34E-13	<i>O.sativa</i> seed
9	18120100	LTR_fam51_osr4_poprice_exp0p1+	80	3.47E-09	<i>O.sativa</i> seed
9	18120200	LTR_fam51_osr4_poprice_exp0p1+	48	5.00E-05	<i>O.sativa</i> seed
10	2994100	DTM_MULE_japo_Os3337	42	2.84E-04	<i>O.sativa</i> seed
10	2994200	DTM_MULE_japo_Os3337	59	1.84E-06	<i>O.sativa</i> seed
10	2994300	DTM_MULE_japo_Os3337	65	2.88E-07	<i>O.sativa</i> seed
10	22300600	LTR_fam51_osr4_poprice_exp0p1+	70	6.04E-08	<i>O.sativa</i> seed
10	22300700	LTR_fam51_osr4_poprice_exp0p1+	76	1.24E-08	<i>O.sativa</i> seed
10	22300800	LTR_fam51_osr4_poprice_exp0p1+	67	1.55E-07	<i>O.sativa</i> seed
10	22300900	LTR_fam51_osr4_poprice_exp0p1+	46	8.97E-05	<i>O.sativa</i> seed
10	22305900	LTR_fam51_osr4_poprice_exp0p1+	55	6.20E-06	<i>O.sativa</i> seed
10	22306000	LTR_fam51_osr4_poprice_exp0p1+	54	8.38E-06	<i>O.sativa</i> seed
11	8983500	DTX-incomp-chim_Osati-B-G3264-Map20.LTR_fam73_exp0p1+	39	6.64E-04	<i>O.sativa</i> seed
12	673100	LTR_fam4_dasheng_osr25_exp0p1+	40	5.01E-04	<i>O.sativa</i> seed

Programmes bioinformatiques

Benchling	https://benchling.com/crispr
BLAST (NCBI programme)	https://www.ncbi.nlm.nih.gov/BLAST/
DMRcaller	https://www.bioconductor.org/packages/release/bioc/vignettes/DMRcaller/inst/doc/DMRcaller.pdf
Domaines conservés	https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi
Emmax	http://genetics.cs.ucla.edu/emmax/
FASTQC	http://www.bioinformatics.babraham.ac.uk/projects/fastqc
FigTree software	http://tree.bio.ed.ac.uk/software/figtree/
IGV	https://www.broadinstitute.org/igv/home
IRGSP1	http://rgp.dna.affrc.go.jp/E/IRGSP/Build5.html
MAFFT	http://mafft.cbrc.jp/alignment/server/
OligoCalc	http://www.basic.northwestern.edu/biotools/oligocalc.html
Primers3	http://www.primer3.ut.ee
Plink	https://www.cog-genomics.org/plink2
RepeatMasker	http://www.repeatmasker.org
R software	http://www.r-project.org/
SEAVIEW	http://doua.prabi.fr/software/seaview
Trimmomatic	http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/TrimmomaticManual_V0.32.pdf