Etude de l'activité transpositionnelle chez le genre Oryza à l'ère des nouvelles technologies de séquençage

Marie-Christine Carpentier

THÈSE

Pour obtenir le grade de Docteur

Délivré par

UNIVERSITÉ DE PERPIGNAN VIA DOMITIA

Préparée au sein de l'école doctorale

« Énergie Environnement » ED 305

et de l'unité de recherche LGDP-UMR 5096

CNRS-UPVD

Spécialité : Biologie

Présentée par Marie-Christine Carpentier

Étude de l'activité transpositionnelle chez le genre *Oryza* à l'ère des nouvelles technologies de séquençage

Soutenue le **10 septembre 2021** devant le jury composé de

Hadi Quesneville, Directeur de Recherche — Rapporteur
Unité de Recherche Génomique Info, INRAE, Versailles

Mathias Lorieux, Directeur de Recherche — Rapporteur
Institut de Recherche pour le Développement, Montpellier

Clémentine Vitte, Chargée de Recherche — Examinatrice
UMR Génétique Quantitative et Evolution, Gif-sur-Yvette

Leandro Quadrana, Chargé de Recherche — Examinateur
Institut de Biologie de l'Ecole Normale Supérieure, Paris

Cécile Bousquet-Antonelli, Directrice de recherche — Examinatrice
Laboratoire Génome et Développement des Plantes, Perpignan

Olivier Panaud, Professeur des Universités — Directeur
Université de Perpignan Via Domitia, Perpignan

Résumé

Les éléments transposables (ET) sont des composants ubiquitaires des génomes eucaryotes. Ce sont des séquences d'ADN mobile, qui ont la capacité de se multiplier et de se déplacer au sein des chromosomes de la plupart des organismes vivants. Ils appartiennent à deux classes : la classe I, des rétrotransposons et la classe II, des transposons. Les rétrotransposons sont prédominants chez les plantes.

Avec l'avancée des nouvelles technologies de séquençage ces dernières décennies, il est maintenant possible d'étudier l'impact structural et fonctionnel de ces éléments au sein de populations naturelles. Néanmoins par leur nature répétée, la détection des polymorphismes associés à l'activité de ces éléments est un défi conceptuel et technique.

Au cours de ma thèse, j'ai développé un pipeline (TRACKPOSON) permettant la détection des insertions polymorphiques d'éléments transposables (en particulier les rétrotransposons) au sein de grand jeux de données que j'ai appliqué aux 3,000 génomes de riz disponibles.

Cette étude a permis de mettre en évidence un paysage transpositionnel très dynamique au sein de l'espèce. A partir de ces résultats, afin d'identifier d'éventuels facteurs génétiques liés à l'activité transpositionnelle, une analyse d'association (GWAS) a été conduite. Celle-ci a permis de mettre en évidence la présence d'une copie « maître » de l'élément lui-même, copie nécessaire à l'activation de la transposition de la famille. Les données de polymorphismes d'insertion des ETs pour l'ensemble des 3000 génomes nous ont également permis d'élucider l'origine de la domestication du riz. Nous avons en effet pu mettre en évidence au moins trois origines génétiques distinctes de l'espèce cultivée.

Dans un second temps, nous avons étudié l'impact des insertions de ces éléments transposables sur certains caractères agronomiques (TE-GWAS). Nous avons mis en évidence qu'une insertion du rétrotransposon *rn215-125* avait un impact significatif sur la largeur des grains des variétés de riz *Indica*. Grâce à l'utilisation de la technologie de séquençage longues lectures Nanopore, nous avons pu caractériser la région dans laquelle se trouve l'insertion et proposons que l'allèle conférant un grain plus large proviendrait plutot d'un évènement d'introgression à partir d'un donneur qui reste à identifier, bien que nos premiers résultats suggèrent qu'il pourrait s'agir de l'espèce sauvage apparentée *O. rufipogon*.

Les analyses conduites au cours de ma thèse permettent de mettre en évidence la forte contribution des éléments transposables sur la dynamique des génomes au sein de la population de 3000 variétés de riz. De plus, l'évolution constante des technologies de séquençage et le développement des outils bioinformatique associés, ouvrent maintenant de nouvelles perspectives pour l'analyse des variations structurales et en particulier l'impact des éléments transposables sur l'évolution des génomes au sein des populations.

Mots-clés : éléments transposables, *Oryza sativa*, génomique, bio-informatique, grand jeux de données, nouvelles technologie de séquençage, étude génomique d'association (GWAS)

Abstract

Transposable elements (TEs) are ubiquitous components of eukaryotic genomes. They are mobile DNA sequences, which have the ability to multiply and move within the chromosomes of most living organisms. They belong to two classes : class I, retrotransposons and class II, transposons. Retrotransposons are predominant in plants.

With the advancement of new sequencing technologies in the last decades, it is now possible to study the structural and functional impact of these elements in natural populations. Nevertheless, due to their repeated nature, the detection of polymorphisms associated with the activity of these elements is conceptually and technically challenging.

During my PhD, I developed a pipeline (TRACKPOSON) allowing the detection of polymorphic insertions of transposable elements (in particular retrotransposons) in large datasets that I applied to the 3,000 available rice genomes.

This study revealed a very dynamic transpositional landscape within the species. Based on these results, an association analysis (GWAS) was conducted to identify possible genetic factors related to transpositional activity. This analysis revealed the presence of a "master" copy of the element itself, a copy necessary for the activation of transposition in the family.

The TE insertion polymorphisms for all 3000 genomes also enabled us to elucidate the origin of rice domestication. We were able to identify at least three distinct genetic origins of the cultivated species.

In a second step, we studied the impact of insertions of these transposable elements on certain agronomic traits (TE-GWAS). We found that an insertion of the retrotransposon *rn215-125* had a significant impact on grain width in *Indica* rice varieties. Using Nanopore long-read sequencing technology, we were able to characterize the region into which the insertion is located and propose that the allele conferring a wider grain is more likely to result from an introgression event from a donor yet to be identified, although our initial results suggest that it may be the wild relative *O. rufipogon*.

The analyses conducted during my thesis show the strong contribution of transposable elements to genome dynamics in the population of 3000 rice varieties. Moreover, the constant evolution of sequencing technologies and the development of associated bioinformatics tools now open new perspectives for the analysis of structural variations and in particular the impact of transposable elements on the evolution of genomes within populations.

Keywords : transposable elements, *Oryza sativa*, genomics, bioinformatics, big data, next generation sequencing technologies, genome-wide association studies (GWAS)

Remerciements

Tout d'abord, je voudrais remercier tous les membres de mon jury qui ont accepté d'évaluer mes travaux de thèse. Merci **Hadi Quesneville** et **Mathias Lorieux** d'avoir rapporté mes travaux. Merci **Cécile Bousquet-Antonelli**, **Clémentine Vitte** et **Leandro Quadrana** d'avoir accepté d'être dans mon jury.

Merci également à tous les membres de mon CSI, **Cristina Viera, Jean-Marc Deragon, Josep Casacuberta** pour leurs remarques pertinentes et toutes les discussions au cours de mes 5 comités de suivi de thèse.

Merci à **Yue-Ie Hsing** et son équipe pour notre collaboration autour du projet des 3000 génomes et d'avoir fait tourner mon pipeline sur leur cluster de calcul.

N'ayant pas une grande fibre d'écrivaine, j'avais au début opté pour un "merci à tous, vous vous reconnaîtrez"…mais après réflexion, je me suis ravisée. D'avance, je m'excuse auprès de toutes les personnes que j'ai oublié de citer…mais sachez que ce n'est pas pour ça que je ne pense pas à vous…c'est juste que je n'ai pas une grande mémoire.

Presque 9 ans au LGDP dont 6 ans en thèse, il s'en est passé des choses : 10 doctorants, 3 présidents de l'UPVD, 2 directeurs de laboratoire, une pandémie…

Un grand merci bien sûr à tous les membres du LGDP (**#lesLGDPiens**) pour votre soutien tout au long de ces années.

Jean-Marc, merci de m'avoir soutenue lorsque j'ai eu la folle idée de commencer un doctorat en poste. Merci pour nos discussions lors des repas à midi à la cafet', c'était le bon vieux temps

Olivier, je me rappelle lorsque que je suis arrivée dans l'équipe, tu m'as dit "le futur d'une ingénieure d'étude c'est de devenir ingénieure de recherche". A l'époque, je l'avais pris à la rigolade mais au final, tu avais peut-être raison. Merci d'avoir été mon directeur de thèse pendant ces 6 années : d'avoir eu confiance en moi quand moi-même je doutais, merci de m'avoir appris tellement choses sur le riz et les éléments transposables. Merci pour toutes nos discussions scientifiques et non-scientifiques autour de la permaculture, des abeilles et bien d'autres. Merci de m'avoir permis de développer mon autonomie dans mon projet de thèse et en même temps d'avoir été disponible quand il le fallait. Merci pour ta bienveillance pendant la correction de mon manuscrit qui n'a pas été de tout repos. Après de nombreux aller-retour : "we did it" !!!

Merci à tous les membres de mon équipe AGE, maintenant devenue MANGO : tel de petits Pokémon, nous avons petit à petit évolué.

Marie, merci pour tes conseils toujours précieux et pertinents, ta bonne humeur et ta bienveillance. Merci pour toutes les corrections de mon manuscrit et ton avis éclairé sur les figures.

Eric, merci pour ta relecture et nos collaborations…maintenant libérée, je vais me remettre à fond au projet OA.

Moaine, Joris, Valy, Emilie et Panpan, merci pour toutes nos discussions en réunions d'équipe ou en dehors toujours enrichissantes.

Christel : merci pour ta gentillesse, ta réactivé, et ton aide indispensable en tant qu'experte manip' Nanopore.

Merci **Dom**, binome du repas de labo, merci pour ta gentillesse et ton aide précieuse pour tous mes bons de commandes ces derniers temps.

Merci au "staff administratif" : **Elisabeth, JR et Myriam**, de m'avoir aidée quand j'étais dans le brouillard administratif avec toutes les démarches pour l'école doctorale, mes missions pour les congrès.

Ce document n'aurait jamais vu le jour (au moins pas sous cette forme), sans toi **Martin**. Merci de m'avoir transmis ta passion pour le LaTeX. Tu m'as presque convaincue à l'utiliser pour mes présentations avec Beamer. J'ai hâte de continuer à être ta petite padawan…

Pendant cette grande aventure, il fallait bien se sorti la tête des publis, projets et lignes de commandes et de s'aérer les idées. Merci aux piliers de la bières : **Christophe, JR, Rémy**, nos traditionnels jeudi ont été de tels moments de rigolade.

Merci **Jean-Jacques**, dixit JJ pour toutes nos discussions autour de la gastronomie et pas que…ça fait du bien de trouver quelqu'un qui adore presqu'autant la nourriture que moi.

Merci **Viviane** pour ton aide dans la culture de nos plants de riz. Mais surtout, merci d'avoir toujours été à l'écoute, de m'avoir rassurée, et aussi bien changer les idées au cours de nos discussions avec nos questions philosophiques, à refaire le monde.

Aux doctorants de maintenant : **Arnaud, Avilien, Clément, Emilie, Edouardo, Jean-Loup, Panpan, Vichet**, merci pour tous ces bons moments passés au labo et en dehors : on sait tous que le doctorat c'est de vrais montages russes…j'avoue c'est desfois bien rude mais je vous assure, ça vaut le coup !

Rémy, pendant ces 6 ans, je ne compte plus nos cafés du matin, nos collaborations, nos marches pour prendre l'air à la boulangerie, le nombre fois où je t'ai dit "je peux arrêter ma thèse maintenant, non ?"…mais surtout pour tous nos moments de rigolade…Merci d'avoir été là à écouter mes lamentations de petite doctorante sans presque rien dire ;) de m'avoir changé les idées en balade à la montagne mais surtout autour d'une bière ou d'un bon verre de vin, quand il le fallait….bref merci d'être mon ami.

J'en ai vu des doctorants arriver et partir du LGDP, tels des chenilles devenues pa-

pillons. Merci donc ”aux anciens” (#lespetits) devenus maintenant amis : **Elodie, Jérémy, Sophie et Ari**. Merci pour votre soutien, même si vous êtes loin. Je me dis qu'au final, vous m'avez sûrement inspiré…

Merci à toute l'équipe du Pentathlon (#PMPC) **Manu, Sophie, Virginie** pour votre super coaching et mes accolytes sportifs **Edouard, Emilie, Jean-Sé, Magali, Nathalie et Tommy** : grâce aux entraînements j'ai pu me vider la tête, ce qui a été indispensable pour mon bien-être mental, surtout cette année.

En dehors du laboratoire, il y a aussi une vie…et oui, on l'oublie parfois..surtout ces derniers temps.

Noémie, Charlotte, Coline tant d'années d'amitié, même si on est un peu loin physiquement les unes des autres, merci d'être vous et d'avoir toujours été à mes cotés dans les bons comme dans les moments plus difficiles. Je suis tellement reconnaissante de vous avoir dans ma vie.

Entre Paris et Perpignan, il y a eu Lyon. **Claire, Fabi, Gilbert**, les anciens *salseros/salsera*, même si on ne se voit pas souvent depuis mon exil dans le sud, chaque fois c'est comme si de rien était : toujours autant de rigolade. Promis, je remonterais plus souvent…

En terres catalanes, j'ai été également très bien entourée. **Naoual, Kristel, Yann** : merci d'avoir été là tout au long de cette aventure et à tous les AMApiens (on dirait vraiment une secte ?) **Anthony, Camille, Elise, Gérald, Micka, Valentine** de m'avoir changé les idées au cours de nos ballades, apéro, resto, mardi de l'AMAP et bien d'autres…pendant ces derniers mois un peu rude.

Et on garde les meilleurs pour la fin, comme on dit. Il n'y a pas de mot pour exprimer toute la gratitude que j'ai envers ma famille.

Florence, merci soeurette, d'avoir été ma *accountability partner* quand j'en ai eu besoin, de m'avoir aidée à relativiser, de m'avoir redonné espoir et confiance, pour toutes nos discussions qui ont duré des heures au téléphone..ne change pas !

Philippe, mon petit fréro, merci à toi de m'avoir écoutée et rassurée, de m'avoir changé les idées en parlant du Japon, des restos à tester sur Paris et de bouffe pour changer (est-ce une obsession ???).

Merci **Papa, Maman**, si j'en suis arrivée là c'est grâce à vous. Merci de m'avoir toujours soutenue et encouragée pendant toutes ces années, d'avoir essayé de comprendre le sujet de ma thèse et ce que je faisais dans mon boulot..même si vous avez au final abdiqué. Vous avez toujours été présent et bienveillant, je ne vous en remercierai jamais assez.

Je suis bien consciente de l'énorme chance de vous avoir tous à mes côtés…un seul mot MERCI pour tout.

Table des matières

Table des figures

Acronymes

ADN Acide DéxoxyriboNucléique.

AP Protéase Aspartique.

ARN Acide RiboNucléique.

ARNt ARN de transfert.

BAC Bacterial Artificial Chromosome.

bp paire de bases.

cDNA ADN complémentaire.

CNV Copy Number Variation.

CPU Unité Central de Traitement.

CRAG Center for Research in Agricultural Genomics.

CRISPR Clustered Regularly Interspaced Short Palindromic Repeats.

ddNTP Didésoxyribonucléotides.

DOC Depth of coverage.

ERV Rétrovirus endogènes.

ET Eléments transposables.

FLC Flowering locus C.

FST Indice de fixation.

GAG Group specific Antigen.

Gb Gigabase - 10^9 bp.

Go Gigaoctets - 10^9 octets.

GPU Unité de Traitement Graphique.

GST Glucosyl transferase.

GWAS Etude de Génomique d'Association.

HSP High-scoring Segment Pair.

INDEL Insertions et Délétions.

INT Integrase.

IRRI International Rice Research Institute.

IS Insertion Sequences.

Kb Kilobase - 10^3 bp.

LINE Long Interspersed Nuclear Element.

LTR Long Terminal Repeat.

Mb Mégabases - 10^6 bp.

MITE Miniature Inverted-repeat Transposable Element.

My Million d'années - 10^6 année.

NGS Next Generation Sequencing.

ORF Open Reading Frame.

PAF Pairwise mApping Format.

PBS Primer Binding Site.

PCR Polymerase Chaine Reaction.

POL Polyproteine.

PPT Poly-purine.

RH RNAseH.

RNASeq Séquençage des ARN.

RT Reverse Transcriptase.

SAM Sequence Alignment Map.

SASAR Super-ASsembly from Assembly Reconciliation.

SINE Short Interspersed Nuclear Element.

siRNA small interfering RNA.

smRNA small RNA.

SNP Single Nucleotide Polymorphisms.

SV Variation Structural.

Tb Terabytes - 10^{12} bytes.

TFBS Transcription Factor Binding Site.

TIP Transposable element Insertion Polymorphisms.

TIR Terminal Inverted Repeats.

UTR Untranslated Transcribed Region.

VLP Virus-like Particule.

WGD Whole Genome Duplication.

Préambule

Depuis décembre 2012, j'ai intégré le Laboratoire Génome et Développement des Plantes en tant qu'ingénieure d'étude CNRS en traitement de données biologiques au sein de l'équipe "Analyse des Génomes et Evolution" (AGE) dirigée par Olivier Panaud. En parallèle, je suis également devenue responsable de la plateforme bio-informatique du laboratoire, m'occupant ainsi de toute l'analyse bio-informatique des projets des différentes équipes du laboratoire.

La thématique de l'équipe de recherche, à l'époque, était l'étude de la dynamique des génomes des plantes en se focalisant sur les élements transposables, et surtout les rétrotransposons.

Au cours de l'année 2014, lorsque le jeu de données de 3,000 génomes de riz cultivé ont été publiés, ce fut l'opportunité de mettre mes compétences en bio-informatique pour analyser ce grand jeu de données et d'étudier la dynamique des éléments transposables au sein de cette population. Ce projet de recherche est ainsi devenu mon projet de thèse.

Grâce au soutien du directeur du laboratoire, Jean-Marc Deragon et de mon directeur d'équipe, Oliver Panaud, j'ai donc décidé de commencer un doctorat en septembre 2015, tout en continuant mon travail au sein de la plateforme bio-informatique. Ainsi j'ai effectué ma thèse à mi-temps.

Au cours de ces 6 années, en parallèle de mon projet de thèse, j'ai participé à de nombreux projets qui m'ont amené à analyser des données transcriptomiques (RNASeq) de génotypes mutants et sauvages de la plante modèle *Arabidopsis thaliana*. Ainsi, avec la plateforme bio-informatique du laboratoire pour répondre aux questions diverses des chercheurs, j'ai dû m'adapter à un nouveau modèle et également développer de nouveaux pipelines et stratégies d'analyses liés aux types de données RNASeq, données qui sont très différentes de celles analysées au cours de ma thèse.

Ces collaborations m'ont permis de participer à de nombreuses publications, 8 au total (Merret *et al.* 2015 ; Pontvianne *et al.* 2016 ; Merret *et al.* 2017 ; Montacié *et al.* 2017 ; Pontier *et al.* 2019 ; Billey *et al.* 2021) dont 3 principales en 1er auteur au cours desquelles j'ai effectué le développement de méthodes majeures dans leur domaine (Carpentier *et al.* 2018, 2020, 2021).

Introduction

CHAPITRE 1

Introduction générale

Selon la théorie Darwinienne, la variabilité génétique au sein des populations est la base de l'adaptation et donc de l'évolution des espèces. L'ADN étant le support de l'information génétique de tous les organismes, les scientifiques ont montré un grand intérêt à la comparaison des espèces au niveau de ce compartiment cellulaire, proposant ainsi de définir les bases génomiques des processus macro-évolutifs. Dans une société humano-centré, alors que l' humain était considéré comme étant le plus complexe des êtres vivants, il a longtemps été pensé que cette complexité biologique devait être corrélée avec la taille du génome (nommé la valeur C). Par ce principe, plus un organisme était (semblait) évolué (complexe), plus il était attendu que sa valeur C soit importante, du fait de la présence de nombreux gènes. Les premières données génomiques ont rapidement infirmé cette hypothèse. Par exemple il a été observé que le génome humain (3,4 Gb) a une taille similaire à celui du maïs (2,1 Gb). Thomas (1971) a formulé cette absence de corrélation entre complexité biologique et taille de génome par le paradoxe de la valeur C.

Prenant en compte ces informations, le nombre de gènes (la valeur G) semblait une réponse à l'explication de la complexité des organismes, mais cette idée a été également rapidement invalidée. Le génome humain comporte plus de 20,000 gènes codant des protéines alors qu'une plante telle que le riz en contient environ 24,000 : c'est le paradoxe de la valeur G.

Tant de paradoxes et une question restant donc à élucider : à quel phénomène était lié la différence de taille de génomes entre les espèces ? Grâce aux données génomiques accumulées au cours des trentes dernières années, il est désormais clair que les deux principales réponses à cette question sont la polyploïdisation et/ou la présence de séquences non géniques répétées.

La polyploïdisation correspond à la multiplication des chromosomes au sein d'une seule cellule par réunion de génomes de la même espèce ou d'espèces différentes (duplication de génome, WGD). Ainsi ce mécanisme, conduit à un doublement de la taille du génome de l'espèce considérée. Cette voie évolutive est assez fréquente chez les plantes, en particulier lors de périodes instables (Murat *et al.* 2015 ; Pont *et al.* 2019). De plus, l'analyse de 106 événements de duplications de génomes au sein des angiospermes a montré que ces changements importants de taille de génomes (amplification puis réduction) étaient étroitement liés à un plus fort taux de diversification. Le taux de spéciation est en effet plus élevé pour les espèces comportant de nombreux WGD (Landis *et al.* 2018).

Les premières analyses de génomes séquencés d'eucaryotes supérieurs, qu'ils soient animaux ou végétaux ont rapidement montré que les gènes ne sont pas les seuls constituants du génome. Ainsi, une autre explication du paradoxe de la valeur C résiderait en la présence de ces séquences non-géniques dont les éléments transposables sont la catégorie majoritaire (Gregory, 2005).

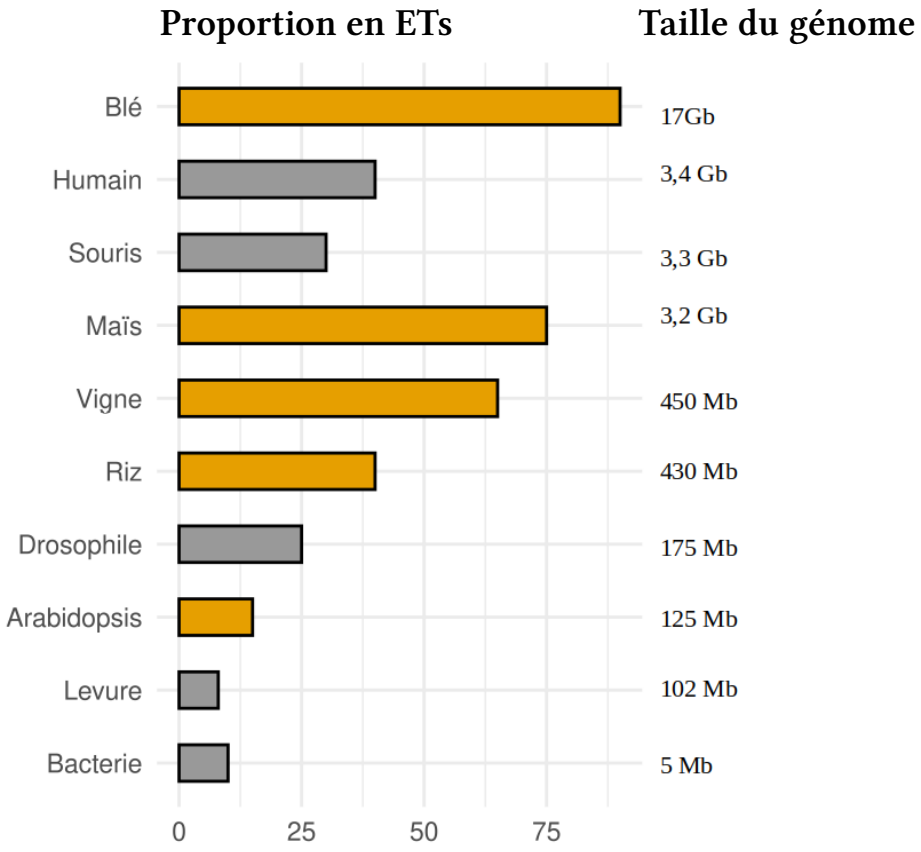


Figure 1.1 – **Distribution de la proportion des éléments transposables en fonction de la taille des génomes.** En jaune sont mis en évidence les plantes. Au sein des plantes, il y a une corrélation entre la taille des génomes et la proportion des éléments transposables.

Par l'analyse de nombreuses espèces, il a été possible d'estimer la proportion de ces éléments transposables au sein des génomes. Ainsi, on observe qu'il y a une corrélation entre la taille des génomes et la proportion en éléments transposables chez les plantes (Figure 1.1). Chez les autres espèces, et notamment les mammifères, les ET sont plus anciens (ils sont moins vite éliminés) et sont très dégénérés. Chez les plantes, la proportion de ces éléments transposables explique bien la différence de taille de génome entre les espèces. Ces derniers sont très présents au sein des plantes, pouvant aller jusqu'à plus de 90 % de la taille du génome chez le blé.

Ainsi, nous allons dans un premier temps décrire les différents types d'éléments transposables (ET).

CHAPITRE 2

Les éléments transposables

2.1 Découverte

Au début des années 50, Barbara McClintock observe chez le maïs, différentes mosaïques de patrons de couleurs de grains et remarque leur instabilité au niveau de leur descendance. Elle identifie deux nouvelles séquences génomiques responsables de ce changement de couleurs : les premiers éléments transposables qu'elle appelle alors "controlling elements" *Activator* (Ac)/*Dissociator* (Ds) chez le maïs. Elle déduit de ses observations que ces deux éléments sont capables de changer de position au sein du chromosome. Elle émet ainsi l'hypothèse que l'instabilité chromosomique pourrait expliquer ce phénomène de patron de coloration des grains, ce qui fit débat à cet époque. L'ADN était en effet considéré comme une molécule stable, transmise à l'identique (à l'exception de quelques mutations) de façon héréditaire et contenant tout le matériel génétique nécessaire à un organisme. Ainsi, la démonstration par Barbara McClintock que la régulation de l'expression des gènes peut être modifiée par la présence d'une séquence d'ADN mobile remet en cause un dogme établi par les généticiens (McClintock, 1953).

Les éléments transposables (ET) sont des séquences d'ADN mobile qui ont la capacité de se déplacer au sein d'un génome. Ils sont présents chez tous les organismes eucaryotes et procaryotes : ce sont des composants ubiquitaires des génomes. Même si l'existence des ET a été confirmée très tôt chez des espèces modèles comme la drosophile avec l'*élément P* (Britten *et al.* 1968) et la levure avec les rétrotransposons *Ty1* et *Ty3* (Cameron *et al.* 1979), les ET ont longtemps été ignorés,considérés comme n'ayant aucun impact biologique significatif ("ADN poubelle"), car ne codant pas pour des fonctions biologiques propres (Doolittle *et al.* 1980 ; Orgel *et al.* 1980).

Grâce aux avancées technologiques de ces dix dernières années dans le domaine de

la biologie, de plus en plus de séquences de génomes d'espèces diverses sont maintenant disponibles. Ainsi, petit à petit, un changement de paradigme s'est opéré à propos des éléments transposables passant du statut d' ADN "poubelle" à ADN avec un intérêt structural (Hua-Van *et al.* 2011 ; Bennetzen *et al.* 2014).

Il s'est en effet avéré que cet ADN mobile contribue significativement à la plasticité du génome, tant au niveau structural que fonctionnel. De plus, plusieurs études ont montré qu'ils peuvent être à l'origine d'innovations biologiques majeures. L'ET *MER20* a contribué à l'origine d'un nouveau réseau de régulation de gènes dédié à la grossesse placentaire chez les mammifères (Lynch *et al.* 2011). Les rétrovirus endogènes (*ERV*), quant à eux ont façonné l'évolution d'un réseau transcriptionnel liés à l'immunité innée (Chuong *et al.* 2016).

2.2 Classification

Les éléments transposables sont très abondants au sein des génomes eucaryotes et notamment chez les plantes, contrairement aux champignons ou métazoaires. Il existe une grande diversité au sein de ces éléments transposables.

Ces éléments ont été divisés en 2 grandes classes selon leur mode de transposition : les rétrotransposons avec un intermédiaire ARN (classe I) et les transposons à ADN (classe II) (Finnegan, 1989). Les mécanismes de transposition des ET de classe I et II sont communément appelés "copier-coller" et "couper-coller" respectivement.

Le séquençage de nombreux génomes eucaryotes a permis une meilleure compréhension des éléments transposables. Ainsi une classification a été faite en se basant sur la similarité de séquences et sur leur origine (Figure 2.1, Wicker *et al.* 2007). On distingue également au sein de chaque classe les ET autonomes des ET non-autonomes. Les ET autonomes ont toute la machinerie (les protéines) nécessaire à leur transposition. Par contre les éléments non-autonomes doivent utiliser les protéines d'autres éléments transposables de la même famille pour pouvoir se déplacer et s'insérer au sein des génomes (Bourque *et al.* 2018).

2.2.1 Les rétrotransposons

Ces éléments transposables de classe I ont une méthode de transposition *via* un intermédiaire ARN, d'où leur nom rétrotransposon. Ils sont présents au sein de tous les eucaryotes. Parmis les ETs de classe I, on distingue les rétrotransposons à LTR (Long Terminal Repeat) des rétrotransposons non-LTR : les LINE et les SINE (Long and Short Interspersed Nuclear Element) (Figure 2.1). A noter que les rétrotransposons à LTR, éléments transposables qui ont été étudiés au cours de ma thèse, sont moins présents chez les animaux et plus abondants au sein des plantes.

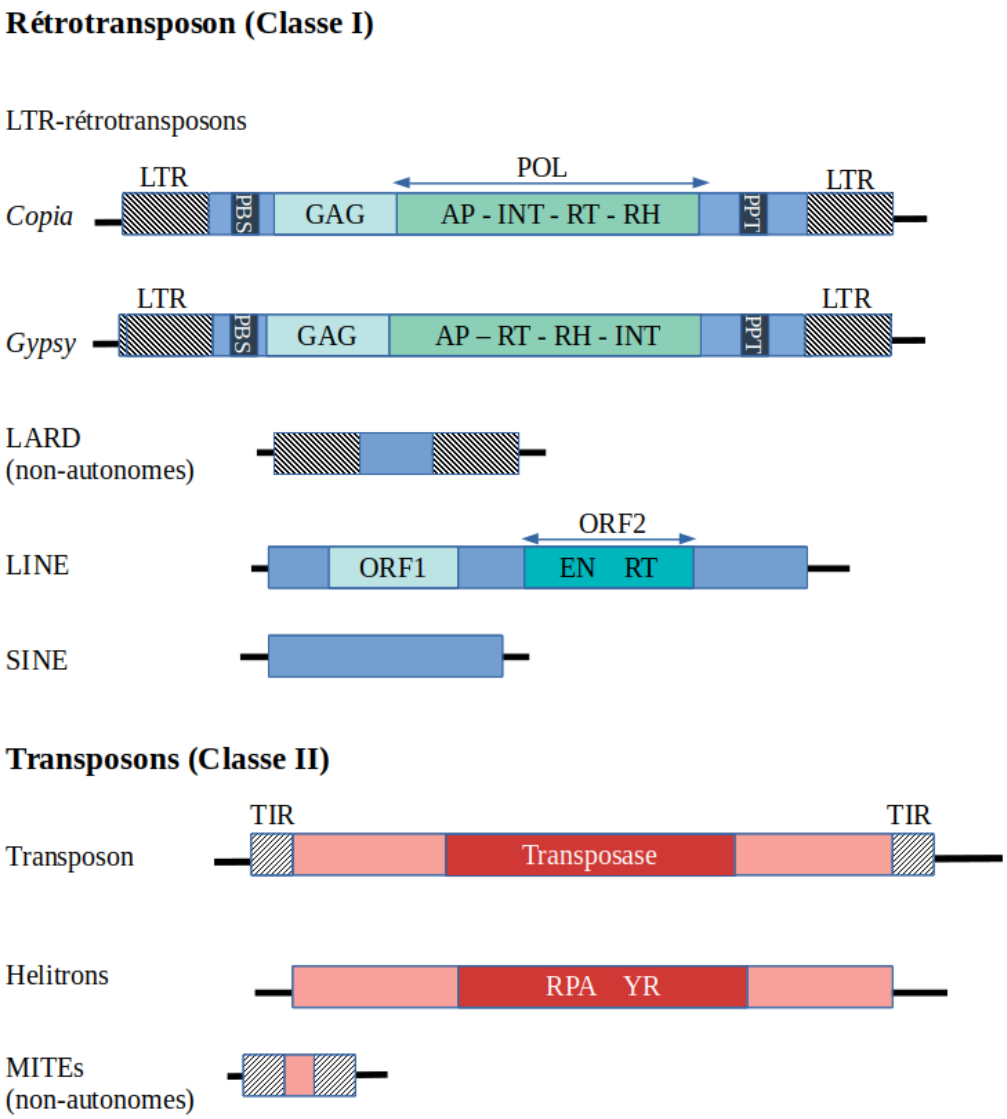


FIGURE 2.1 – **Classification simplifiée des éléments transposables**. Adapté de WICKER *et al.* 2007. Les ET sont divisés en 2 grandes classes selon leur mode de transposition : les rétrotransposons (classe I) et les transposons (classe II). Au sein de chaque classe, des super-familles ont été définies selon la similarité des séquences des ET.
LTR : Long Terminal Repeat, PBS : primer-binding site, POL : polyprotéine, AP : protease aspartique, RT : réverse transcriptase, INT : intégrase, RH : RNAaseH, PPT : poly-purine, EN : enveloppe, TIR : Terminal Inverted Repeat, RPA : Replication protein A, YR : Y2-Tyrosine recombinase

Les rétrotransposons sont constitués d'une séquence codante flanquée ou non par des LTR. La séquence codante de l'élément contient deux cadres de lecture ouverts (*Open Reading Frame, ORF*) : l'un codant la GAG, une protéine structurale qui permet la formation de l'enveloppe (VLP) et l'autre codant une polyprotéine (POL) clivée par la suite en quatre protéines avec des fonctions enzymatiques distinctes, nécessaires au cycle de rétrotransposition de l'ET. Cette polyprotéine apporte toute la machinerie enzymatique nécessaire à la transcription inverse et à l'intégration de l'ET au sein de son génome hôte. Ces protéines, dont les fonctions seront décrites par la suite, sont une protéase (AP), une transcriptase inverse (RT), une RNAseH (RH) et une intégrase (INT).

Il existe 2 grandes familles de rétrotransposons à LTR : *Gypsy* et *Copia*. La différence majeure entre ces familles est la place de la protéine intégrase au sein de la polyprotéine (Figure 2.1).

Le cycle de transposition des rétrotransposons à LTR s'effectue en cinq grandes étapes (Figure 2.2). Celui-ci est semblable à celui des rétrovirus.

1. La transcription de l'élément s'effectue par l'ARN polymérase du génome hôte, à partir du promoteur (Primer Binding Site, PBS) situé en aval du LTR en 5'(Figure 2.1).
2. L'ARN transcrit migre ensuite vers le cytoplasme pour être traduit en protéines : GAG et POL, nécessaire à la transcription inverse et à l'intégration. Les protéines GAG vont se polymériser pour ainsi former une particule de type virale (la VLP) qui permettra de protéger les transcrits de l'ET. Au niveau de la polyprotéine, la protéase va effectuer le clivage de cette dernière en trois protéines distinctes : la transcriptase inverse, la RNAseH et l'intégrase.
3. L'ARNm de l'élément sert de matrice pour la synthèse du brin complémentaire par la transcriptase inverse (RT). La transcription inverse est amorcé par fixation d'un ARN de transfert (ARNt) au niveau du PBS du rétrotransposon, en aval du LTR en 5'. Le LTR en 3' contient le site de terminaison et de poly-adénylation.
4. L'ARN sera par la suite digéré par la RNASeH pour permettre la synthèse du second brin par la RT.
5. Le double brin d'ADN rétrotranscrit migre vers le noyau pour ensuite s'intégrer (grâce à l'intégrase) au niveau d'une nouvelle région génomique.

Ainsi à la fin du cycle de transposition des rétrotransposons à LTR, il y a une insertion d'une nouvelle copie de l'élément qui est identique en tout point avec la copie « maître» dont il est issu. A noter que les LTR et les régions non-codantes des ET sont les parties qui évoluent le plus rapidement au cours du temps, contrairement à la transcriptase inverse, qui semble sous contrainte sélective plus forte vu son rôle dans le cycle rétrotranspositionnel.

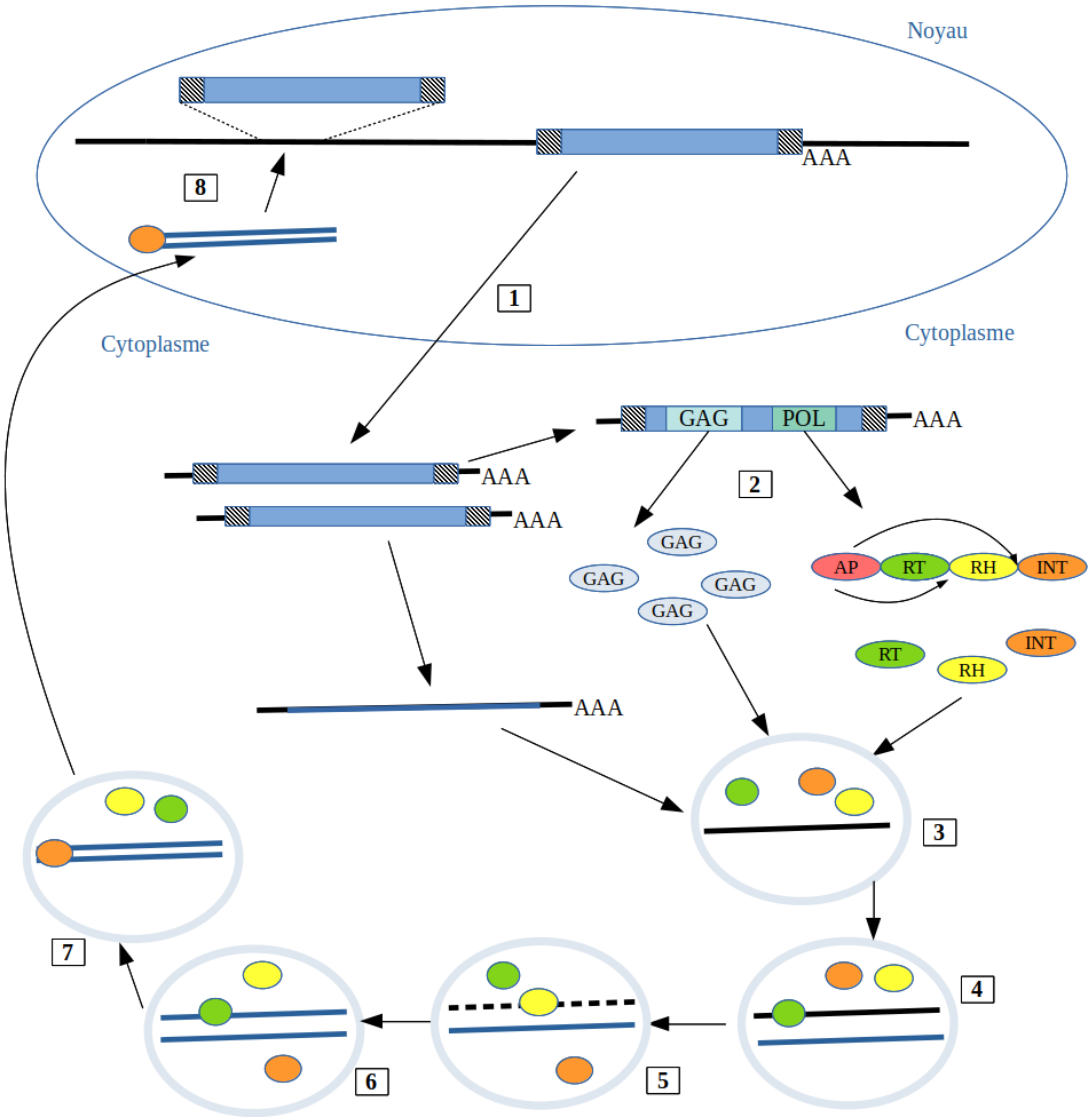


FIGURE 2.2 – **Cycle de transposition d'un rétrotransposon à LTR**. Inspiré de SABOT *et al.* 2006. Le cycle de transposition d'un rétrotransposon à LTR s'effectue en différentes étapes. La transcription du rétrotransposon **(1)** se fait dans le cytoplasme par l'ARN polymérase II. Les transcrits sont utilisés comme matrice pour la traduction **(2)** et comme matrice pour la réverse transcription. Après la traduction, la polyprotéine (POL) est autoclivée par la protéase (AP) en trois protéines : une transcriptase inverse (RT), une RNAse (RH) et une intégrase (INT). Les protéines GAG **(3)** vont s'agglomérer en particule (VLP) qui va permettre la protection des transcrits et des trois protéines. Ainsi, la réverse transcription du transcrit du rétrotransposon peut débuter. Cette transcription via la RT est initiée par liaison d'un ARNt de l'hôte au PBS situé en aval du LTR en 5' **(4)**. La RNAse va ensuite dégrader la matrice ARN **(5)** et le second brin va être synthétisé par la RT **(6)**. Ainsi, on obtient un ADN double brin. Celui-ci va se lier avec l'intégrase **(7)** qui va permettre l'intégration du rétrotransposon au sein du noyau. Cette nouvelle copie pourra par la suite s'intégrer au sein du génome hôte **(8)**.

Les différentes étapes du cycle de transposition des rétrotransposons sont dorénavant bien décrites (Schulman, 2013 ; Sabot *et al.* 2006). Mais les mécanismes de l'intégration de l'élément au niveau de sa nouvelle localisation restent encore à élucider. L'hypothèse d'une forme intermédiaire circulaire à la fin du cycle de rétrotransposition a été proposée (Hirochika *et al.* 1995 ; Lanciano *et al.* 2017).

Les rétrotransposons non LTR

Les LINE (Long Interspersed Nuclear Elements) sont constitués de deux cadres de lecture ouverts, dont un codant une transcriptase inverse (RT) (Figure 2.1). Comme pour les rétrotransposons à LTR, leur intégration est couplée à une transcription inverse qui va synthétiser l'ADN à partir d'une matrice ARN. Les LINE varient en diversité chez les eucaryotes, mais prédominent sur les rétrotransposons LTR chez de nombreux animaux. La famille *L1* compte environ 500,000 copies chez l'humain, soit environ 15% du génome humain (Rodić *et al.* 2013 ; International Human Genome Sequencing Consortium, 2001)

Les SINE (Small Interspersed Nuclear Elements) sont des ET très petits (80-500pb). Le plus connu des SINE est l'élément *Alu* qui est présent à plus d'un million de copies au sein du génome humain, représentant ainsi plus de 10 % de la séquence totale .

Les élément de classe I, non autonomes sont dérivés d'éléments autonomes. Par exemple, chez le riz, l'élément *Dasheng* serait issu de l'élément *RIRE2* qui a évolué et perdu une partie de sa séquence interne, son intégrase (Jiang *et al.* 2002). Ainsi, il est devenu non fonctionnel et nécessite l'intégrase de la famille de l'élément dont il est issu pour pouvoir transposer au sein du génome hôte.

2.2.2 Les transposons

Les éléments transposables de classe II, sont communément appelé transposons. Comme pour les éléments de classe I, il en existe deux classes qui ont été définies selon leur similarité de séquence. Les transposons sont trouvés au sein de la majorité des eucaryotes, mais également chez les procaryotes, comme les bactéries, où ces derniers sont appelé IS (*insertion sequences*).

Les transposons contiennent dans la majorité des cas une seule séquence codante, correspondant à la transposase qui est flanquée par des courtes répétitions inversées (TIR, Terminal Inverted Repeats) (Figure 2.1). La transposition est coordonnée par la transposase qui reconnaît les TIRs aux extrémités et coupe de part et d'autre. Ainsi, le transposon excisé peut ensuite s'insérer à un nouveau locus au sein du génome par ligation.

Les classes ont été déterminées selon la présence ou absence d'une triade catalytique DDE/DDD au sein de la transposase. Ainsi, la classe I possédant cette triade est représentée par les éléments de la famille *Tc1-Mariners* (Plasterk *et al.* 1999). Les autres

transposons ne possédant pas cette triade sont représentés par la famille *hAT, Mutator (Mu) et CACTA*.

L'autre classe de transposons correspond aux Hélitrons. Ces transposons, contrairement aux *Tc1-Mariner* utilise un mécanisme de cercle roulant pour leur cycle de transposition.

La famille *Maverick*, non présente chez les plantes, fait également partie de cette classe particulière de transposons.

La famille de transposons non autonomes la plus connue sont les MITE (*Miniature Inverted-repeat Transposable Elements*). Les MITE sont des éléments très courts, d'une centaines de paires de bases, qui sont flanqués par des TIRs et qui ont perdu leur séquence codante pour la transposase (Figure 2.1). On les retrouve le plus fréquemment au sein ou près des gènes (Feschotte *et al.* 2003). Ils tiendraient leur origine d'un autre transposon, *Tc1-Mariner*.

Avec leur système de transposition par couper-coller, on s'attendrait à ce que le nombre des transposons soit fixé au sein des génomes au cours du temps. Mais en analysant les génomes de différents eucaryotes, il a été observé que les transposons représentaient une fraction non négligeable des séquences répétées. Chez le riz, les *MITE* sont la classe la plus importante, en terme de nombre copies. Ainsi, on peut se demander comment ces éléments ont pu envahir les génomes. Après excision du transposon, la cellule met en place tous les mécanismes moléculaires pour la réparation de l'ADN au niveau de la région où se trouvait le transposon. Il est possible que la cellule utilise comme matrice homologue, un chromosome contenant un transposon pour cette étape. Ainsi à la fin de la réparation, il y a présence d'une nouvelle copie du transposon (Wicker *et al.* 2010). Il est à noter que c'est par ce mécanisme que les transposons peuvent augmenter leur nombre au sein des génomes pendant la réplication des chromosomes (Lisch, 2002)

Par l'analyse des insertions de MITE au sein d'une population de riz cultivé, Castanera et al ont démontré que l'amplification de ces ET n'était pas déterminée seulement par la présence des transposases mais que celle-ci était liée à la réplication de seulement quelques copies « maîtres », comme c'est le cas pour les rétrotransposons (Castanera *et al.* 2021).

Dans cette première partie, nous avons abordé la grande diversité des éléments transposables. Du fait de leur grande proportion au sein des génomes de plantes, de leur mobilité et de leur capacité à induire de la plasticité génomique, plusieurs auteurs ont proposé que les ET pourraient être de bons facteurs d'adaptation chez les plantes (E. Casacuberta *et al.* 2013 ; Rey *et al.* 2016). Mais cela nécessite cependant d'établir que la diversité génomique induite par la transposition a un impact fonctionnel. Ce dernier aspect sera abordé dans la seconde partie introductive de ma thèse.

2.3 Impact fonctionnel des éléments transposables

Précédemment, nous avons vu que les éléments transposables par leur mobilité contribuaient à la dynamique des génomes. Après leur transposition, les ET génèrent donc des mutations qui pourraient influencer l'expression des gènes au voisinage de leur point d'insertion, constituant ainsi de nouveaux allèles pour ces gènes. Si cela confère à l'hôte un avantage sélectif, la mutation provoquée par l'élément transposable sera conservée car sous pression de sélection positive. Par contre, si elle est délétère, les individus portant cette mutation seront très rapidement éliminés au sein de la population. Par conséquent cette insertion finira par être éliminée au cours du temps.

Jusqu'à présent quatre principaux mécanismes de régulation de l'expression des gènes par les ET ont été mis en évidence : la disruption des gènes (knock-out), la domestication des ET avec la formation de co-transcrits gènes-ET, la propagation de méthylation entre l'ET et les gènes avoisinants et la régulation à distance de l'expression des gènes (*cis-régulation*) (Feschotte, 2008 ; Galindo-González *et al.* 2017 ; Nishihara, 2019).

2.3.1 La disruption des gènes (*knock-out*)

Lors de la transposition d'un élément transposable, celui-ci peut s'insérer au sein de la séquence codante d'un gène et par conséquent modifier directement l'expression de celui-ci.

Par exemple chez l'humain, les insertions d'ET sont liées à certaines maladies génétique ou types de cancer. Certaines insertions de transposons *Alu* sont liées au cancer du sein, tandis que certaines insertions du rétrotransposon *L1* sont liées au cancer du colon (Scott *et al.* 2016).

Récemment, Hof et al. (Hof *et al.* 2016) ont montré que la présence d'un transposon *Carbonaria* au sein du gène *Cortex* de la phalène du bouleau induit la surexpression de celui-ci et augmente la pigmentation du papillon passant de blanc à gris/noir (Figure 2.3-A), mettant ainsi en évidence qu'un mouvement d'ET est à l'origine de l'un des processus adaptatifs par sélection le mieux connu et le plus illustratif de la théorie Darwinienne.

Chez les plantes, certains gènes sont préférentiellement ciblés par les ET, comme ceux impliqués dans la date de floraison. Il a été observé de nombreuses insertions d'ET différents au niveau du locus *FLC* (flowering locus C) chez *Arabidopsis thaliana* (Quadrana *et al.* 2016). Ces insertions induisent une diminution de l'expression du gène qui aboutira à une floraison précoce de la plante (Horváth *et al.* 2017).

Chez la carotte, l'insertion d'un transposon *Tc1/Mariner* au sein du gène *DcMYB7* diminue l'expression de ce dernier et altère la régulation des anthocyanes, impliquées dans la pigmentation de la racine. Ainsi les carottes avec cette insertion seront non pas de couleur violette mais de couleur jaune-orangée (Xu *et al.* 2019).

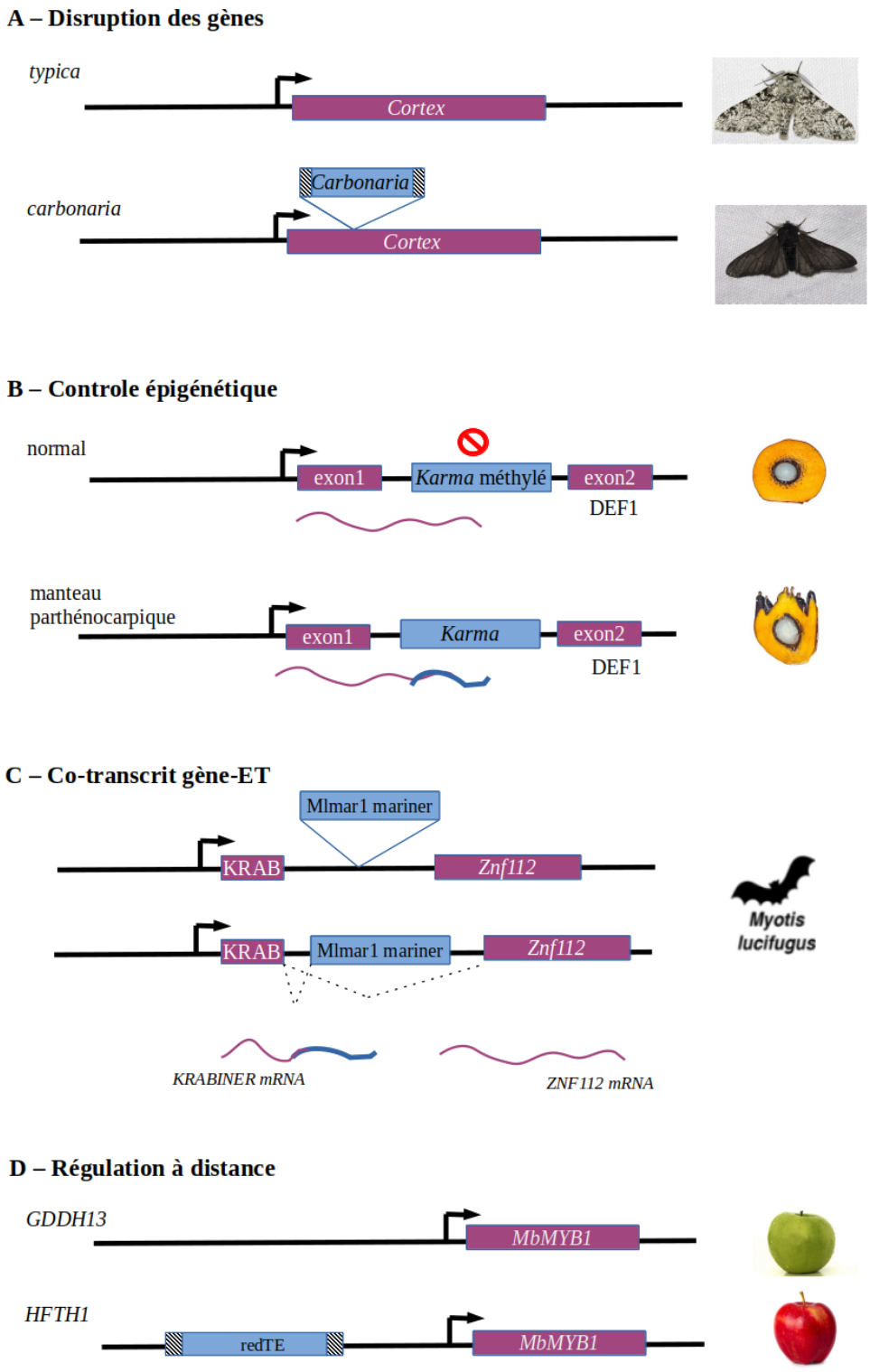


FIGURE 2.3 – **Impact fonctionnel des éléments transposables**. En prune et en bleu sont représentés les gènes et les ET, respectivement. **A** - Disruption des gènes : Chez la phalène du boulot, l'insertion d'un rétrotransposon *Carbonaria* au sein du gène *Cortex* induit la surexpression de celui-ci et augmente la pigmentation du papillon, qui aura ainsi une coloration plus foncée (HOF *et al.* 2016). **B** - Contrôle épigénétique : Chez le palmier à huile, l'expression de l'ET *karma* provoque un transcrit alternatif issu du gène *DEF1* avec une terminaison précoce. Cette dérégulation épigénétique va modifier la forme du fruit du palmier, qui ne sera plus fertile (ONG-ABDULLAH *et al.* 2015). **C** - Co-transcrit gène-ET : Chez la chauve-souris, l'insertion d'un transposon *mariner Mlmar1* au niveau des domaines régulateurs KRAB induit un évènement d'épissage alternatif, qui engendre ainsi de nouveaux régulateurs d'expression (COSBY *et al.* 2021). **D** - Régulation à distance : Chez la variété de pomme *GDDH13*, l'insertion du rétrotransposon *redTE* près du gène *MbMYB1*, impliqué dans les voies de pigmentation du fruit, conduit à une nouvelle variété rouge *HFTH1* (ZHANG *et al.* 2019).

Chez le millet, de multiples événements d'insertions d'un rétrotransposon *TSI* au sein du gène *waxy GBSS1* conduisent à une modification du taux d'amylase au sein de l'endosperme du grain de millet. Ainsi les grains de cette céréale seront plus collants. Ce phénotype *sticky millet* a été sélectionné par les agriculteurs lors de la domestication du millet en Asie (Kawase *et al.* 2005).

Chez le palmier à huile, l'hypométhylation du transposon *karma* induit l'expression de celui-ci. Le transposon *karma* est inséré au sein de l'intron du gène *DEF1*, impliqué dans la mise en place des organes mâle de la plante. Cette expression va provoquer un transcrit alternatif avec une terminaison précoce (Figure 2.3-B). Cette dérégulation épigénétique va modifier la forme du fruit du palmier, ce nouveau phénotype est appelé "manteau parthénocarpique" (*mantled abnormality*). Par conséquent, le fruit ne sera plus commercialisable (Ong-Abdullah *et al.* 2015).

2.3.2 Domestication des ET : co-transcrit gène-ET

Les éléments transposables peuvent être également domestiqués, co-optés par l'organisme hôte. L'ET nouvellement inséré va être ainsi considéré comme un exon et pris en charge par la machinerie transcriptionnelle de l'hôte. Ainsi, des co-transcrit gènes-ET peuvent être produits. Les activités biochimiques des protéines dérivées de l' ET ont été co-optées à plusieurs reprises au cours de l'évolution pour favoriser l'émergence des innovations cellulaires convergentes dans différents organismes.

Chez la plante modèle *Arabidopsis thaliana*, *DAYSLEEPER* encode une transposase domestiquée de la super-famille *hAT* (classe II), qui est essentielle pour le développement de la plante (Bundock *et al.* 2005). De plus, cette transposase a été retouvée chez de nombreux autres angiospermes, comme le riz et la vigne (Knip *et al.* 2012).

Chez les tétrapodes, Cosby *et al.* ont mis en évidence la capture de domaine de transposase d'élément de classe II par épissage alternatif : phénomène qui s'est produit plusieurs fois indépendamment au cours des 350 millions d'année d'évolution. La transposase se fusionne préférentiellement avec les domaines régulateurs KRAB (*Krüppel-associated box*) (Figure 2.3-C) qui engendrent ainsi de nouveaux répresseurs transcriptionnels (Cosby *et al.* 2021).

Dans le génome du murier, il a été montré que de multiples insertions de *MITE* étaient fréquemment associées à des évènements d'épissage alternatif et en particulier d'éxonisation (acquisition de nouveaux exons issue de séquences d'ET) (Xin *et al.* 2019).

2.3.3 La diffusion de méthylation

En conditions normales, les éléments transposables sont sous contrôle épigénétique. L'expression de ces ET est réprimée *via* de nombreux mécanismes épigénétiques tels que la méthylation, la régulation par petits ARN (smRNA) ou la conformation de la chromatine. Toutes ces modifications peuvent avoir un impact sur les gènes au voisinage de l'ET. Par diffusion de cette répression épigénétique des ET vers les gènes, l'expression de ces derniers sera modifiée, souvent de manière négative (sous-expression).

Cela a été mis en évidence, par exemple, en comparant les deux espèces proches *Arabidopsis thaliana* et *Arabidopsis lyrata*. *A. lyrata* comporte 2 fois plus de copies d'ET comparé à l'espèce modèle *A.thaliana* (son génome a d'ailleurs une taille deux fois supérieure à cette dernière). La répression de ces ET *via* petits ARN interférents (siRNA) est associée à une diminution de l'expression des gènes à proximité. De plus, au cours du temps, ces ET seront soumis à une sélection purifiante (Hollister *et al.* 2011).

2.3.4 La régulation à distance (*cis-regulation*)

Les éléments transposables comportent des sites de liaison de facteurs de transcription (*TFBS : transcription factor binding site*). Les rétrotransposons à LTR portent ces motifs TFBS au sein des LTR. Ils peuvent être considérés comme des sources de diversité des réseaux de régulation de l'expression des gènes. A noter que les *LINE* contiennent également cette séquence de fixation au niveau de la région 5'UTR. Ils contribuent d'ailleurs à une fraction non négligeable de sites de fixation de facteur de transcription au sein des génomes. 25 % des promoteurs des gènes chez l'humain contien-nent des séquences d'ET (Jordan *et al.* 2003). Chez le riz 58 % des gènes sont associés à des MITE (C. Lu *et al.* 2012). La transposition peut représenter ainsi un mécanisme d'ajout de nouveaux sites de fixation, qui pourraient être à l'origine d'innovation génétique dans les populations. De plus, au cours de l'évolution, les ETs vont accumuler des mutations qui pourront modifier leur site de fixation et donc leur activité régulatrice. Par la présence de ces TFBS à proximité des gènes, la fixation de facteur de transcription va modifier l'expression des gènes avoisinants.

C'est ce processus qui est à l'origine de l'une des innovations biologiques majeure de la lignée des mammifères : la forte propagation du rétrotransposon *MER20* comme activateur d'expression de gènes endométriaux chez l'ancêtre des mammifères conduit à une reprogrammation massive des gènes de développement à l'origine de la grossesse placentaire (Lynch *et al.* 2011).

La conservation au cours de l'évolution de ces sites de liaison indique que les rétrotransposons font partie intégrante des voies de régulation d'expression des gènes chez l'hôte, notamment chez les plantes, du fait de leur grand nombre de copies au sein des génomes. De nombreux traits phénotypiques des espèces domestiquées sont liés à la pré-

sence d'ET. Les gènes cibles sont des gènes de la voie de régulation de biosynthèse des anthocyanes, gènes impliqués dans la pigmentation des fruits. Ainsi ces insertions d'ET vont être sélectionnées par les agriculteurs et fixées au cours de l'évolution de l'espèce. L'insertion de l'ET induit une diversité phénotypique.

Comme par exemple chez la pomme avec l'insertion du rétrotransposon *redTE* près du gène *MbMYB1*, impliqué dans les voies de pigmentation du fruit, qui conduit à une nouvelle variété rouge *HFTH1*. (Zhang *et al.* 2019, Figure 2.3-C) Il existe de nombreux exemples similaires liés à la couleur des fruits au sein des plantes cultivées, comme l'insertion d'un rétrotransposon *Gret1* au sein de la région promotrice du gène *MYBA1* chez la vigne créant ainsi la nouvelle variété à grains blancs *Chardonnay* (Kobayashi, 2004) ou suite à son élimination non totale, la variété *Ruby Okuyama* à grains rosés (Lisch, 2013). Chez l'orange, l'insertion d'un rétrotransposon va induire la surexpression du gène *Ruby* à proximité. Ce gène est impliqué dans la voie de biosynthèse des anthocyanes. Ainsi, la surexpression de ce régulateur va donner la pigmentation rouge à l'orange, typique de l'orange sanguine de la variété domestiquée *Moro* (Butelli *et al.* 2012).

Les insertions d'ET peuvent également influencer la morphologie des plantes, comme chez la tomate avec le rétrotransposon *Rider* qui contribue à la duplication génomique de 24,7kb dont le gène *SUN*, l'un des principaux gènes contrôlant la forme allongée des fruits. Cet événement a ainsi modifié le contexte génomique du *locus*, qui va induire une surexpression de celui-ci, conduisant au final à l'élongation du fruit (Xiao *et al.* 2008).

L'insertion de l'ET peut avoir un effet à longue distance. C'est le cas chez le maïs domestiqué avec l'insertion d'un rétrotransposon *Hopscotch*, 60kb en amont du gène de domestication *tfb1* (teosinte branched). Cette insertion induit la surexpression du gène qui va permettre la croissance apicale du maïs. Cette caractéristique est une des étapes clés de la domestication du maïs (Studer *et al.* 2011).

On peut également observer ce phénomène au sein de plantes sauvages, telle que la plante modèle de laboratoire *Arabidopsis thaliana* chez laquelle il a été montré que des insertions anciennes d'ETs contenaient également des sites de facteur de transcription (TFBS). De plus, ces insertions étaient conservées au sein des *Brassicaceae*, avec une surreprésentation au niveau des régions promotrices des gènes (5'UTR) impliquées dans la floraison. Ainsi cela suggère que ces ET ont eu un impact fonctionnel à long terme sur le développement de ces plantes et donc ont été conservés au cours du temps (Baud *et al.* 2019).

La diversité des ET, leur mode de transposition et leur caractéristiques intrinsèques impliquent ainsi de nombreuses sources de mutations et de régulateurs au sein des génomes.

Que ce soit aux niveaux micro- ou macro-évolutifs, on voit ainsi que les ETs peuvent jouer un rôle important dans l'évolution des eucaryotes en contribuant à la diversité adaptative des populations mais également à l'émergence d'innovations biologiques ma-

jeures qui sont à la base de nouvelles lignées (Chuong *et al.* 2017). Mais pour pouvoir avoir un tel impact fonctionnel sur leur hôte, ces derniers doivent d'abord être exprimés, ce qui pose la question sous-jacente des conditions nécessaires à leur activation.

Nous aborderons ce point dans la partie suivante.

2.4 Activation de la transposition par le stress

Les ETs sont majoritairement inactifs en condition normale car ils sont réprimés par plusieurs mécanismes épigénétiques du génome de l'hôte (méthylation de l'ADN, ET au sein de l'hétérochromatine, régulation par les petits ARNs) (Rigal *et al.* 2011).

Les éléments transposables sont ainsi sous-contrôle du génome hôte. Au cours du temps, ce controle épigénétique conduira à l'élimination rapide de l'ET. Ainsi l'élement est face à un ultimatum : "sauter ou mourir". Pour survivre, l'élement peut sauter par transfert horizontal au sein d'un génome naïf (ie, n'ayant pas eu d'insertion de cet élément) (El Baidouri *et al.* 2014 ; Gilbert *et al.* 2018) ou grâce à sa réactivation en condition de stress, il pourra être sélectionné par la suite pour son effet adaptatif.

Lorsqu'une plante subit un stress environnemental (biotique ou abiotique), celle-ci se "défend" en enclenchant des voies cellulaires de réponse à ce stress dont une est la diminution du niveau de méthylation de l'ADN qui va permettre l'activation de gènes de défense (Lanciano *et al.* 2018). Ces modifications épigénétiques globales du génome vont conduire à une réactivation transcriptionnelle des ETs (Secco *et al.* 2015). Suite au relâchement de cette répression épigénétique, certains rétrotransposons à LTR vont ainsi pouvoir débuter leur cycle de rétrotransposition qui va générer une ou plusieurs nouvelles copies de l'élément (Figure 2.4-A). Selon la nature du stress, seulement certaines familles d'ET vont êtres activées. De plus, au sein de ces familles, seules certaines copies seront actives (Figure 2.4-C).

Chez le tabac, la présence de micro-organismes va réactiver l'expression du rétrotransposon *Tnt1*, dont les LTR contiennent des motifs régulateurs liés à la défense contre les pathogènes. Ainsi sa réactivation permet d'activer les mécanismes de défense de la plante (Grandbastien *et al.* 2005).

Les ET peuvent également être activés par un changement de température. Ainsi la chaleur va permettre l'expression du rétroélément *ONSEN* chez *Arabidopsis thaliana* qui contient au sein de ces LTR des séquences régulatrices HRE (*heat response elements*) (Cavrak *et al.* 2014). Ces séquences régulatrices au sein des ET sont retrouvées chez la plupart des *Brassicacae*, ainsi peut-on penser que cette réponse adaptative liée à cet ET a été sous sélection positive. Malgré cette activation, les insertions des nouvelles copies d'*ONSEN* n'ont été observées qu'au sein de plantes dont les voies épigénétiques étaient non-fonctionnelles (Ito *et al.* 2011). Ceci suggère donc que le stimulus externe (ici la chaleur) est une condition nécessaire mais pas suffisante à la réactivation de l'élément, le fond génétique jouant également un rôle important.
C'est la combinaison de ces deux facteurs : un épigénétique et un stimulus externe qui vont permettre la transposition des copies de l'ET néo-transcrit (Grandbastien, 2015).

La culture cellulaire *in vitro*, peut également induire la transposition. Le premier

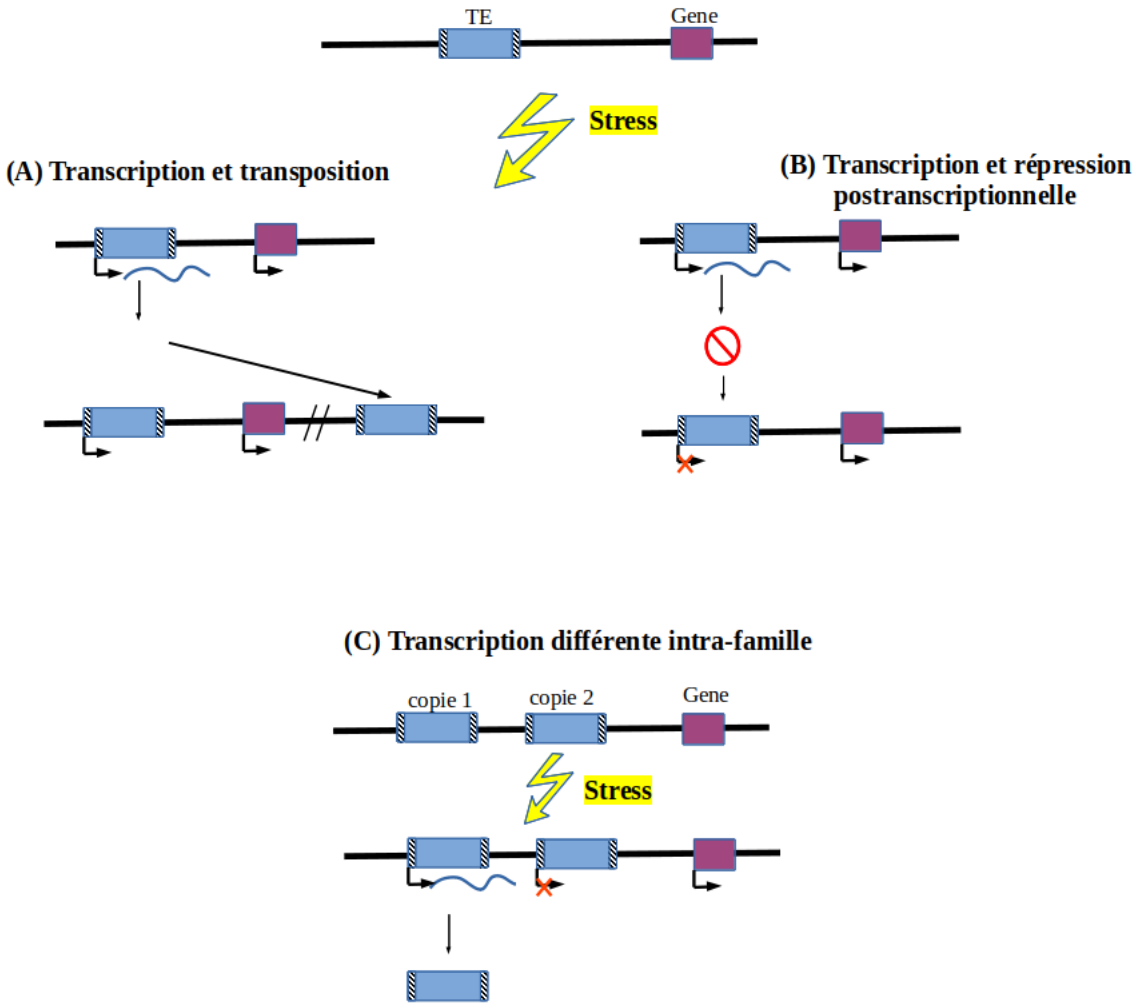


FIGURE 2.4 – **Les éléments transposables et le stress**. En prune et en bleu sont représentés les gènes et les ET, respectivement. **(A)** La transcription de l'ET peut être activée en condition stress. Le transcrit récemment produit peut conduire à une nouvelle copie de l'élément qui pourra transposer au sein d'une autre région du génome. **(B)** Suite au stress, certains ET peuvent être activés mais seront rapidement réprimés par les voies épigénétiques (méthylation). **(C)** Au sein d'un génome, seulement certaines familles d'ET vont réagir à un un stress donné. De plus, au sein même de la famille, seulement certaines copies seront activées quand d'autres seront réprimées. Ainsi on peut dire que l'activation des ET par le stress est famille et copie dépendante.

exemple de réactivation des éléments transposables a été mis en évidence chez le tabac, avec le rétrotransposon *Tnt1* dans les cultures de protopplastes (Grandbastien *et al.* 1989). Chez le riz *Oryza sativa Nipponbare*, la culture de cals a permis de mettre en lumière la réactivation de deux rétrotransposons à LTR *Tos17*, le premier rétrotransposon détecté actif chez le riz (Hirochika *et al.* 1996) et *Tos19/Lullaby* (Picault *et al.* 2009). Dans ces deux exemples, il a été observé qu'une seule copie du rétrotransposon (celle du chromosome 7 pour *Tos17* et celle du chromosome 6 pour *Lullaby*) sur les deux copies natives présentes au sein du génome de référence, est active.

Les ET de classe II peuvent également être actifs en condition de culture cellulaire. Chez le riz, la famille de transposons constituée de *Ping et mPing* est réactivée. *mPing* représente la version non autonome du transposon *Ping* : il utilisera ainsi toute la machinerie de ce dernier pour sa transposition. Ainsi, on observe leur nombre de copies augmenter fortement au sein des cultures cellulaires, suite à un stress froid (L. Lu *et al.* 2017).

Bien que l'activation de certains ET puisse jouer un rôle adaptatif temporaire comme nous venons de le montrer, elle peut avoir au contraire un impact négatif sur l'hôte. Par exemple, les rétrotransposons humains *HERV-W* sont surexprimés au sein des cellules cancéreuses, induisent la production de nombreuses protéines d'enveloppe qui peuvent affecter l'activité des cellules saines comme celles du cerveau ou celles du système immunitaire (Ruprecht *et al.* 2008).

Les différents exemples cités ci-dessus illustrent l'impact fonctionnel que les ETs peuvent avoir sur leur hôte. Au sein de la partie suivante, je vais maintenant présenter les connaissances que l'on a de leur impact sur la structure des génomes.

2.5 Impact structural des éléments transposables

Contrairement à ce que l'on pourrait supposer au sein des plantes, la diversité des ET n'augmente pas avec la taille du génome (Elliott *et al.* 2015). Ce sont en effet seulement quelques familles qui sont présentes en multiples copies au sein des génomes. Par exemple, chez la plante *Vicia parmonica*, une unique famille de rétrotransposon (nommée *Ogre*) représente près de 38% du génome (Neumann *et al.* 2006). A noter que la majorité de ces copies sont inactives car elles sont réprimées via des mécanismes épigénétiques (Figure 2.4-B). Chez le riz sauvage *Oryza australiensis*, il a été observé un doublement de la taille du génome comparé à l'espèce cultivée apparentée *O. sativa*. Ce changement de taille de génome peut être expliqué par l'amplification de seulement 4 familles principales de rétrotransposons : *Dingo, Wallabi, Kangourou et RIRE1* (Piegu *et al.* 2006).

Comme énoncé auparavant, les rétrotransposons jouent un rôle important dans l'évolution de la taille des génomes en se copiant d'un endroit à un autre du génome. La question qui se pose est quels sont les mécanismes qui permettent de limiter la taille des génomes et donc la propagation des ET ?

À la suite d'un événement de transposition, les mécanismes de recombinaison sont activés pour limiter le nombre d'ET au sein des génomes des plantes. Ainsi ces derniers sont éliminés via recombinaison homologue intra ou inter-rétrotransposons via les LTR ou à travers des cassures double-brins (Figure 2.5, Ma, 2004 ; Vitte *et al.* 2007 ; Novák *et al.* 2020). Le taux de délétions des rétrotransposons par exemple chez le riz est estimé à 3,62 kb/million d'années par élément (Stein *et al.* 2018). Au cours de l'évolution, la proportion d'éléments transposables entiers et actifs va diminuer, et par conséquent la présence d'ET partiels, tel que les solo-LTR, les éléments non-autonomes (comme les MITE) vont augmenter au sein des génomes. L'ET au cours du temps va accumuler des mutations (des petites insertions et délétions) au sein de sa séquence, jusqu'à devenir non-fonctionnel. Ces ET seront donc fragmentés et inactifs, devenant par conséquent de l'ADN non caractérisé, appelé aussi « dark matter » du génome (Maumus *et al.* 2016).

Tout ces mécanismes mettent en évidence la forte contribution des ET à la dynamique structurale des génomes (Anderson *et al.* 2019).

En plus de leur rapide évolution, les ET ne sont pas distribués aléatoirement dans le génome. Selon leurs caractéristiques, ils ont des préférences d'insertions au niveau de certains compartiments génomiques, mais ils ont tout de même tendance à s'insérer au niveau des régions péri-centromériques (Bourque *et al.* 2018 ; Quesneville, 2020). Ce phénomène s'observe surtout chez les rétrotransposons. De par leur nature et leur mode de transposition par multiplication, les éléments transposables insérés aux niveaux des régions non exprimées (hétérochromatine) seront moins éliminés par rapport à des ET dont la présence va diminuer la *fitness* de l'organisme. Ainsi l'hôte va mettre en place

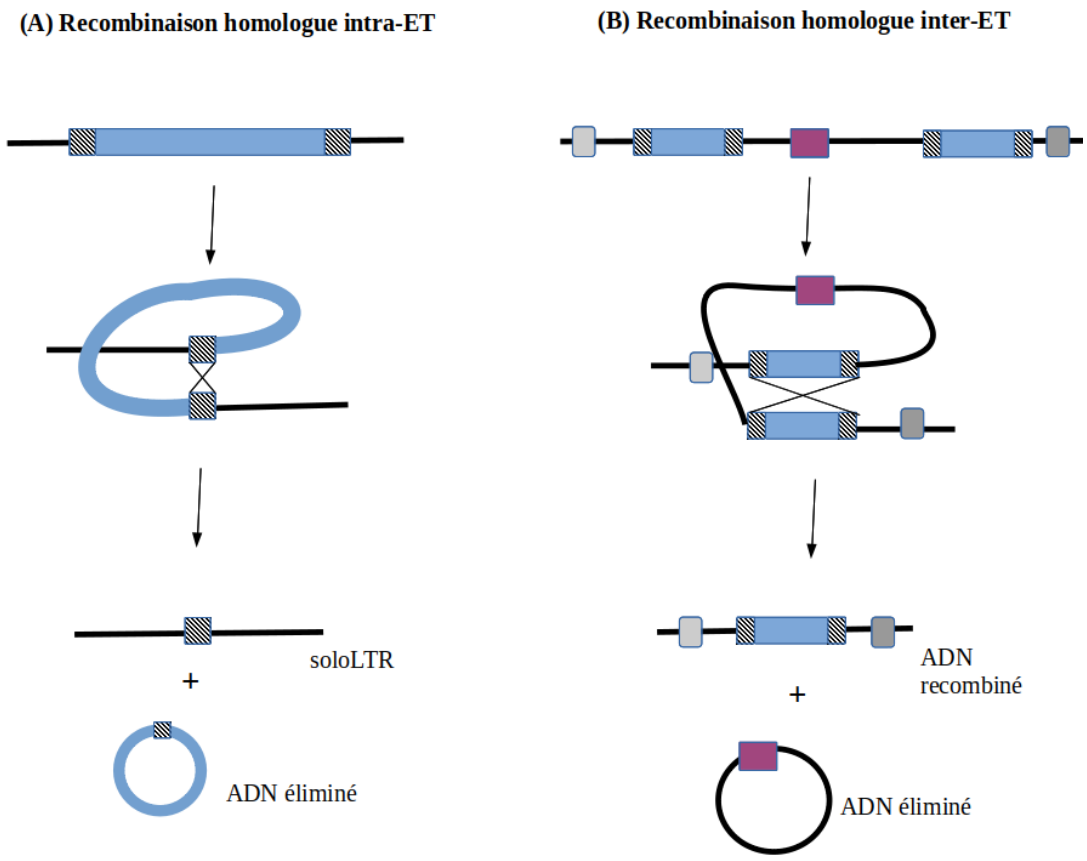


FIGURE 2.5 – **Elimination des rétrotransposons.** En bleu : les rétrotransposons, en rose : un gène. **A**-La recombinaison homologue intra-ET : au sein d'un rétrotransposon à LTR, il y a recombinaison entre les 2 LTRs qui conduit ainsi à l'élimination du rétrotransposon avec un seul LTR. A la fin de la recombinaison, il reste donc un solo-LTR. **B**-La recombinaison homologue inter-ET : 2 copies de rétrotransposons à LTR similaires vont recombiner conduisant à l'élimination de la séquence génomique située entre les 2 rétrotransposons, qui peut contenir des gènes. A la fin de cet évenement, il restera donc une copie du rétrotransposon.

tous les mécanismes nécessaires (voies épigénétiques, élimination par recombinaison) à l'élimination rapide de l'ET donné. De plus, il a été observé qu'au sein des rétrotransposons, la famille des *Copia* s'insère plus près des gènes que ceux de la familles des *Gypsy* (J. M. Casacuberta *et al.* 2003).

Le succès et la diversité des ET dans un génome sont façonnés à la fois par des propriétés intrinsèques aux éléments ainsi que par des forces évolutives agissant au niveau de l'espèce (Sanchez *et al.* 2017).

Nous avons abordé l'importance du rôle des éléments transposables au niveau macro-évolutif et leur grande dynamique sur le long terme entre les espèces. Il existe une grande diversité des ET selon leur nature et également en fonction de l'espèce hôte. Ainsi, on peut se demander quel est l'impact de ces éléments transposables au niveau de l'espèce et au sein d'une population, c'est à dire au niveau micro-évolutif.

Les ET, du fait de leur dynamique sont une grande source de polymorphismes génétiques entre les individus proches. Par conséquent la majorité des ET n'est pas fixée au sein des populations (Goubert *et al.* 2020). Il y a une grande diversité des insertions d'éléments transposables au sein d'une espèce. De ce fait une part non négligeable du génome est très instable et spécifique à chaque variété (environ 20% chez *Arabidopsis thaliana*), montrant ainsi l'évolution rapide des éléments transposables au sein des génomes de plantes (Quadrana *et al.* 2019). Chez le riz, par comparaison de 3 000 génomes, la même observation a été mise en évidence (Fuentes *et al.* 2019). Cette forte dynamique des ET au sein d'une population de riz cultivé va être décrite par la suite du au sein du chapitre 6 de mon manuscrit (Carpentier *et al.* 2019).

Par l'évolution rapide des techniques de séquençage et la disponibilité croissante de séquences génomiques, cet impact au niveau structural peut de plus être étudié par des approches de génomique comparative au sein des populations. Ce point sera abordé par la suite dans la partie introductive. Je montrerai en effet qu'il est maintenant possible d'étudier la diversité du paysage transpositionnel entre les variétés d'une même espèce.

CHAPITRE 3

Nouvelles technologies de séquençage et détections des éléments transposables

3.1 Les nouvelles technologies de séquençage

Le séquençage de l'ADN consiste à déterminer l'ordre de l'enchaînement des nucléotides d'une molécule d'ADN donnée. Depuis la première méthode de séquençage Sanger en 1970, les différentes technologies ont fortement évolué. La méthode développée par Sanger est basée sur la réaction de polymérisation de l'ADN avec des didésoxyribonucléotides (ddNTP).

La première étape est l'amplification d'un fragment d'ADN par polymérisation à partir d'une amorce marquée. L'incorporation des ddNTP stoppe la réaction, générant ainsi des fragments de différentes tailles. Puis ces fragments sont traités par électrophorèse. A chaque extrémité se trouve un fluorochrome spécifique selon la base (ddNTP). De ce fait par détection de couleur et l'alignement des fragments selon leur longueur, il est possible de lire la séquence dans son ensemble. Cette technique permet la lecture d'un fragment d'ADN d'une taille maximum de 1000pb environ.

Au cours des trentes dernières années, de nouvelles technologies de séquençage ont vu le jour, révolutionnant le domaine de la génomique en ayant toutes en commun une réduction drastique du coût de la séquence. Actuellement, nous sommes à la 3ème génération de technologie de séquençage.

3.2 Séquençage haut-débit (2ème génération)

La seconde génération est apparue en 2005 pour pallier le prix élevé et le faible débit du séquençage Sanger. C'est le début du séquençage haut-débit, développé par les entreprises Roche et Illumina. Ce séquençage consiste tout d'abord à une fragmentation de l'ADN puis une amplification par PCR et se finit par à une incorporation de nucléotide à chaque cycle. Toutes ces étapes se font en quelques jours (2-3 jours). A noter que la technologie Roche a été très rapidement remplacée par la technologie Illumina qui permet un meilleur débit de séquençage à faible coût.

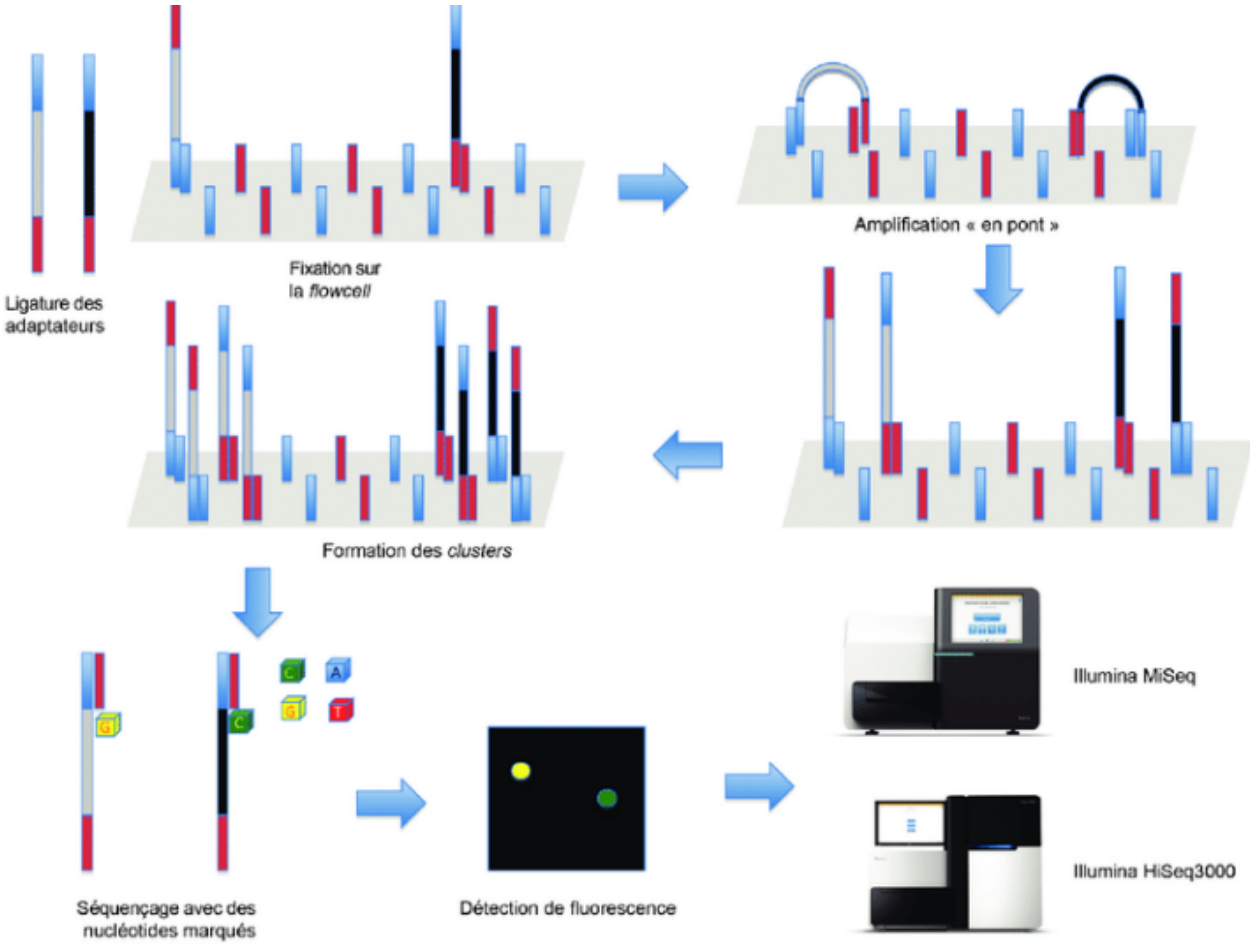


FIGURE 3.1 – **Le séquençage Illumina**. Après fragmentation de l'ADN, il y a ligation des adapteurs de séquençage qui vont permettre la fixation des fragments produits sur la plaque. Une amplification "en pont" s'effectue à chaque cycle de séquençage qui conduit à la formation de cluster. Grâce au marquage par fluorescence de chaque base, par analyse d'image, il est possible de déterminer quel nucléotide a été incorporé et d'obtenur la séquence du fragment.

Le séquençage développé par Illumina se fait sur plaque. Après fragmentation de l'ADN, celui ci est ligué à des adaptateurs en 3' et 5'. Ensuite le fragment est fixé à la

plaque de séquençage (flowcell) via complémentarité avec les sondes déjà présentes sur celle-ci. Une amplification "en pont" s'effectue et au cours des cycles de séquençage des clusters de fragments identiques se forment. Le séquençage des clusters s'effectue par incorporation de nucléotides marqués : chaque base a une fluorescence spécifique, qui permet ainsi la lecture des fragments d'ADN par analyse d'images (Figure 3.1). Depuis une dizaine d'années, la société Illumina a pris le monopole du séquençage de 2ème génération en améliorant continuellement le débit des séquenceurs tout en en diminuant le coût. Leur dernier séquenceur NovaSeq produit plus de 20 milliards de lectures en 2 jours à un très faible coût (moins de 10€ le Gb). En augmentant le débit, le prix de séquençage d'un génome (du Gb) va sûrement continuer à décroître avec le temps, permettant ainsi de nouvelles opportunités de séquençage : espèces avec de grande taille de génome, multiplication des échantillons.

L'avantage de cette technologie est en premier lieu le faible coût du séquençage par base, car à l'issu du séquençage on obtient un grand volume de données (centaines de millions de lectures en moyenne). De plus ces lectures comportent un taux d'erreur très faible (inférieur à 1%). Cependant la limite de cette technologie est la courte longueur des lectures : une à trois centaines de bases seulement, ce qui est un frein pour l'assemblage d'un génome et la détection de variants structuraux, comme nous allons le voir dans la suite de cette thèse.

Néanmoins, avec cette nouvelle technologie il est possible de re-séquencer des génomes de plantes. De nombreuses génomes de plantes cultivées ont été (re)séquencées tel que le maïs en 2009, la pomme de terre en 2011... Les génomes des espèces modèles comme *Arabidopsis* ou le riz, ont déjà été séquencées en utilisant la première génération de séquençage via le séquençage de BAC (Bacterial Artificial Chromosome). Mais ils ont été reséquencés avec cette nouvelle technologie. Ces nouveaux projets de reséquençage ont un coût et un temps de séquençage incomparablement plus faible. Néanmoins les régions complexes telles que les centromères, les télomères, les régions avec de nombreuses séquences répétées sont difficiles à assembler si l'on dispose que de courtes lectures. Pour améliorer la qualité de l'assemblage, une stratégie de séquençage hybride est utilisée en mutualisant les séquençages de courtes lectures avec des cartes physiques et longues lectures comme les BACends ou celles générées par des séquenceurs de troisième génération (voir plus bas) pour permettre l'ancrage des séquences Illumina.

L'une des conséquences de la baisse continue des coûts de séquençage est l'émergence de projets de génomique des populations : tels que le séquençage de 1135 génomes d'*Arabidopsis* (1001 Genomes Project, 2016), 1001 génomes humain (IGSR, 1KGP Project, 2010) , 3000 génomes de riz cultivés *Oryza sativa* (2014), 1011 génomes de levure *Saccharomyces cerevisiae* (2018), 1100 génomes de drosophiles *Drosophila melanogaster* (Drosophila Genome Nexus, 2016), 1000 génomes de vaches (1000 Bull Genomes Project, 2019) ou plus récemment le grand projet à visée épidémiologique de séquençage humain

de 18,000 génomes par an (Goldfeder *et al.* 2017) et le projet *Tree of Life* qui consiste à séquencer plus de 60,000 espèces eucaryotes des îles britanniques.

A l'aide de la génomique et la génétique des populations, l'analyse de l'adaptation des individus d'une même espèce dans des niches écologiques différentes par des approches de génomique comparative est désormais possible. L'objectif est de mettre en évidence les régions communes entre les individus d'une même espèce (core génome), de celles qui sont plus spécifiques ou partagées par quelques individus de l'espèce (*dispensable genome*) (Carlos Guimaraes *et al.* 2015). Cela était déjà possible auparavant par analyse de FST (Indice de fixation) mais sur certains locus seulement, maintenant il est possible d'effectuer cette analyse tout au long du génome. Au vu des caractéristiques des éléments transposables que nous avons vu précédemment, il est attendu que ces derniers représentent une part non négligeable de ce pangénome.

3.3 Séquençage de 3ème génération

Les technologies de séquençage de 3ème génération ont été développées pour augmenter la longueur des lectures (considérée beaucoup trop faible pour l'assemblage de génomes complexes par exemple). Elles sont basée sur le séquençage d'une seule molécule d'ADN en temps réel. Contrairement aux générations précédentes, il n'est plus nécessaire d'amplifier l'ADN et de le fragmenter, la molécule est directement « lue ».

Les deux principales technologies disponibles en 2021 sont Pacific Biosciences (PacBio) et Oxford Nanopore (Nanopore) (Figure 3.2).

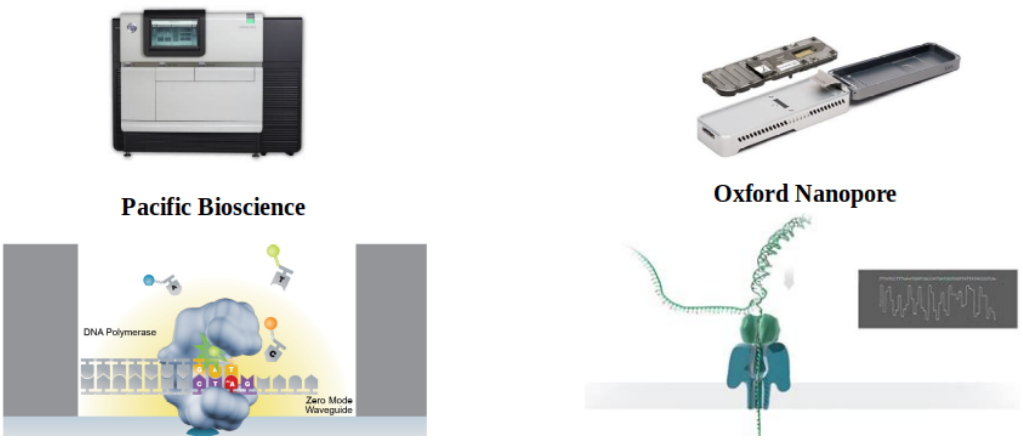


Figure 3.2 – **Les deux techniques de séquençage de 3ème génération**. Le PacBio effectue la synthèse de l'ADN en temps réel liée à une détection de fluorescence. Le MinIon d'Oxford Nanopore détecte le potentiel électrique de chaque base lors du passage de la molécule d'ADN au sein d'un nanopore.

PacBio suit en temps réel la synthèse de l'ADN par la polymérase présente au fond du puit de la flowcell. Chaque nucléotide est couplé à un fluorochrome distinct, ainsi lors de

son incorporation la fluorescence est détectée, permettant de déterminer quel nucléotide a été incorporé. La longueur des séquences produites est d'une dizaine de kilobases en moyenne, mais avec un débit d'au maximum 500,000 lectures à quelques millions avec le Sequel II dernière génération. Grâce à de nouveaux protocoles, les lectures PacBio ont maintenant un taux d'erreur très faible (environ 1%) proche de celui des lectures Illumina. Par contre, le prix du séquençage et la qualité ultra-pure nécessaire de l'ADN génomique est un des inconvénients du PacBio.

De son côté, la technologie d'Oxford Nanopore détecte les bases au fur et à mesure qu'une molécule d'ADN simple brin passe à travers un nanopore (pore protéique). La détection des bases s'effectue grâce à la variation du potentiel électrique spécifique d'un groupe de bases. Les séquences produites ont une taille moyenne de 20 kilobases et environ 10Go-20Go de données sont produites (environ 1 million de lectures). De plus, le petit format du séquenceur MinIon, similaire à une clé USB (Figure 3.2) permet une grande portabilité pouvant aller jusqu'au séquençage direct sur le terrain. C'est cette technologie qui a été utilisée pour ma thèse.

Les lectures obtenues par séquençage par le MinIon sont de très grandes longueurs, mais celle-ci contiennent de nombreuses erreurs (environ 10%). Le nombre de séquences est assez élevé (environ 15Go) et le séquençage ne dure que quelques heures. De plus le coût est assez faible, environ 1000€ pour un run de MinIon. Ce n'est pas une technologie adaptée pour l'étude de polymorphisme de type SNP du fait du taux d'erreur qui reste assez important, mais elle s'avère particulièrement efficace pour la détection de variants structuraux.

Avec l'amélioration des différents séquenceurs au sein d'Oxford Nanopore Technology, il est maintenant possible de séquencer une population d'individus. Avec le PromethIon, qui comporte jusqu'à 96 flowcells (= 96 MinIon), il est maintenant possible de produire non plus des gigabytes (Gb) de données mais des terabytes (Tb) à faible coût. Ainsi 100 génomes de tomates ont été séquencés en 100 jours par l'équipe de Michael Schatz pour étudier les variations structurales au sein de cette population (Alonge *et al.* 2020). 11 génomes issus de lignées cellulaires humaines ont également été séquencées en 9 jours produisant 2,3Tb de séquences avec un N50 à 42Kb (Shafin *et al.* 2020).

De plus l'amélioration ces dernières années de l'algorithme de lecture du potentiel électrique lié à la molécule d'ADN a permis une nette amélioration de la qualité des lectures Nanopore passant de 10% d'erreur à seulement 2% d'erreur. Néanmoins, cela nécessite de grandes capacités de calculs spécifiques avec l'acquisition d'un système avec une carte graphique puissante dont les calculs sont basés sur une unité de traitement graphique (GPU) et non centrale (CPU). Au cours de ma thèse, le laboratoire a acquis un serveur GPU, ce qui m'a permis d'optimiser mes protocoles d'analyse de données Nanopore (Chapitre 7).

Ainsi, cette nouvelle technologie est l'outil adéquat pour l'analyse des variations de structure d'un génome et également pour l'assemblage *de novo*, en couplant avec la 2ème génération pour corriger le taux d'erreur des longues lectures, même si les dernières versions de *basecalling* permettent d'envisager la seule utilisation des données Nanopore. Ces dernières années, le couplage de ces 2 technologies a été utilisé pour l'assemblage de génome *de novo* de différentes plantes dont des génomes complexes : comme 2 variétés de riz *Basmati et sandri* (Choi *et al.* 2020), 2 génomes de *Brassica nigra* (Perumal *et al.* 2020) ou le noyer (Marrano *et al.* 2020). A noter qu'un assemblage haute qualité d'*Arabidopsis thaliana* a également été effectué avec seulement une seule flow cell de séquençage Nanopore (Michael *et al.* 2018). Panpan Zhang, en doctorat actuellement au sein de l'équipe a dévéloppé un outil de meta-assemblage SASAR (*Super-ASsembly from Assembly Reconciliation*) qui permet d'optimiser et réconcilier les assemblages *de novo* à partir de lectures Nanopore de différents logiciels. Ainsi, l'assemblage final obtenu pour l'espèce *Arabidospis thaliana* est de meilleure qualité que la référence actuelle *TAIR10* : les régions complexes, tels que les télomères et régions péricentromériques ont réussi à être mieux assembler (données non publiées).

Ces dernières technologies constituent une avancée significative dans le domaine de la génomique structurale en permettant la détection de variants structuraux à grande échelle, ce qui pourra ainsi déboucher sur l'étude de l'impact fonctionnel de ces derniers. Par contre, la gestion des gros volumes de données produites et l'analyse bio-informatique de celles-ci sont maintenant les nouveaux challenges dans le domaine de la génomique. Il est nécessaire de développer de nouveaux outils bio-informatiques dédiés à la détection de ces variations. Actuellement la majorité des outils de détection des insertions des ET au sein des génomes est basée sur les courtes lectures, ce qui sera abordé dans le chapitre suivant. Ainsi ces nouvelles technologies et leur lectures longues nécessitent le développement de nouvelles stratégies et concepts qui prennent en compte la possibilité d'un alignement d'une seule longue lecture sur une insertion entière d'élément transposable.

3.4 Méthodes de détection des néo-insertions au sein de génome reséquencé

Le reséquençage d'un génome consiste à séquencer un individu, principalement en utilisant la technologie Illumina pour étudier son polymorphisme en le comparant avec le génome de référence pour mettre en évidence les différences comme les mutations uniques (SNP), les petites insertions et délétions (INDEL). Il également possible de mettre en évidence les nouvelles insertions d'éléments transposables (ET) au sein de cet individu, mais cela reste plus complexe.

Au cours des études génomiques, les ET sont souvent ignorés ou masqués du fait de leur nature très répétée : leur analyse est donc difficile surtout en utilisant des lectures courtes Illumina (Goerner-Potvin *et al.* 2018). Les ET insérés au sein des cellules germinales qui sont donc partagés par toutes les cellules d'un individu seront plus facile à détecter que les insertions somatiques qui sont présentes uniquement dans une sous-population de cellules. Ainsi, les analyses de séquences montreront une grande différence de couverture entre ces deux types d'insertions. Pour cette raison, la majorité des logiciels d'analyse de transposition se limitent aux évènements de transposition ayant lieu dans la lignée germinale.

Il existe trois principales stratégies pour la détection des insertions d'éléments transposables au sein d'un génome reséquencé (en utilisant les courtes lectures Illumina) : les lectures discordantes, les lectures coupées et la différence de couverture. Lors de la détection de ces néo-insertions d'ET, il faut néanmoins faire la part entre le taux de faux positif (la spécificité) et le taux de faux négatif (la sensibilité). La spécificité peut être diminuée par la présence d'ET paralogues (copies d'ET d'une même famille) qui sera difficile à différencer lors de la détection. La sensibilité est plutôt liée à la complexité de la région d'insertion, comme le nombre de lectures au niveau de cette région ou de la mappabilité de celle-ci (notion qui sera abordé par la suite).

Au préalable de ces 3 stratégies, une étape d'alignement des lectures issues du génome de l'individu reséquencé contre le génome de référence est indispensable. Lors de cette étape, il est observé que tout au long du génome de référence, la couverture n'est pas homogène. Ainsi, la notion de "mappabilité" d'un génome doit être prise en compte (Rishishwar *et al.* 2017). Certaines régions du génome sont plus accessibles que d'autres : les régions fortement répétées, comme les régions péri-centromériques sont très faiblement détectées. Pour limiter l'impact de ce phénomène, il est donc préconisé d'obtenir une profondeur de séquençage d'au moins 20-30X.

Au cours de la première étape d'alignement, il est très difficile de distinguer les copies d'ET d'une même famille. Même si au cours de l'évolution ces éléments ont accumulé

des mutations, insertions ou délétions, du fait de la faible longueur des lectures, c'est un challenge de pouvoir déterminer sur quelle copie la lecture va s'aligner. Il est donc très difficile de distinguer les copies paralogues entre elles. Ainsi pour limiter ce problème, il est préconisé d'effectuer un séquençage apparié (séquençage des 2 côtés du fragment d'ADN au sein d'un cluster) pour avoir une meilleur sensibilité lors de l'alignement et avoir une meilleure idée de la localisation des lectures au niveau des régions répétées. Il faut toutefois noter que toutes ces limitations sont abolies avec l'utilisation de technologies longues lectures.

3.4.1 Les lectures discordantes (*paired-end mapping*)

Lors de l'alignement des lectures appariées contre le génome de référence, les polymorphismes d'insertion d'ET sont mis en évidence en ne gardant que les paires de lectures qui sont mal alignées (Figure 3.3 - A). Un mauvais alignement ou un alignement discordant correspond au fait qu'au sein de la même paire de lectures, les deux lectures

- ne s'alignent pas au même endroit
- sont trop éloignées l'une de l'autre, par rapport à la taille de l'insert attendue
- une seule des lectures s'aligne contre le génome de référence

Ainsi parmi ces lectures discordantes, si au sein d'une paire, une lecture correspond à un ET et l'autre correspond à une région génomique : ces lectures définissent une nouvelle insertion d'un ET au sein du génome reséquencé (Sabot *et al.* 2011 ; Elbaidouri *et al.* 2013 ; Kofler, 2016).

Cette technique a une forte sensibilité par contre il y a beaucoup de faux positifs, du fait de la forte similarité des copies d'une même famille d'ET entre elles : les copies paralogues. De plus, pour l'identification du point de la néo-insertion, on ne peut définir qu'une région d'insertion (la distance entre les 2 lectures appariées) et non la position exacte.

3.4.2 Les lectures coupées (*split-reads*)

Cette stratégie consiste comme précédemment à aligner les lectures contre le génome de référence. Mais cette fois-ci, les lectures sont considérés plutôt comme des lectures uniques que comme des lectures appariées. Au sein des lectures si une partie s'aligne sur le génome et l'autre s'aligne sur un ET, celles-ci seront mises en évidence. Ainsi la lecture tronquée correspond à la jonction entre l'insertion de l'ET et la séquence de référence (Figure 3.3 - B). Contrairement à la stratégie des lectures discordantes, les lectures coupées peuvent déterminer l'insertion des ET à la base près. Mais la couverture est moindre et cette méthode est très sensible à la taille des lectures.

Les différents outils de détection utilisent habituellement les 2 stratégies pour optimiser au maximum la détection des néo-insertions. Ils commencent par une stratégie de

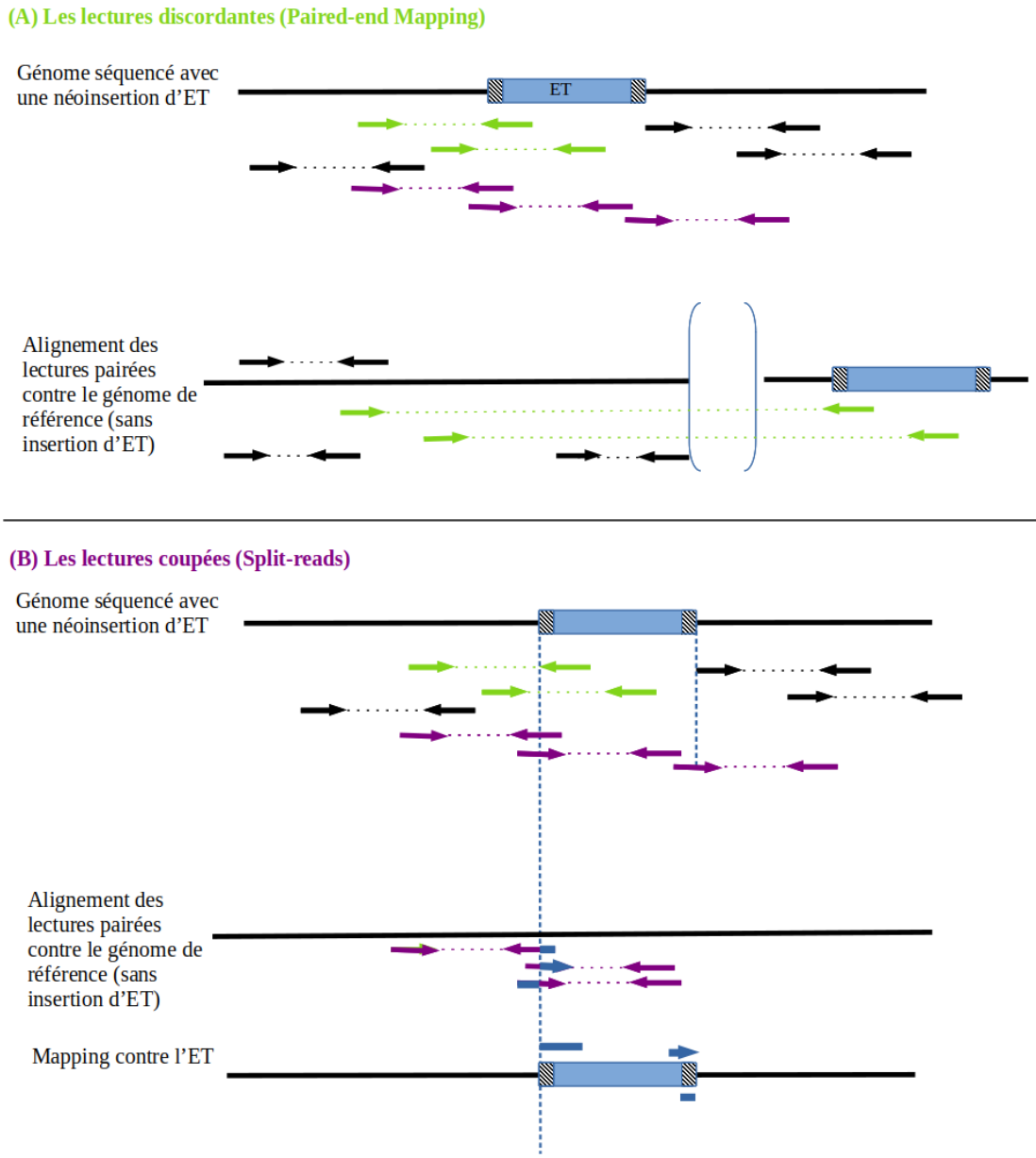


FIGURE 3.3 – **Stratégies de détection des néo-insertions au sein d'un génome reséquencé avec des lectures courtes**. **A** - La méthode des lectures discordantes, consiste à l'identification des paires de lectures dont une s'aligne sur le génome de référence et l'autre sur l'ET qui se situe à un autre endroit du génome. **B** - La méthode des lectures coupées, consiste à détecter les lectures dont une partie correspond au génome de référence et une autre à l'ET. Ces lectures correspondent au point d'insertion de l'ET, elles sont à la jonction entre le génome de référence et la néo-insertion d'un ET.

lectures discordantes, puis affinent l'identification du site d'insertion grâce aux lectures coupées (Hénaff *et al.* 2015 ; Goerner-Potvin *et al.* 2018 ; Baduel *et al.* 2020)

Basés sur ce concept, certains outils sont dédiés plus spécifiquement à l'analyse de grand jeux de données pour étudier la dynamique des insertions des ET au sein d'une population comme TEPID, T-Lex , PopoolationTE2, MELT (Kofler, 2016 ; Stuart *et al.* 2016 ; Gardner *et al.* 2017 ; Bogaerts-Márquez *et al.* 2019). En parallèle de ces nombreux développements de logiciel de détection d'ET, des plateformes répertoriant ces pipelines ont été mises en place tel que l'outil McClintock (Nelson *et al.* 2017).

3.4.3 La différence de couverture (DOC)

Si au sein du génome reséquencé, il y a eu de nouvelles insertions de certains éléments transposables déjà présents au sein de la référence, on s'attend à avoir un plus grand nombre de lectures correspondant à cette famille d'ET.

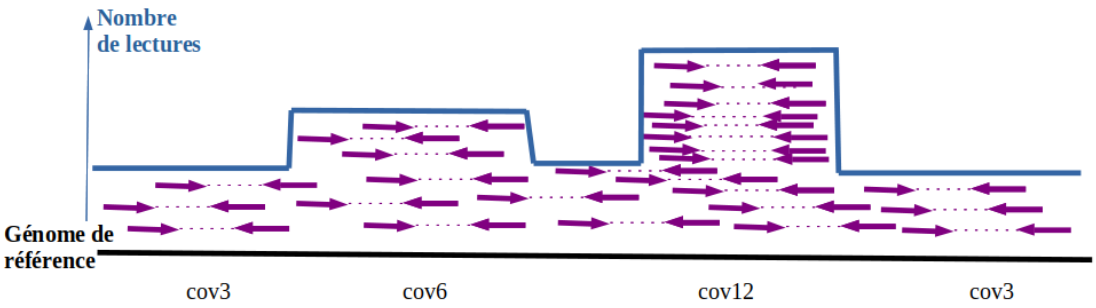


Figure 3.4 – **La différence de couverture (DOC)**. La profondeur de couverture (nombre de lectures par position) est calculée tout au long du génome. Ainsi, on pourra observer une différence de couverture au niveau des ET néo-insérés. Ici, on observe une couverture moyenne du génome équivalente à 3 lectures (3X). Par conséquent, une profondeur de couverture égale à 6X correspond à 2 copies d'un ET, tandis qu'une profondeur de couverture de 12X correspond à 3 copies de l'ET.

La méthode du DOC est basée sur ce postulat. La couverture au niveau de chaque ET est calculée au sein de la référence et de l'individu reséquencé. Ainsi s'il y a une augmentation significative de celle-ci au sein du nouveau génome, on peut en déduire qu'il y a eu des néo-insertions de la famille d'ET en question au sein du génome de l'individu. Pour estimer le nombre de copies, il sera nécessaire de normaliser la couverture au niveau de l'ET par la couverture de gènes uniques.

L'avantage de cette méthode est sa rapidité : les lectures sont alignées directement sur le génome et ensuite un comptage du nombre de lectures par ET est effectué. Néanmoins, cette méthode ne met en évidence que la nature de l'ET qui s'est amplifié mais il n'est pas possible de connaître la localisation de ces néo-insertions. C'est une méthode plus exploratoire comparée aux deux autres décrites ci-dessus, et elle ne peut s'appliquer qu'aux ETs de classe I (qui transposent via un mode copier/coller).

3.4.4 Les longues lectures

Grâce au séquençage « longues lectures », l'alignement (la mappabilité) contre le génome de référence est considérablement amélioré. De ce fait la sensibilité liée à la détection des insertions d'ET est meilleure. Contrairement aux courtes lectures, qui ont une faible sensibilité et un fort taux de faux positif, il est possible maintenant d'avoir au sein d'une longue lecture une insertion d'ET en entier et ses deux séquences flanquantes, permettant de définir précisément le point d'insertion (Figure 3.5).

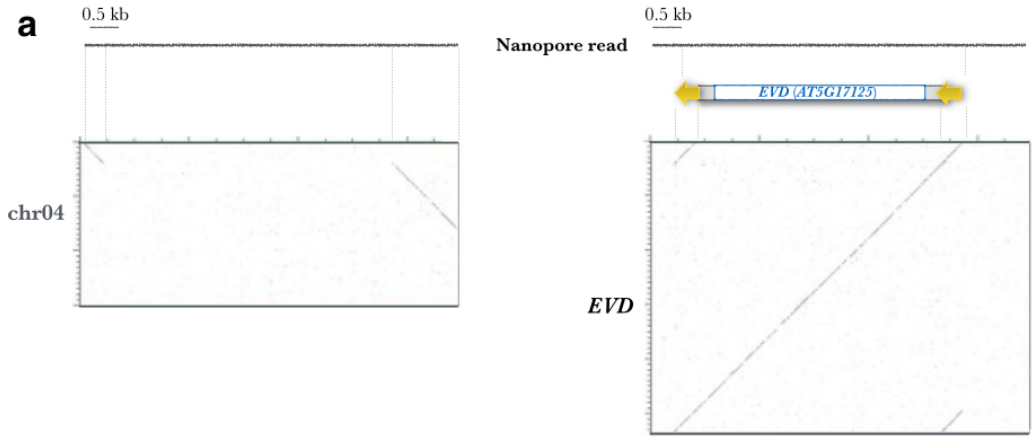


FIGURE 3.5 – **Caractérisation d'une néo-insertion d'un rétrotransposon à LTR au sein d'une lecture Nanopore**. Figure 1 issue de DEBLADIS *et al.* 2017. Alignement de la lecture Nanopore de 7kb contre la région d'insertion au niveau du chromosome 4 de *A.thaliana* (à gauche) et contre la séquence du rétrotransposon *EVD* (à droite). On observe au sein de la même lecture le rétrotransposon *EVD* en entier et les régions flanquantes de la néo-insertion.

De plus les longues lectures ont le potentiel de diminuer le taux de faux positifs du fait de la grande taille de leur alignement contre le génome de référence. Finalement, avec cette nouvelle technologie, il est désormais possible de détecter des événements d'insertions plus complexes, comme les insertions d'ET nichées, les variations structurales de grandes longueurs telles que les inversions ou duplications génomiques. Ces dernières années de nombreux logiciels de détection de variations structurales ont été développés tels que NanoSV, Sniffles, Dygsu (CRETU STANCU *et al.* 2017 ; SEDLAZECK *et al.* 2018 ; CLEAL *et al.* 2021) qui montrent une nette amélioration de la sensibilité de la détection des insertions d'ET avec les longues lectures Nanopore, comparée aux courtes lectures Illumina (A. ZHOU *et al.* 2019).

CHAPITRE 4

Le riz cultivé *Oryza sativa*

Au cours de ma thèse, mon modèle d'étude a été le riz cultivé *Oryza sativa*. Il existe actuellement deux espèces de riz cultivé : le riz africain *Oryza glaberrima*, domestiqué il y a 3 000 ans et le riz asiatique *Oryza sativa*, domestiqué il y a 10 000 ans (Purugganan, 2019). Ces 2 espèces ont une histoire évolutive très différente. Dans la suite de ma thèse je me concentrerais sur le riz asiatique (*Oryza sativa*).

Harlan disait « Quiconque veut étudier la diversité d'une plante cultivée doit comprendre son histoire », c'est pourquoi dans un premier temps je vais aborder l'histoire de la domestication du riz asiatique et dans un second temps l'activité des éléments transposables au sein du génome de *Oryza sativa*.

4.1 Domestication du riz (*Indica* et *Japonica*)

La domestication des plantes est un des phénomènes les plus important lié à l'histoire de l'Homme. Après la dernière ère glaciaire, il y a 12,000 ans, au début du Holocène, il y a eu un réchauffement du climat qui a coincidé avec l'apparition de l'agriculture. La domestication des plantes et des animaux, à l'origine de cette mutation profonde des sociétés humaines a consisté en l'acquisition d'un ensemble de caractéristiques morpho-physiologiques adaptées à leur culture (ou leur élevage) appelé syndrome de domestication. La domestication est un processus qui demande des années avant que les caractères phénotypiques soient fixés au sein d'une population (Purugganan, 2019).

Le riz asiatique est la première céréale vivrière au monde. Il est cultivé sur les cinq continents dans des écosystèmes très divers. Il existe sept types variétaux de riz asiatique, divisés en 5 grand groupes : deux principaux *Japonica tempéré et tropical et Indica*, ainsi que deux autres *Aus/Boro, Basmati/sadri* (Figure 4.1).

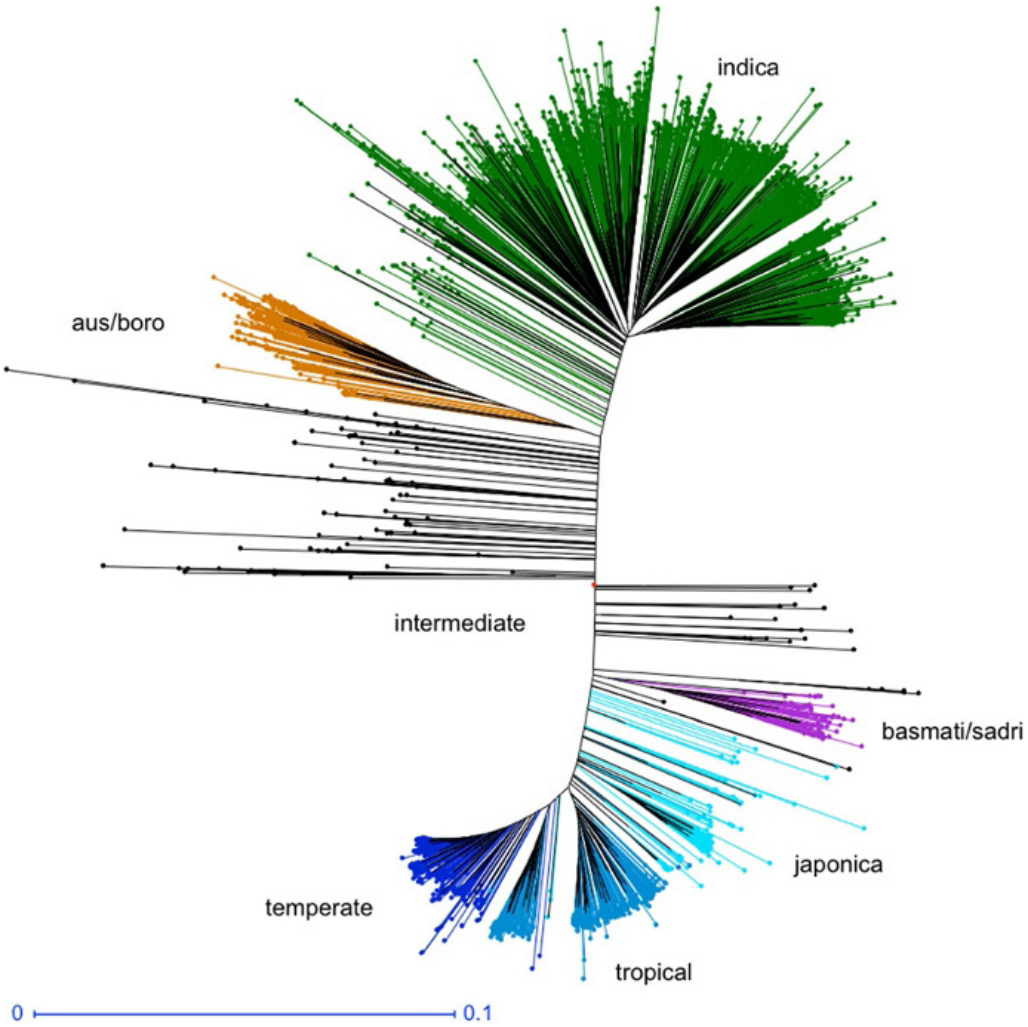


FIGURE 4.1 – **Arbre phénétique des 3000 génomes de riz cultivé** (THE 3,000 RICE GENOMES PROJECT, 2014). Classification des 3000 accessions de riz cultivés regoupés en 7 groupes variétaux distincts réalisée à partir de 5 sets aléatoires de 200,000 SNPs, parmis les 18,9 millions SNP détectés au total.

Le riz asiatique *Oryza sativa* est l'aliment de base de la moitié de la population mondiale, par conséquent c'est une des céréales les plus importantes pour l'Homme. Le riz asiatique a été domestiqué à partir de l'espèce sauvage *Oryza rufipogon* et *Oryza nivara*. Les différences entre le riz sauvage *O. rufipogon* et le riz cultivé *O. sativa* se reflètent aux niveaux morphologiques et physiologiques comme mentionné plus haut (Figure 4.2).

Chez cette espèce, le syndrome de domestication est similaire à celui d'autres céréales comme l'absence d'égrenage à maturité permis par la disparition de la couche d'abscission fonctionnelle à la base du grain et une modification de l'architecture de la panicule, l'absence de dormance du grain, l'homogénéité de la date de floraison (HUANG *et al.* 2012 ; ISHIKAWA *et al.* 2020).

Le déterminisme génétique de certains de ces caractères a été étudié et plusieurs

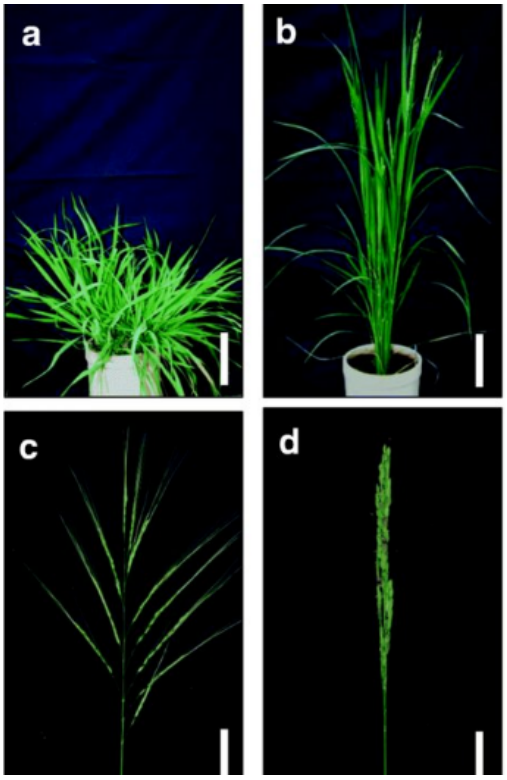


FIGURE 4.2 – **Comparaison de la morphologie des plantes entre le riz sauvage *Oryza rufipogon*, et le riz cultivé *O. sativa ss Japonica Nipponbare***. D'après ISHIKAWA *et al.* 2020. Le panel de gauche correspond au riz sauvage *O.rufipogon*, celui de droite correspond au riz cultivé *O.sativa*. **(a, b)** Architecture végétale. **(c, d)** Forme de la panicule.

gènes majeurs ont été caractérisés, comme le gène *sh4* impliqué dans le non-égrenage (non-shattering) des grains (C. LI, 2006) ou *sd1* pour la nanisme du riz (FERRERO-SERRANO *et al.* 2019).

L'origine et l'histoire de la domestication du riz sauvage est un sujet très débattu au sein de la communauté, du fait que l'allèle du gène de non-égrenage des formes cultivées – *sh4* – , gène clé de la domestication est le même au sein des deux variétés principales *Japonica et Indica*. Ce qui pourrait suggérer une domestication unique. Pourtant en analysant génétiquement ces deux variétés, on observe une réelle différenciation : ce qui pourrait suggérer au contraire plutôt deux évènements de domestications distincts (K. ZHAO *et al.* 2011 ; HUANG *et al.* 2012)

Dans l'hypothèse d'une domestication unique, le riz *Japonica* a été domestiqué il y a environ 9,000 ans à partir du riz sauvage *O.rufipogon* dans la vallée de *Yantgtze*. De l'autre côté de l'Himalaya, dans la vallée du Gange, il y a environ 8,000 ans à partir d'un autre riz sauvage *O.nivara*, une sous espèce "proto-*indica*" est apparue. Par migration des populations via la route de la soie, le riz domestiqué *Japonica* s'est hybridé avec le riz *proto-indica*, créant ainsi le type *Indica* actuel avec les gènes de domestication acquis par introgression (Huang *et al.* 2012 ; Gross *et al.* 2014 ; Q. Zhao *et al.* 2018 ; Purugganan, 2019) (Figure 4.3).

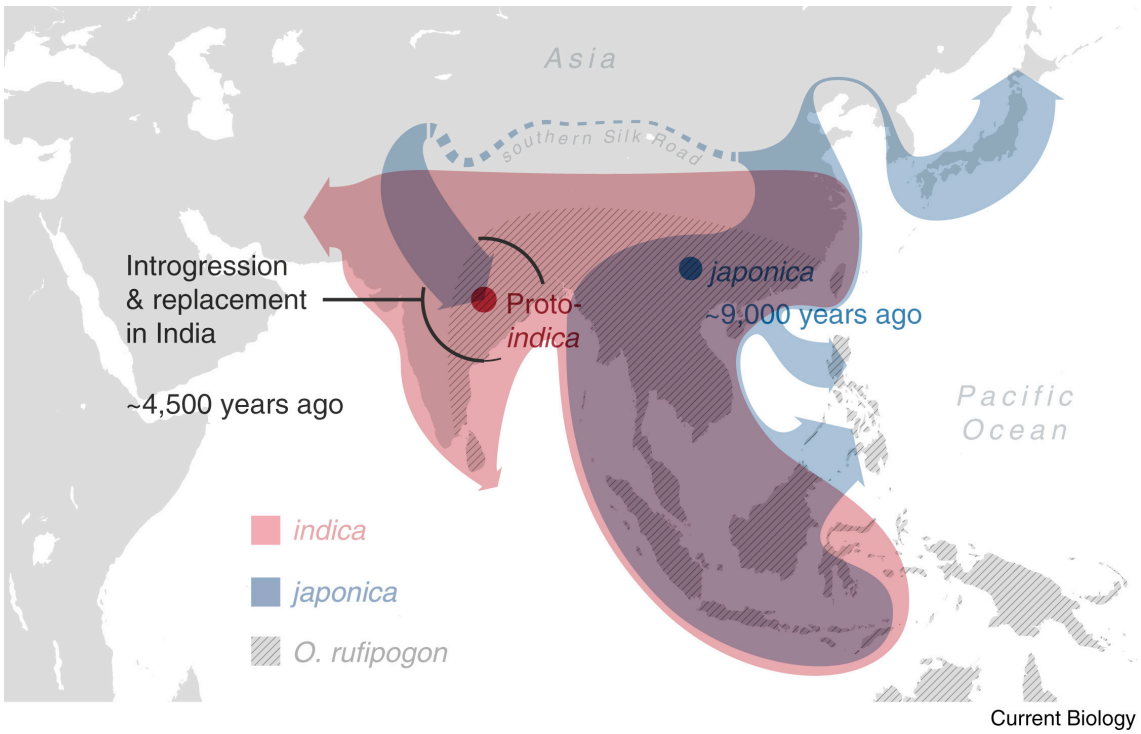


Figure 4.3 – **L'hypothèse de la domestication unique et l'origine du riz *Indica*.** D'après Purugganan, 2019. Dans ce scénario de domestication unique, le riz *Japonica* a été domestiqué il y a ~9,000 ans à partir de *O. rufipogon* et un *proto-indica* a peut-être débuté il y a ~8,000 ans à partir de *O. nivara*. On pense que le riz *Japonica*, se déplaçant via la route de la soie, est entré dans le Nord-Ouest de l'Inde et s'est hybridé avec le *proto-indica* non domestiqué, fournissant des allèles de domestication et conduisant à l'*Indica*.

Dans l'hypothèse de domestications multiples, les deux sous-espèces *Japonica et Indica* proviennent du même ancêtre commun *Oryza rufipogon*. Mais l'espèce s'est scindée en deux pools génétiques entre 500,000 et 800,000 ans de part et d'autre de l'Himalaya, le type *Indica* au Sud et le type *Japonica* au Nord (Figure 4.4) (Civáň *et al.* 2015 ; Wang *et al.* 2018 ; Carpentier *et al.* 2019).

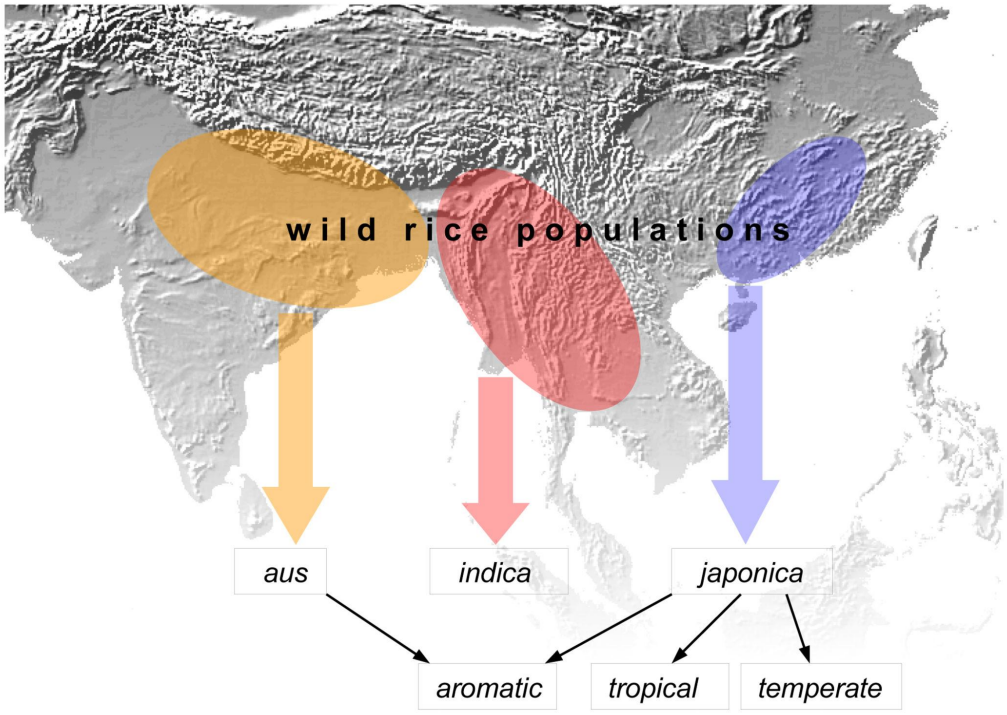


Figure 4.4 – **L'hypothèse de la domestication multiple du riz** D'après Civáň *et al.* 2015. Schéma des origines du riz domestiqué dérivé de l'analyse phylogéographique de 31 variétés de riz. Des domestications séparées ont donné les variétés *Indica, Japonica et Aus*, ces domestications ayant eu lieu dans différentes parties du sud-est et du sud de l'Asie, avec une hybridation ultérieure entre *Japonica et Aus* donnant le riz aromatique, et les versions tempérées et tropicales de *Japonica* évoluant comme des adaptations ultérieures.

Avec la dissémination globale de la culture du riz, les deux types *Japonica et Indica* sont souvent cultivés dans les mêmes régions, mais chaque variété garde ses caractéristiques propres comme la longueur et la largeur de grains spécifiques à chaque variété (Gutaker *et al.* 2020). L'hybridation ente les deux types est possible mais limitée, ce qui explique leur différenciation génétique. Néanmoins, des traces d'introgression ont été identifiés chez de nombreux cultivars de riz, rendant difficile la caractérisation de types « purs » *Indica ou Japonica*. Ainsi, on peut penser que l'unique allèle de non-égrenage (*sh4*) est le résultat d'une introgression entre ces deux variétés. Ceci pourrait ainsi réconcilier cette seconde hypothèse avec les observations génétiques. On peut considérer ce phénomène d'introgression et d'échange de matériel génétique comme une source de nouveauté permettant une meilleure adaptation aux niches écologiques.

4.2 Dynamique des éléments transposables chez le genre *O.sativa*

Les éléments transposables, toutes classes confondues, représentent environ 40% de la taille du génome du riz de référence *O.sativa ssp japonica, Nipponbare* (Jiang *et al.* 2013).

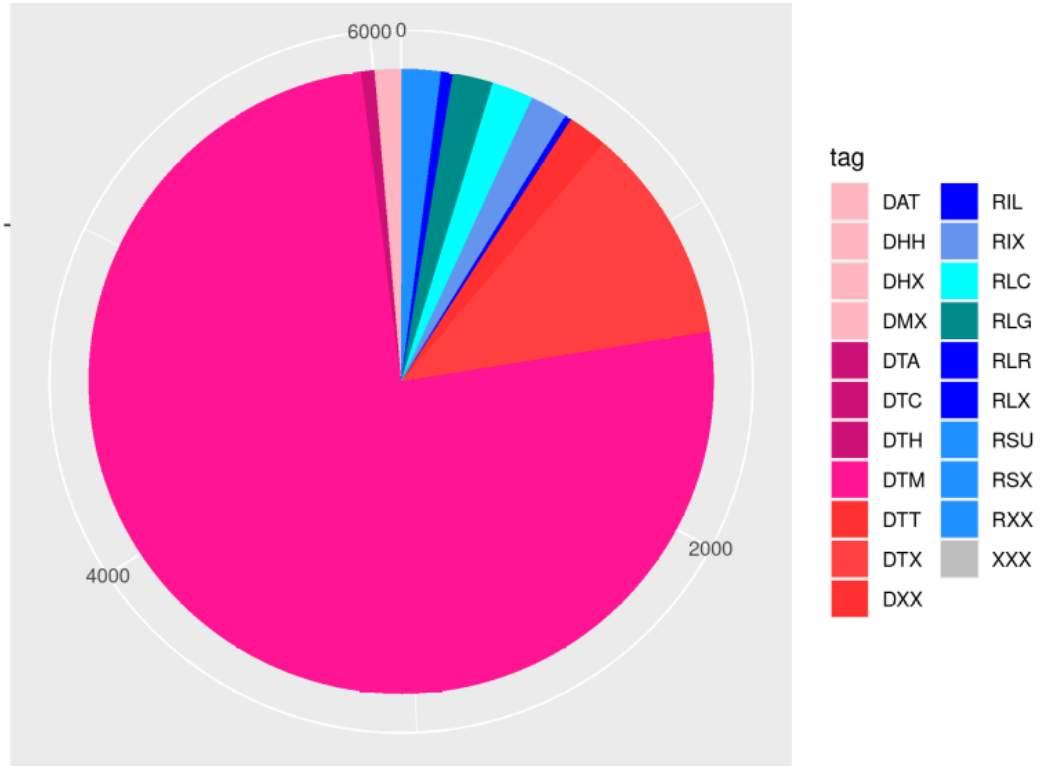


Figure 4.5 – **Distribution des différentes catégories d'éléments transposables chez le riz *O. sativa ss japonica*.** En dégradé de bleu sont représentés les rétrotransposons (ET classe I) et en dégradé de rose, les transposons (ET classe II). Le tag de chaque famille d'ET provient de l'annotation de Wicker *et al.* 2007.

Les transposons et plus précisément les MITE (#DTM de la Figure 4.5) sont la famille qui est la plus abondante en terme de copies. Mais les rétrotransposons (notamment les rétrotransposons à LTR) représentent quand à eux le plus important nombre de paires de bases (25% du génome), du fait de leurs grandes longueurs. Avec la forte dynamique des ET au sein du génome du riz, le turn-over de ces ET au sein de l'espèce est très rapide (El Baidouri *et al.* 2013 ; Stein *et al.* 2018). Leur élimination par recombinaison homologue (Figure 2.5) engendre une grande abondance de soloLTR, qui sont présents au sein des chromosomes du riz de référence *Nipponbare Japonica*. Grâce aux différentes ressources génomiques disponibles (International Rice Genome Sequencing Project), une base de données expertisée des éléments transposables a été développée (Copetti *et al.* 2015). Cette base de données, la "RiTE-db", contient les séquences consensus de tous les ET

entiers présents chez le genre *Oryza*. De plus, avec l'analyse des 12 génomes du genre *Oryza* (Stein *et al.* 2018), il est également possible de connaître l'histoire évolutive de ces éléments, c'est à dire si ces copies sont partagées par tous les espèces du genre ou si celles-ci sont spécifiques à seulement certaines variétés. Cette base de données est un catalogue de tous les éléments transposables au sein du genre *Oryza*.

Au sein de l'équipe, à partir de la "RiTE-db", une base de données expertisée des éléments transposables du génome de référence *Nipponbare, O.sativa ss Japonica* a été également développé. L'annotation des ET a été expertisée manuellement et une séquence consensus a été conservée pour chaque famille. On obtient ainsi un total de 6087 séquences d'ET dont la majorité sont des MITE (Figure 4.5).

En parallèle de ces données génomiques, certains ET actifs avaient été mis en évidence à l'aide de méthodes plus classiques, comme le séquençage des cDNA de la transcriptase inverse (RT) pour *Tos17* (Hirochika *et al.* 1995), l'hybridation sur puces pour *Lullaby* (Picault *et al.* 2009), les southern blot de *Ping/Mping* (L. Lu *et al.* 2017) ou la mise en évidence de cercles extra-chromosomiques du rétrotransposon *PopRice* au sein de l'endosperme (Lanciano *et al.* 2017). A noter que la réactivation du rétrotransposon *Tos17* en culture cellulaire de cals de riz a été exploitée pour créer des banques de mutants chez le riz (Hirochika *et al.* 2004).

CHAPITRE 5

Objectifs de la thèse

Nous avons vu dans l'introduction l'impact et l'importance des éléments transposables au sein des génomes de plantes au niveau macro et micro-évolutif.

Vu son fort intérêt socio-économique, de nombreuses études ont été menées sur le riz pour analyser son histoire et sa diversité. De plus, de nombreuses ressources génomiques de haute qualité (séquences et annotations) sont disponibles, incluant des séquences génomiques pour 3000 variétés de riz cultivé (The 3,000 rice genomes project, 2014). Grâce à sa taille de génome relativement faible (comparé au maïs dont le génome fait 2,1Gb) mais suffisamment grande pour héberger une part significative d'ETs (40%), le riz est un excellent modèle pour l'étude de l'impact génmique de la transposition. Ma thèse a été consacrée à l'étude de l'activité transpositionnelle chez l'espèce *Oryza sativa* à l'ère des nouvelles technologies de séquençage.

Dans un premier temps, il a été nécessaire de décrire le paysage transpositionnel et surtout les éléments actifs ayant contribué à la genèse de la diversité génomique au sein de la population de 3000 génomes de riz cultivés. Ceci sera décrit au sein du chapitre 6 avec la présentation de TRACKPOSON, pipeline de détection de polymorphismes d'insertions d'éléments transposables (TIP) au sein de grands jeux de données que j'ai développé au cours de ma thèse.

Dans un second temps, nous nous sommes intéressés à l'impact de ces variations structurales. Pour répondre à cela, une analyse d'association entre les insertions d'ET et un phénotype (TE-GWAS) au sein des 3000 génomes de riz cultivé a été conduite par un post-doctorant de l'équipe. Une association significative a été analysée plus particulièrement : c'est le cas de l'insertion du rétrotransposon *rn215-125* au niveau du chromosome 4 et son impact fonctionnel sur la largeur du grain au sein des variétés de riz *Indica*. Mon

implication dans ce travail a été de fournir toutes les ressources génomiques nécessaires pour l'étude, et de mettre en œuvre une stratégie basée sur l'utilsation des données Illumina et Nanopore pour la validation *in silico* de ces résultats et la caractérisation de cette région d'intérêt. Ceci fait ainsi l'objet du chapitre 7 de ma thèse.

Résultats

CHAPITRE 6

Activité transpositionnelle au sein d'une population de 3000 génomes de riz cultivés, *Oryza sativa*

6.1 Introduction

En 2014, un projet de séquençage massif de 3000 génomes de riz cultivé a été entrepris par un consortium international regroupant l'Institut International de Recherche sur le riz (IRRI) et l'académie d'agriculture de Chine (The 3,000 rice genomes project, 2014). Le riz asiatique, *Oryza sativa*, du fait de son grand intérêt socio-énonomique en tant que pilier de la sécurité alimentaire du globe, a été largement étudié. Ainsi depuis la publication de la première séquence de génome de riz (International Rice Genome Sequencing Project *et al.* 2005), de nombreuses ressources génomiques ont été générées, telles que plusieurs génomes assemblés de haute qualité (Y. Zhou *et al.* 2020).

Le génome du riz est constitué à plus de 40% de séquences répétées dont la majorité (en termes de paires de bases) sont des éléments de classe I, les rétrotransposons à LTR. Grâce à ce grand jeu de données issues de la 2ème génération de technologie de séquençage, il est donc possible d'étudier la dynamique des éléments transposables au niveau d'une population.

Ce séquençage est un séquençage Illumina effectué avec de séquences courtes, seulement 75 paires de bases. Ainsi le défi de cette analyse est donc de réussir à détecter les insertions d'ET au sein des 3000 génomes de riz simultanément, en ayant le minimum de faux positifs. Comme mentionné dans l'introduction, il est très difficile de détecter les néo-insertions d'ET à partir de ce type de séquençage. Les logiciels existant de détections d'ET au sein d'un génome reséquencé se basent sur l'alignement des lectures issues du génome reséquencé contre le génome de référence. S'en suit alors l'identification des

néo-insertions d'ET à partir des lectures discordantes et la présence de lectures coupées (*split-read*).

Ainsi grâce à ces outils, il est possible de mettre en évidence toutes les néo-insertions d'ET présentes au sein du nouveau génome mais cela peut prendre quelques heures pour un génome de la taille du riz (environ 390Mb) et nécessite une capacité de stockage conséquence (>10Go par génome x 3000 pour les fichiers d'alignement bam). De plus, le taux de faux positifs est assez important : il est très difficile de distinguer des insertions d'ET qui sont récentes et donc qui ont par conséquent des taux de pourcentage d'identité très proches.

Il était donc inenvisageable d'utiliser de tels outils sur un grand jeux de données tel que les 3,000 génomes de riz : c'est pour cela que j'ai développé mon pipeline, TRACKPOSON. Contrairement aux outils standards, TRACKPOSON se base sur un alignement contre une séquence consensus de référence d'une famille d'ET, et non contre le génome entier de référence. Ainsi seulement une faible partie des lectures est gardée, et donc cela permet de diminuer considérablement le temps de calcul, passant de quelques heures à quelques minutes. Cependant, TRACKPOSON permet de détecter toutes les insertions d'une seule famille d'ET, mais au sein des 3,000 génomes de riz. Ainsi, il est nécessaire de réitérer la détection par TRACKPOSON pour chaque famille d'ET souhaitée. Grâce aux données longues lectures Nanopore, il a été possible de valider les résultats trouvés par TRACKPOSON et d'estimer la sensibilité et la spécificité de celui-ci qui est de 81% et 94,5% respectivement. Comme espéré, on a peu de faux-positif. Par contre, dû aux problèmes de mappabilité des courtes lectures sur le génome, la sensiblité est nettement diminuée : on a ainsi plus de faux-négatif.

Ces résultats ont fait l'objet d'une publication dans Nature Communications, début 2019.

6.2 Résultats

Dans cet article, à l'aide du pipeline développé TRACKPOSON, nous avons identifié les insertions de 32 familles d'ET de classe I au sein des 3,000 génomes de riz cultivés (N=53,262). Ainsi, cela nous a permis d'observer un fort polymorphisme d'insertion des ET et que la dynamique de ces derniers au sein de cette population était famille dépendant, distinguant ainsi les rétrotransposons de la famille des *Gypsy* et *Copia*.

A noter que la majorité des insertions sont présentes en faible fréquence au sein de la population des 3,000 génomes de riz : les insertions sont spécifiques à une ou deux variétés seulement. Ainsi, peut-on penser qu'il existe une activité transpositionnelle très récente, au champ, *in agro*. Alternativement, cette faible fréquence peut être le résultat d'une élimination efficace des insertions par un processus de sélection qui élimine les individus qui les portent.

Au vu de ce résultat, une analyse de génomique d'association (GWAS) a été conduite pour déterminer s'il y avait un génotype qui permettait d'expliquer le nombre d'insertions de chaque famille d'ET. Finalement, nous n'avons identifié aucune cause génétique à l'activation de la transposition, comme proposé pour une population d'*Arabidopsis* (Quadrana *et al.* 2016). Mais cette analyse d'association a permis d'identifier la copie "maître" de l'ET, copie nécessaire à l'activation de la transposition de la famille. Cette activation pourrait être une activation *via* un stimulus externe, mais cela reste très spéculatif.

Nous nous sommes également intéressés à l'histoire évolutive des insertions détectées par TRACKPOSON. Grâce à la structure particulière des rétrotransposons à LTR et au mécanisme de leur cycle transpositionnel, il est possible de dater l'insertion des copies. En comparant le pourcentage d'identité entre les 2 LTR et avec une horloge moléculaire spécifique aux ET (1.3x10-8 substitution/site/an) (SanMiguel *et al.* 1998), il est possible de traduire ce taux de divergence en date d'insertion. Ainsi avec la datation des toutes les insertions détectées, il a été possible de mettre en évidence plusieurs événements de domestication indépendantes du riz cultivé asiatique (Civáň *et al.* 2015), qui valide la robustesse de notre approche et de mon outil TRACKPOSON.

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-018-07974-5 OPEN

Retrotranspositional landscape of Asian rice revealed by 3000 genomes

Marie-Christine Carpentier[1], Ernandes Manfroi[2], Fu-Jin Wei[3,4], Hshin-Ping Wu[3], Eric Lasserre[1], Christel Llauro[1], Emilie Debladis[1], Roland Akakpo[1], Yue-Ie Hsing[3] & Olivier Panaud[1,5]

The recent release of genomic sequences for 3000 rice varieties provides access to the genetic diversity at species level for this crop. We take advantage of this resource to unravel some features of the retrotranspositional landscape of rice. We develop software TRACK-POSON specifically for the detection of transposable elements insertion polymorphisms (TIPs) from large datasets. We apply this tool to 32 families of retrotransposons and identify more than 50,000 TIPs in the 3000 rice genomes. Most polymorphisms are found at very low frequency, suggesting that they may have occurred recently in agro. A genome-wide association study shows that these activations in rice may be triggered by external stimuli, rather than by the alteration of genetic factors involved in transposable element silencing pathways. Finally, the TIPs dataset is used to trace the origin of rice domestication. Our results suggest that rice originated from three distinct domestication events.

[1] Laboratoire Génome et Développement des Plantes, UMR CNRS/UPVD 5096, Université de Perpignan Via Domitia, 52 Avenue Paul Alduy., 66860 Perpignan Cedex, France. [2] Faculdade de Agronomia, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS 90040-060, Brazil. [3] Institute of Plant and Microbial Biology, Academia Sinica, 128, Section 2, Yien-chu-yuan Road, Nankang 115 Taipei, Taiwan. [4] Department of Forest Molecular Genetics and Biotechnology, Forestry and Forest Products Research Institute, 1 Matsunosato, Tsukuba 305-8687 Ibaraki, Japan. [5] Institut Universitaire de France, 1 rue Descartes, 75231 Paris Cedex 05, France. Correspondence and requests for materials should be addressed to O.P. (email: panaud@univ-perp.fr)

One of the major discoveries of the last decades of genomic research is that genes are outnumbered by transposable elements (TEs) in most eukaryotic genomes[1,2]. In flowering plants, the distribution of genome size among 7500 angiosperm species indeed shows that 99% of the species have a genome larger than 200 Mbp/1 C (i.e., at least twice the size of their gene space, ~100 Mbp on average)[3].

TEs are very diverse both in terms of structure and modes of transposition[4], but they all can be mobilized and amplified within the genomes, which explains their propensity to densely populate eukaryotic chromosomes[5]. TEs have long been considered useless or even deleterious, but these views have recently been challenged. Experimental evidence have shown that they could be beneficial to organisms and have, in some cases, been domesticated by their host genome in various eukaryotic lineages to create biological novelty[6,7]. With the advent of new sequencing technologies and the availability of genomic resources for many organisms (and even populations), the molecular mechanisms involved in this process have begun to be unraveled and show that both the regulatory sequences and the proteins encoded by TEs can be exapted[8,9]. In a shorter time scale, TEs have been shown to play a role in adaptation in natural populations[10] and in crops[11,12].

Several aspects of TE dynamics at the population level remain unclear. On the one hand, comparative genomic studies in various lineages show that TEs contribute significantly to genome diversification, suggesting that TE insertion polymorphisms (TIPs) could be frequent enough in natural populations to serve as a source of adaptive variation[13]. On the other hand, recent advances in epigenetics clearly show that transposition is strictly controlled in planta by several transcriptional and post-transcriptional pathways[14], which raises the questions of the conditions of transposition activation *in natura* and of the actual dynamics of transposition in natural populations.

The exhaustive characterization of TIPs in a given gene pool requires genomic data for a comprehensive sample of individuals and at least one good-quality reference genome sequence from which TEs have been well characterized. These resources are available for a few model species. Rice (*Oryza sativa*), is well suited for such study with one high quality, physical map-based genome assembly[15], from which TEs have been annotated and curated[16,17]. More importantly, with a genome size of 430 Mbp, rice is close to the most representative plant genome in terms of size (and thus TE content), as evidenced by the modal value of the distribution of 7500 angiosperm lineages, which is ~500–600 Mbp[3]. Rice genome indeed harbors hundreds of TE familes of both class I and class II types[16,18,19]. Long Terminal Repeats (LTR)-retrotransposons constitute the largest part of rice mobilome in terms of percentage of genome they represent[20]. In a previous study, we showed that retrotranspositional activity in *Oryza* genus was at the origin of significant variations in genome size among diploid species, suggesting that this particular type of TEs plays a major role in transposition-driven genome dynamics in this lineage[21]. Rice genome harbors ~300 families of LTR-retrotransposons, belonging to either *Gypsy* or *Copia* superfamilies[16]. These families differ in copy number, ranging from singletons to large families (e.g., over 100 complete copies for *Hopi* and *Houba* families). In addition to the high-quality map-based genome assembly of the Nipponbare variety, three other high-quality PacBio-based assemblies are publicly available[22] and the raw Illumina-based genome sequences of 3000 accessions have been released[23]. This offers the opportunity to study genome dynamics at intra-specific level and at an unprecedented scale, although this requires the development of bioinformatic tools that are suitable for the analysis of very large datasets.

Conceptually, the detection of TIPs from populations based on Illumina data is straightforward, although prone to high false discovery rate (FDR) for large genomes (several 100 Mbp) due to the small reads size[24]. The methods commonly used are paired-end mapping (PEM) and split reads. For the first method, all paired reads are mapped onto a reference genome and discordant pairs (i.e., both reads from the same "amplicon" mapping at different locations) should correspond to structural variants and declared as TIPs if one end matches a known TE. The second method consists in identifying reads for which one part maps onto the reference genome sequence while the other matches a TE[25–27]. The main limitation of these approaches (besides FDR) is that the mapping step of all paired reads onto a reference genome is computationally intensive.

Therefore, we developed TRACKPOSON software for the efficient detection of TIPs of known TEs in large datasets. Using this new tool, we successfully characterized TIPs among 3000 rice genomes for 32 retrotransposons families. We chose this sample as being representative of the 300 families found in the rice genome in terms of superfamilies, copy number and transpositional activity[16,28–31], thus unraveling the retrotranspositional landscape of a plant genome at the species level. Using this data, we tentatively looked for genetic factors that may trigger transposition in agro in rice and conclude that for most TE families, such activation likely originated from external stimuli, rather than from a mutation within a genetic factor involved in the control of transposition.

The origin of rice domestication has been strongly debated in the past decades. Many studies of the genetic diversity of Asian rice clearly show a diphyletic origin of the crop, thus suggesting two independent domestications[32,33]. However, the cloning and characterization of some key domestication loci, such as shattering locus *Sh4*[34], showed that all rice types harbor the same domesticated allele, which suggests a single domestication event and that the differenciation of rice into the two major *Indica* and *Japonica* types may result from introgressions between this first domesticate and wild relatives in southern Himalayas[35,36]. Alternatively, one could hypothesize that there were indeed two distinct domestications, followed by some introgressions between the two domesticated gene pools, as the result of human selection for the best cultivated phenotype[37]. Finally, the domesticated alleles could have been present in the populations of the wild progenitor of rice long before the split of the two gene pools that gave rise to the domesticated forms and been selected for at least two times independently[38]. Here, exploiting TIPs as genomic paleontological records, we show that *Indica, Japonica*, and *Aus/Boro* groups originate from wild relatives that diverged long before upper neolithic, which supports the hypothesis that they originated from multiple domestication events.

Results

TRACKPOSON is a tool for the detection of TIPs in gene pools. The new strategy we developed consists in first mapping all reads of a given accession onto each TE family represented by a single consensus sequence (as opposed to mapping them onto the complete genome, as in the case of conventional PEM procedures) and then mapping the unmapped paired reads onto the rice reference genome, split into 10 kb windows (Fig. 1). In this regard, TRACKPOSON is not designed for a full characterization of all structural variations, but rather as a fast tool to unravel the transpositional activity of known TE families from very large genomic dataset. We first tested TRACKPOSON on a rice mutant for which TIPs had been previously characterized by using our previous PEM-based software[39] and wet-lab validated by PCR amplification and sequencing (Supplementary Figure 1). All TE

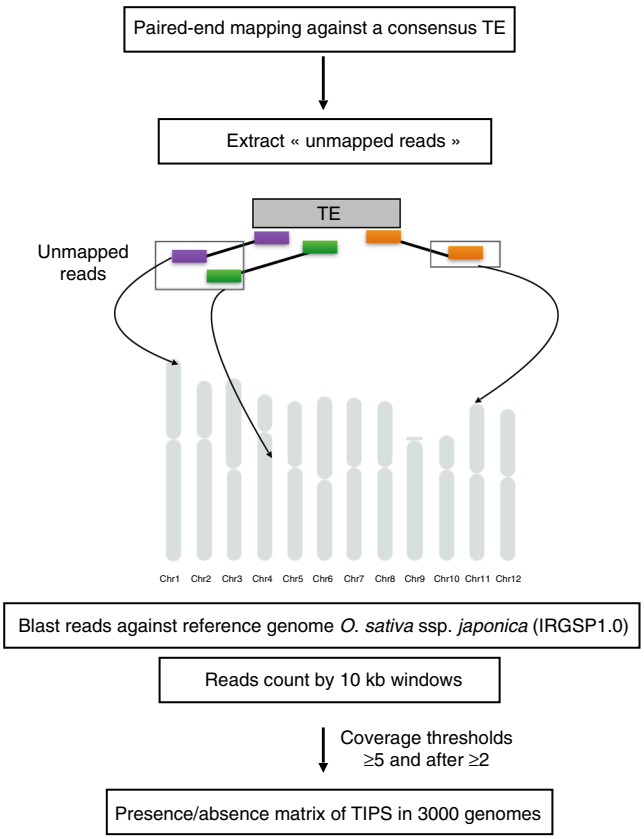


Fig. 1 TRACKPOSON method. Schematic representation of the pipeline

insertions were detected, which suggests that TRACKPOSON is a robust detection procedure with high sensitivity, given that sufficient genome coverage has been achieved (the mutant was sequenced at 30× depth). We should emphasize that the genome coverage of the 3000 rice accessions is 11.6× on average, with some as low as 7× (the first quartile being <9.4×). We therefore had to adapt our method to improve its sensitivity (low false-negative rate). First, only events supported by at least five paired reads were declared as TIPs, thereby opening the corresponding 10 kb window for the rice variety for which the polymorphism was identified. A second step consisting of another complete scan of the 3000 rice varieties, albeit with a detection threshold of 2 paired reads, was performed, and only for the 10 kb windows previously opened (supported by 5 paired reads). This considerably improved the sensitivity of detection (Supplementary Figure 2), with little risk that it increased the false-positive detection rate because the probability that two chimeric amplicons match the exact same 10 kb window as the one previously detected in another accession should be very low.

In order to check for the performance of this method, we resequenced one rice accession (Mutha Samba::IRGC 49924–2) that was originally sequenced in the 3000 genome collection at a 8.3× depth, i.e., among the lowest coverage in the dataset. We used Nanopore long-read sequencing technology (Minion device). One flowcell was used and produced 4.77 Gbp of sequence (11× genome coverage). We manually checked for the presence of TIPs of *Hopi*, *Tos17*, and *Karma*, from the reads thus generated[40]. We chose these three families because they are representative of the diversity of the retrotransposons found in the reference rice genome, i.e., a low (*Tos17*) and a high (*Hopi*) copy-number LTR-retrotransposon families and a moderately repeated LINE family (*Karma*). We should however emphasize that TRACKPOSON could only unravel the transpositional activity of TE families identified from the reference genome sequence and therefore that we could not test the efficiency of this software on a retrotransposon family with genomic features that are very different from *Hopi*, *Tos17*, or *Karma*, in case such family exists in one of the 3000 rice genomes represented in the dataset. TRACKPOSON identified 501, 8, and 9 insertions of *Hopi*, *Tos17*, and *Karma*, in Mutha Samba cultivar, respectively. Hundred percent of the insertions of both *Tos17* and *Karma* identified by TRACKPOSON were validated. In the case of *Hopi*, 473 insertions were validated. These results show that the specificity of TRACKPOSON is ~94.5%. The rate of false negatives (sensitivity) was also estimated. While there was a 100% overlap between the insertions detected by TRACKPOSON and those detected using Nanopore reads for both *Tos17* and *Karma*, which is indicative of 100% sensitivity for these two families, we detected a total of of 581 *Hopi* insertions from the Nanopore dataset, which corresponds to a drop of sensitivity to 81% for this family. However, the mappability of a complex genome such as that of rice with small reads obtained on Illumina platforms is expected to not be 100% because of the presence of repeats that impede unambiguous mapping at unique sites[41]. In the case of rice, the mappability of 100 bp windows at a $1e^{-20}$ threshold (i.e., that used for the second step of the detection, see Methods section) was estimated to be 63.5%, on average. This estimation fits with previous estimates of the repeat content of *Japonica* rice genome (i.e., ~40%[15]). We determined the mappability of the regions where we detected *Hopi* insertions with Nanopore reads. We found an average of 58% for all insertions. However, the insertions found only in the nanopore data (and not detected by TRACKPOSON) have a significant lower mappability of 42%. This may explain the lower sensitivity of our software for this particular family.

Retrotranspositional landscape of Asian rice. TRACKPOSON was used to detect TIPs for 32 retrotransposon families in the 3000 rice genomes dataset. The number of complete elements for these families in the reference rice genome is given in Table 1. We chose a sample of families that are representative of the diversity found in this accession in terms of number of repeats, from very moderately (e.g., *Tos17* with two complete copies in the reference genome) to highly repeated families (e.g., *Houba* with 150 complete copies). Therefore, we consider that the features revealed by this sample of retrotransposons should be representative of the complete retrotranspositional landscape of rice genome shaped by the whole 300 LTR-retrotransposon families[16]. In total, we identified 53,262 and 47,007 TIPs for the 3000 genomes and the 1067 traditional varieties included in the dataset, respectively (Table 1). Each family showed some variation of copy number among the 3000 rice accessions (Fig. 2a). Overall, the total number of insertions when all families were considered varied from 3324 to 12,380 with an average of 6225. No rice accession was found to be an outlier in terms of LTR-retrotransposon content.

We then analyzed the map position of all the TIPs. As mentioned above, we first determined the mappability of the rice genome in order to avoid bias in the interpretation of our results. As shown in Fig. 3, the mappability of TIPs decreased in pericentromeric regions, known to be repeat-rich. The mapping of all TIPs on the 12 rice pseudomolecules however shows an insertion bias in these regions, especially for *Gypsy* elements (Fig. 3), which confirms previous studies[42,43]. However, the mapping data shows that no region of the genome is devoid of TIPs, which suggests that retrotransposition contributes to genome diversity in all chromosomes regardless their position.

Table 1 TRACKPOSON results for 32 transposable elements families in 3000 rice genomes[a]

TE family	Families	Total TE insertions number	Total TE insertions number in traditional rice	Mean insertion in *japonica* varieties	Mean insertion in *indica* varieties	Distance from gene (kb)	Activity
Poprice	*Copia*	2324	1678	98.3 ± 12.1[b]	91.6 ± 14.0[b]	22.5 ± 31.4[b]	Recent
Tos17	*Copia*	181	121	7.6 ± 3.6	7.9 ± 2.9	21.1 ± 35.8	Recent
Houba	*Copia*	5976	5112	523.8 ± 65.1	388.8 ± 57.0	22.6 ± 32.0	Recent
Fam89_osr7	*Copia*	587	425	47.4 ± 8.7	32.0 ± 7.0	23.4 ± 33.3	Recent
Fam35-fam36	*Copia*	1756	1678	669.0 ± 63.7	637.2 ± 77.0	22.9 ± 32.2	Old
Fam67_echidne	*Copia*	438	371	105.8 ± 20.5	102.2 ± 24.4	23.4 ± 33.4	Recent
Rn304	*Copia*	52	30	2.2 ± 1.5	1.9 ± 1.2	24.8 ± 36.9	Continuous
Scaff6	*Copia*	29	23	1.7 ± 0.5	1.9 ± 0.9	27.9 ± 41.0	—
Lullaby	*Copia*	153	92	5.6 ± 2.1	3.1 ± 1.8	26.8 ± 37.4	Recent
Fam51_osr4	*Copia*	1788	1328	104.6 ± 11.0	100.0 ± 12.2	22.6 ± 31.6	Recent
Fam90	*Copia*	285	166	17.9 ± 2.9	16.6 ± 4.1	24.4 ± 32.5	Recent
Fam93_ors14	*Copia*	2194	1521	22.9 ± 13.5	56.4 ± 12.9	22.0 ± 31.8	Recent
Fam98_rn81	*Copia*	153	123	20.6 ± 5.1	10.8 ± 4.0	26.4 ± 35.1	Recent
Hopi	*Gypsy*	5152	5027	695.7 ± 179.5	701.3 ± 215.0	23.5 ± 33.7	Continuous
Dagul	*Gypsy*	2924	2742	571.8 ± 67.9	527.8 ± 75.6	22.6 ± 31.9	Recent
Fam17_Rn215_125	*Gypsy*	7096	7006	575.9 ± 175.5	930.1 ± 319	23.8 ± 34.6	Continuous
Fam80_rire7	*Gypsy*	382	338	66.5 ± 12.8	69.9 ± 14.2	23.4 ± 32.8	Continuous
Dasheng	*Gypsy*	5723	4806	586.4 ± 60.5	486.7 ± 61.1	22.8 ± 32.3	Recent
Rire2	*Gypsy*	3061	2785	304.6 ± 53.0	295.9 ± 70.3	22.8 ± 32.2	Recent
Fam81-fam82	*Gypsy*	2758	2558	474.2 ± 74.3	463.3 ± 94.3	23.6 ± 34.5	Continuous
Fam31_osr37	*Gypsy*	3368	2797	393.1 ± 53.4	430.7 ± 57.0	23.4 ± 33.2	Recent
Fam49_osr29	*Gypsy*	968	904	304.4 ± 36.0	284.0 ± 44.6	23.5 ± 33.6	Old
Fam124_rn208	*Gypsy*	112	68	11.02 ± 1.9	10.47 ± 1.8	26.3 ± 35.8	Continuous
Fam108	*Gypsy*	594	570	201.5 ± 31.3	188.1 ± 36.6	26.2 ± 37.6	Continuous
Fam106	*Gypsy*	213	148	15.8 ± 3.7	9.3 ± 3.7	23.7 ± 32.7	Recent
Fam86	*Gypsy*	506	450	202.1 ± 16.7	192.0 ± 20.8	23.9 ± 33.5	Continuous
Fam79_rn206	*Gypsy*	601	461	125.8 ± 10.8	120.9 ± 13.3	23.5 ± 32.4	Recent
Rn60	*Gypsy*	16	15	1.3 ± 1.4	1.2 ± 0.9	25.6 ± 38.7	—
Scaff3	*Gypsy*	56	28	1.4 ± 1.1	1.9 ± 1.3	32.8 ± 48.1	Recent
Scaff5	*Gypsy*	19	1	1 ± 0	0.7 ± 0.5	27.0 ± 42.6	—
Rire3	*Gypsy*	3719	3576	458 ± 147.0	408.2 ± 152.3	23.9 ± 34.9	Continuous
Karma	*LINE*	78	59	3.7 ± 1.4	4.2 ± 2.2	25.2 ± 36.7	Recent
32 families		53,262	47,007				

[a]The transpositional history of each TE families was based on the histogram in Fig. 2
[b]Numbers after "±" denote standard deviation

Furthermore, the position of the TIPs in relation to the closest gene differed significantly between *Gypsy* and *Copia* elements (*t* test, *p* value $<1.6^{e-7}$), with the *Copia* elements insertions being closer to genes than the *Gypsy* elements (Table 1; method described in Supplementary Table 3).

The level of polymorphism generated by the 32 retrotransposon families was assessed by using the frequency of insertions found in the 1067 traditional varieties (Fig. 2). Surprisingly, a large portion of TIPs are specific to only one variety (Fig. 2c). The sensitivity of TRACKPOSON is high but not 100% for highly repeated families. We therefore may not exclude the possibility that some insertions may have been missed in some accessions, but this may certainly not change the L-shape distribution observed in Fig. 2c. Therefore, our results strongly suggest that transposition may have occurred in agro, after domestication. Moreover, these low-frequency TIPs were found in all varietal groups, regardless of their geographical origin, which further suggests that transposition is triggered in various agro-environments. However, the 32 families did not contribute to genome diversification in the same fashion (Fig. 2b): some families (e.g., *Houba*) exhibit only very low-frequency insertions (L-shaped distribution of the number of accessions sharing the insertions), which suggests a recent transpositional activity or a segregation of ancient polymorphisms via lineage sorting. However, some, like *Hopi*, exhibited insertions at frequencies ranging from ~0.001 (insertions found in only one variety) to ~1 (ancestral insertions found in all traditional varieties). This finding suggests that such families have undergone transposition continuously since domestication and that the low frequency of *Houba* insertions is likely not due to lineage sorting, but to recent activity. Finally, for a few families (e.g., *Dasheng*), most insertions were found at high frequencies, which suggests more continuous activity.

The frequency of each TIP in the 1067 traditional varieties, together with its map location clearly shows that the TIPs found at high frequency are mostly pericentromeric (Fig. 3). Such high frequency TIPs could be considered as old insertions (because they are shared by many varieties) as opposed to the ones found at low frequency. The higher TE density usually found in pericentromeric regions in plant genomes could thus result from retention of TEs in addition to an insertional bias in these regions for *Gypsy* elements.

No genetic factor for transposition activation in rice. By showing that retrotransposition contributes to genomic diversification in rice and to such a large extent, we confirm that it is a major force driving genome evolution in this crop and likely in plants in general. However, in rice, like in the other model species *Arabidopsis thaliana*, transposition is strictly controlled by several epigenetic pathways[44], which raises the question of whether transposition is triggered in agro by mutations in key genes

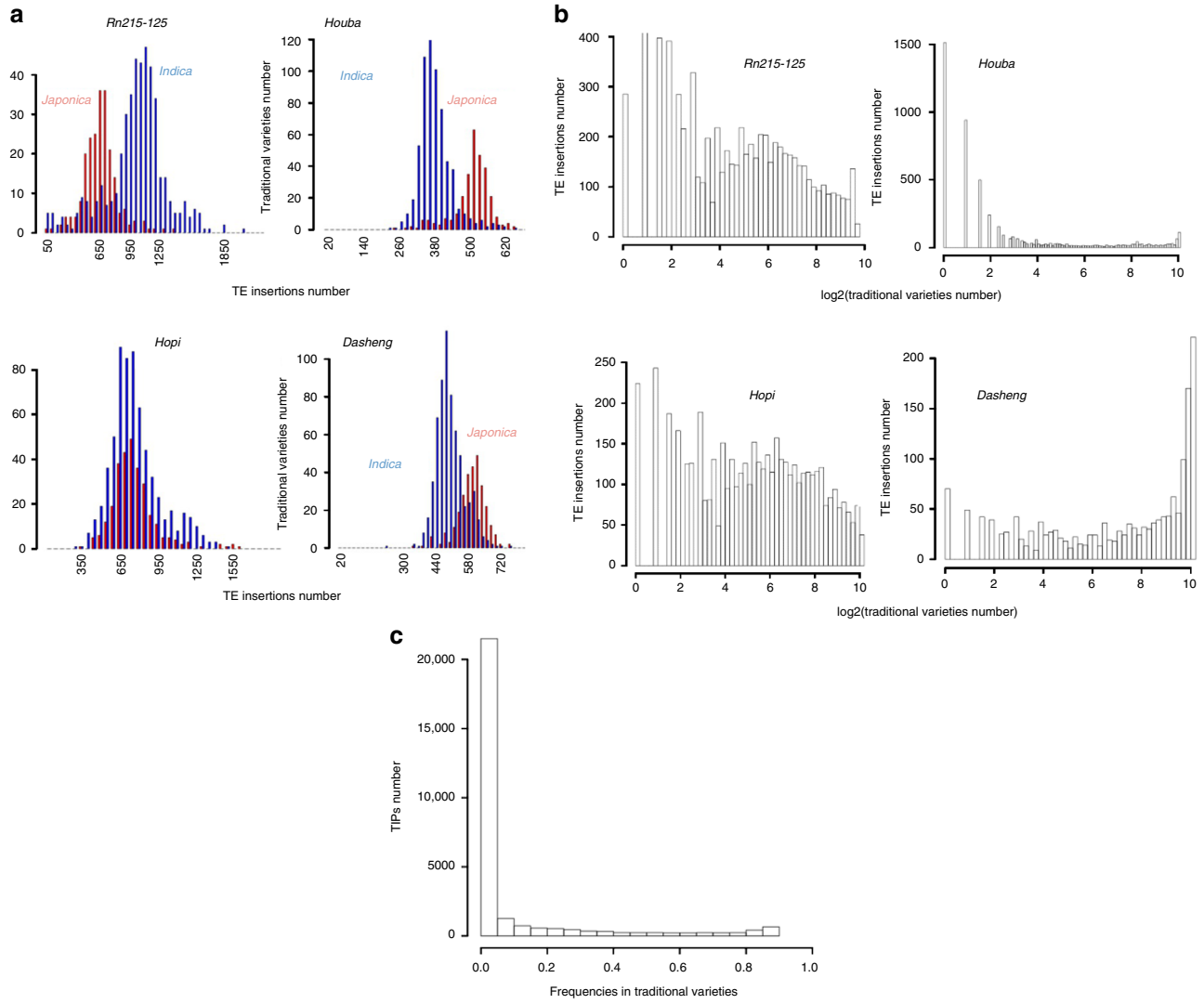


Fig. 2 History of TE families. **a** Distribution of the number of TE insertions in traditional varieties. *x* axis represents the number of TE insertions for each TE family. *Indica* and *Japonica* varieties are shown in blue and red, respectively. **b** Distribution of traditional varieties by TE insertion. The *x* axis represents the number of rice traditional varieties in log2 scale and *y* axis the number of TE insertions. For all the TE families, the peak on the left corresponds to TE insertions present at very low frequencies in rice varieties, which suggests recent insertions. **c** Distribution of frequencies of TIPs in traditional varieties. *x* axis shows the frequency of TE insertion in traditional varieties and *y* axis shows the frequencies for all TE insertion in all 32 families. Most TIPs are present with low frequencies (<0.01), i.e., TE insertions are only in one or two rice varieties

involved in these pathways in cultivated populations or, alternatively, by external stimuli such as biotic or abiotic stress that may modify the epigenetic landscape of the genome, mimicking the impediment of these pathways. We tested both hypotheses by using a genome-wide association study (GWAS), similar to that of Quadrana et al.[45]. Using the single-nucleotide polymorphism (SNP) dataset generated in the framework of the 3000 rice genome project[46], we sought associations between these markers and the number of copies (taken here as a phenotype) for each retrotransposon family (Fig. 4 and Supplementary Figures 3–14). Significant association peaks were found for the 32 families, although the clearest results were obtained for the 12 less repeated families (e.g., *Tos17*, Fig. 4a). For the others, e.g. *Rire2* or *Rire3*, the peaks were too numerous. For the 12 less repeated families, a total of 26 significant peaks were found (Supplementary Figures 3–14). Twenty fell into a region with a copy of the element identified using our TRACKPOSON software. The remaining five peaks (*Fam86*x2, *Fam89*, *Fam124*, *Karma*, and *Houba*) did not fall within a locus harboring a copy of the retrotransposon. However, the mappability in these regions was <60% because of the presence of repeats, which suggests that some insertions may not have been detected with TRACKPOSON.

That a majority of association peaks overlapped with a TE insertion suggests that the copy number of a given family depends on the presence of an insertion of a member of the family at a site where it can be activated in planta. We should stress out that such insertion may be distinct from the previously characterized active copies of well-known families, like in the case of *Tos17* and *Karma*: Nipponbare genome harbors two copies of *Tos17*, one on chromosome 10 (position 15.4 Mbp) and the other on chromosome 7 (position 26.7 Mbp). It was previously shown that the latter was the one activated during callus culture[29] and therefore, it is often referred to as the active copy of the family. The peak that we identified on chromosome 7 is located at ~20 Mbp (Fig. 4a) and thus does not overlap with that copy. However, we identified another *Tos17* insertion, shared by 404 varieties (366 *Indica*s and 17 *Japonica*s), in the region of the peak. This suggests that there is a copy of *Tos17* found in rice germplasm,

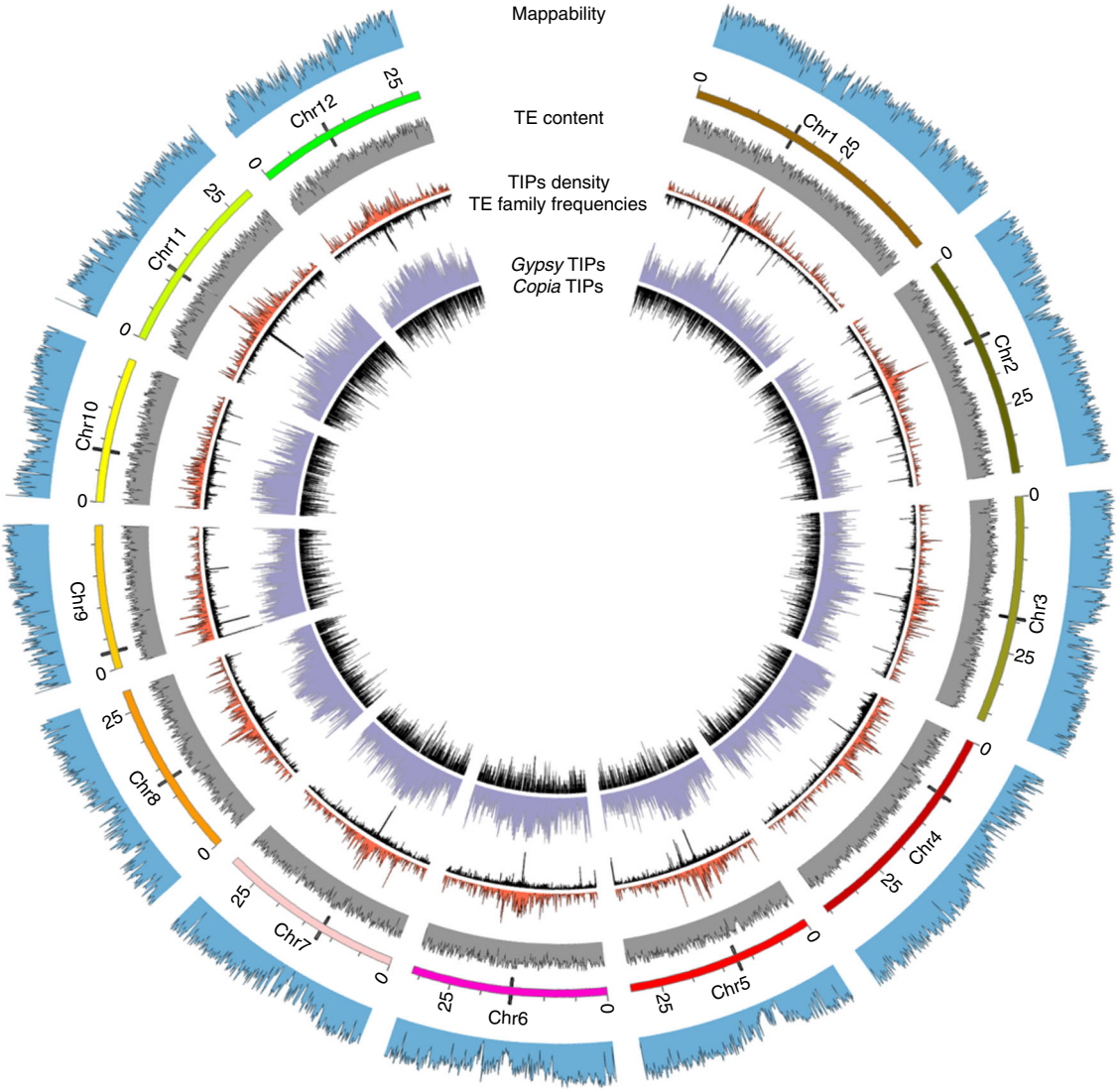
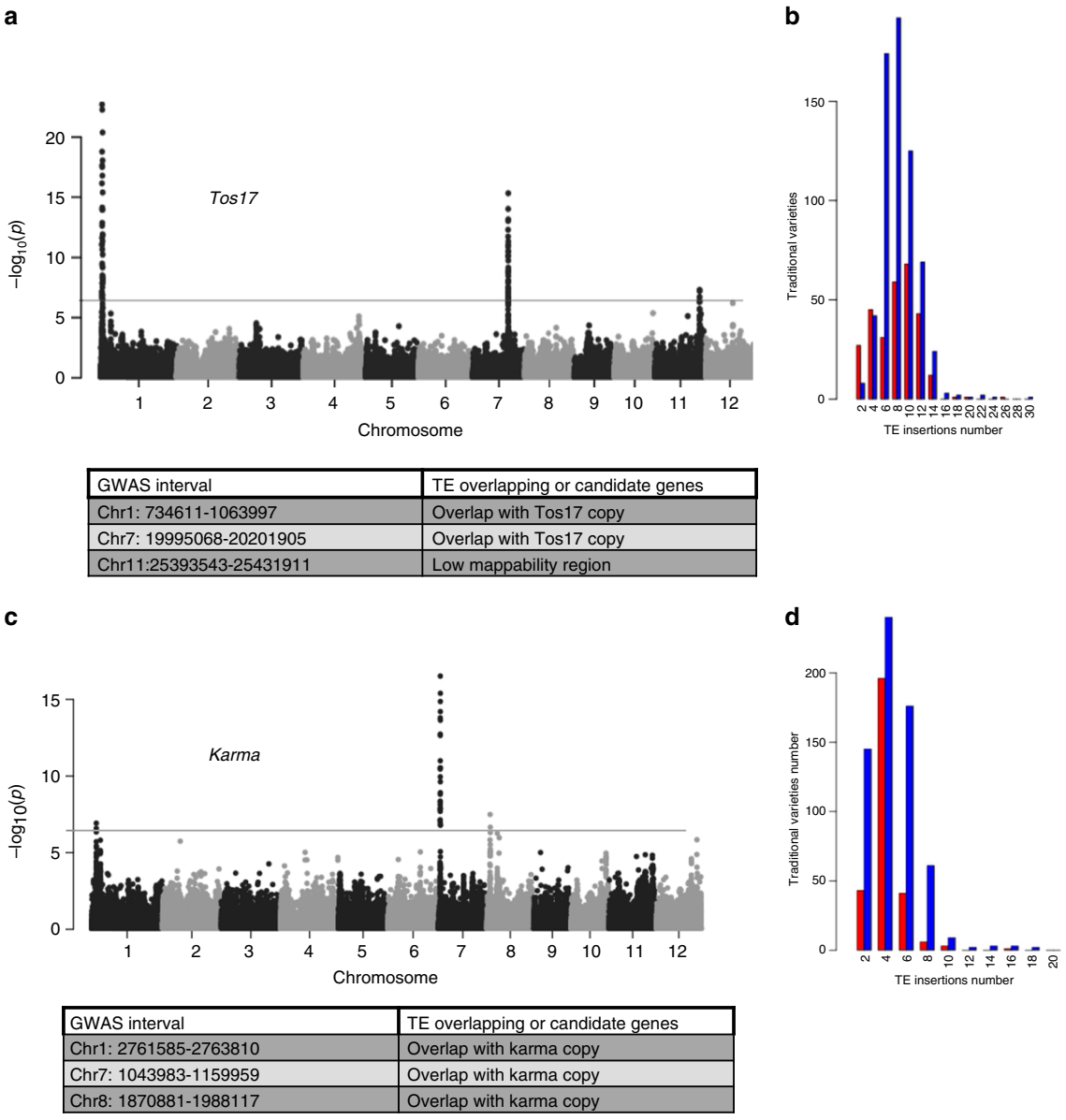


Fig. 3 Circos representation. From outside to inside: the first circle corresponds to the mappability of the 12 rice chromosomes (see Methods). The second circle represents the 12 chromosomes of rice genome. For each chromosome, the black tips correspond to the centromere. The third circle represents the TE content: all TE insertions present in rice annotation file. The fourth circle (in red) corresponds to the mapping of the TIPs for all the 32 TE families and opposed, below, in black, the fifth circle shows the number of different TE families for each TE insertion polymorphisms (TIPs), (i.e., the scale is from 1 to 32). The sixth circle represents the distribution of TE insertions per LTR-retrotransposon type: *Gypsy* (purple) or *Copia* (black)

which may be more active than that of Nipponbare on chromosome 7, or at least which may be active in agro since its presence is correlated with an increase in copy number of the element in rice germplasm. The conditions of its transpositional activation remain however to be elucidated. In the case of the LINE *Karma*, Nipponbare genome harbors only one complete copy located on chromosome 11 (at ~27 Mbp). That particular insertion is found in 1558 rice lines (among which 821 from *Indica* type and 568 from *Japonica* type). It also harbors an inactive truncated copy (often found with LINEs) on chromosome 5 (at ~13.2 Mbp). *Karma* was identified as transpositionally active in rice calli[30], like for *Tos17*. The peak that we obtained in our GWAS is located on chromosome 7 (Fig. 4b). It overlaps with an insertion of *Karma*, shared by 697 lines (among which 469 from *Indica* type and 62 from *Japonica* type). It therefore appears that the most active *Karma* copy is not the one previously identified in Nipponbare, similar to what observed for *Tos17*.

The gene annotation of the 26 regions for which a significant association was found did not reveal the presence of any genetic factors known to be involved in the control of transposition[14], unlike what was is found in *Arabidopsis*[45]. This result, together with the fact that most peaks overlap with a TE insertion, suggests that one cause of transposition activation in agro is the presence of an active copy of the element, rather than an alteration of a genetic factor controlling a cellular pathway for transposition control (i.e., epigenetic silencing). In addition, the presence of such active copy may not be sufficient, as transposition has been shown to be triggered in particular physiological states, such as biotic or abiotic stress in plants that may modify the methylation status of TEs[47]. In the case of rice, the exact nature of such external stimuli and the dynamics of the plant's response to it remain to be elucidated.

Cultivated rice may originate from distinct domestication events. We first used the TIPs data for principal coordinate analysis of the 3000 rice accessions, and showed that TE insertions, like SNPs, clearly discriminate *Indica* and *Japonica* varieties, while the *Aus/Boro* group appeared to be more similar to the *Indica* group (Fig. 5a). We then tentatively dated the origin of



GWAS interval	TE overlapping or candidate genes
Chr1: 734611-1063997	Overlap with Tos17 copy
Chr7: 19995068-20201905	Overlap with Tos17 copy
Chr11:25393543-25431911	Low mappability region

GWAS interval	TE overlapping or candidate genes
Chr1: 2761585-2763810	Overlap with karma copy
Chr7: 1043983-1159959	Overlap with karma copy
Chr8: 1870881-1988117	Overlap with karma copy

Fig. 4 GWAS analysis of *Tos17* and *Karma* TE families. GWAS analysis for *Tos17* (**a**) and *Karma* (**b**) TE families. Manhattan plot represents the log10 *p* value for each association between the single-nucleotide polymorhisms and the TE insertion (see Methods). TE insertion-SNP association *p* value >6 are significative. **c**, **d** The TE insertion distribution in rice traditional varieties (Fig. 3) reflects a polymorphism in TE copy number along rice varieties, considered a quantitative phenotype

the three varietal groups *Japonica*, *Indica*, and *Aus/Boro* using a genomic paleontology approach[32]. For this, we first identified all insertions of full elements in Nipponbare (for *Japonica*), IR8 (for *Indica*), and N22 (for *Aus/Boro*) high-quality genome assemblies[22] for the nine TE families showing the highest number of TIPs, i.e., *Dagul*, *Dasheng*, *Hopi*, *Houba*, *Osr37*, *Rire2*, *Rire3*, *RN215*, and *Poprice* (see Methods section for the details of the procedure). These TIPs were then classified into seven distinct categories: (1), (2), and (3): *Indica*-, *Japonica*-, and *Aus/Boro*-specific insertions, respectively; (4) ancestral insertions present in all traditional varieties, regardless of their varietal group; (5) insertions that are common between *Japonica* and *Indica* but not present in *Aus/Boro*, (6) insertions that are common between *Indica* and *Aus/Boro*, but absent from *Japonica*, and (7) insertions that are common between *Japonica* and *Aus* but absent from *Indica*. As previously mentioned, most TIPs are found at low frequency and therefore the insertions used for this analysis represented a small fraction of all TIPs (Fig. 6). Each TIP thus identified was dated using the method of SanMiguel et al.[48] with a molecular clock of 1.3×10^{-8} substitution/site/year[49]. In total, we successfully dated 1476 TIPs from the seven categories (Fig. 6a). As expected, the insertions of the fourth category (i.e., common among all varieties) are more ancient than those belonging to the other categories (*Indica*- and *Japonica*-specific, respectively), which illustrates a split of a common wild ancestor into three distinct gene pools. The distribution medians for *Indica*, *Japonica*, and *Aus/Boro*-specific TE insertions are 99.4% identity (~230,000 years ago), 99.2% identity (~310,000 years ago), and 99% identity (~380,000 years ago), respectively. These values represent a peak of transpositional activity after the split of the lineages that gave rise to the three cultivated types and could therefore be used to estimate the lower limit of the time of divergence between the

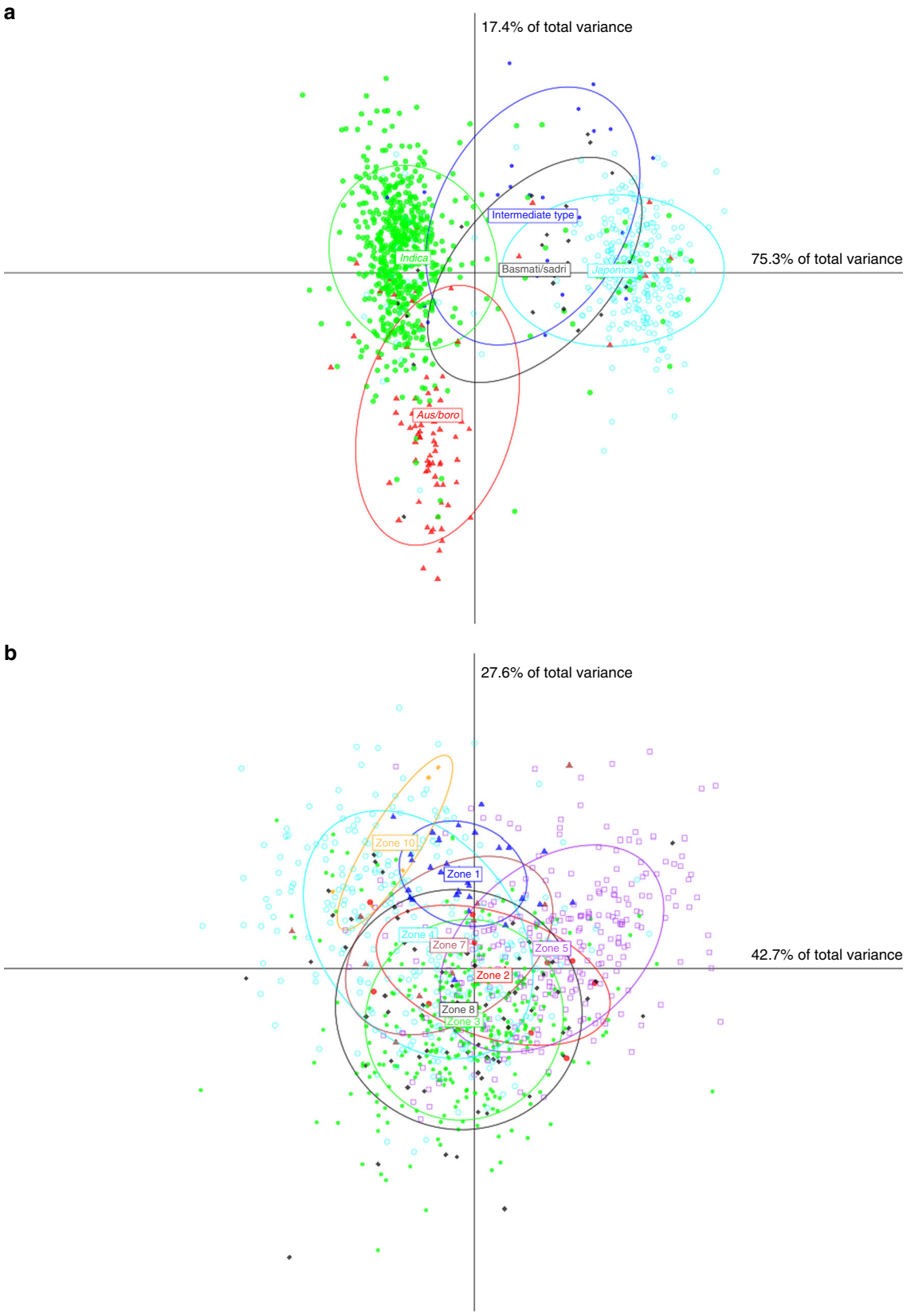


Fig. 5 Discriminant analysis of the 3000 genomes using TIPs. Discriminant analysis of principal components was performed using TE insertions for all the 32 TE families as a function of varietal groups in green for the *Indica* varieties and in blue for the *Japonica* varieties (**a**) or in function of geographical zones (**b**). Zone 1 (blue): Japan and Korea; zone 2 (red): China; zone 3 (green): South-East Asia—Thailand, Laos, Malaysia, Cambodia, and Myanmar; zone 4 (cyan): Asian peninsula—Taiwan, Indonesia, Philipines, and Australia; zone 5 (purple): Pakistan, India, Bangladesh, Nepal, Egypt, and Sri Lanka; zone 6: Europe; zone 7 (brown): Madagascar; zone 8 (black): Africa; zone 9: North America; and zone 10 (orange): South America

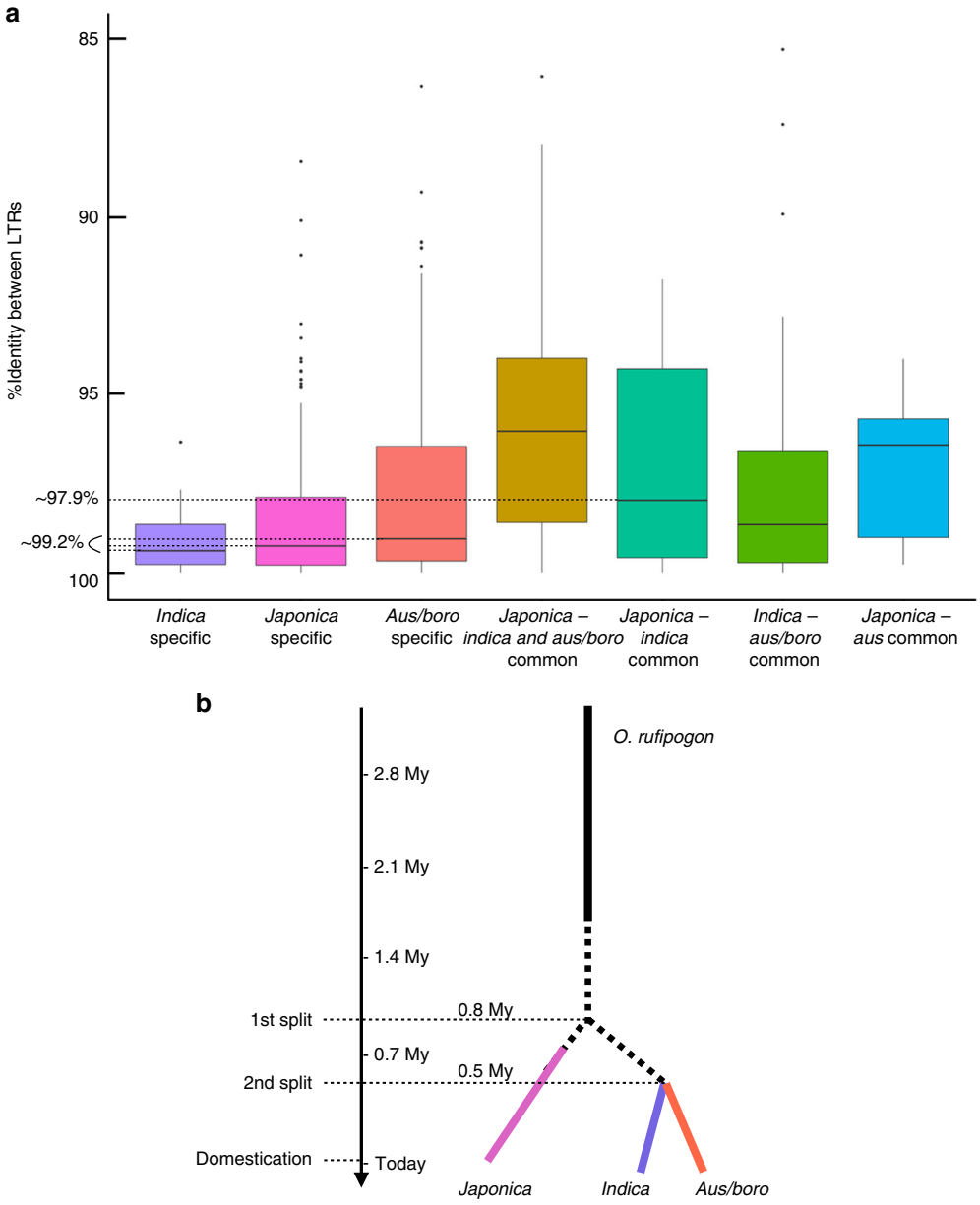


Fig. 6 Origin of rice domestication. **a** The box plots represent the distribution of percent identity between LTRs of retrotransposons for insertions that are *Indica*-specific ($N = 91$), *Japonica*-specific ($N = 266$), *Aus/Boro*-specific ($N = 216$), common to the three groups ($N = 138$), common to *Indica/Japonica* ($N = 32$), common to *Indica/Aus* ($N = 30$), and common to *Japonica/Aus* ($N = 12$), respectively. The three horizontal dashed lines correspond to the mode of distribution of *Indica*-, *Japonica*-, and Aus-specific insertions. **b** Representation of the history of the three domestications of Asian rice

genomes of the wild progenitors of the three groups (Fig. 6b). The three values are significantly older than 10,000 years, i.e., the origin of rice cultivation at late Neolithic. The distribution median of *Indica–Japonica* common insertions is 97.9%, which translates into a date of 800,000 years ago. This confirms that the two gene pools of *Oryza rufipogon* from which both cultivated types originated were separated long before the origin of agriculture. This value also provides an estimate of the upper limit of the date of divergence between *Indica* and *Japonica* progenitors which, combined with the estimated date of the *Indica*- and *Japonica*-specific categories leads us to propose that the date of the divergence ranges from ~300,000 years ago to 800,000 years ago. The split of the *Indica/Aus* lineages appears to be more recent with a median at 98.6% (~540,000 years ago). Combined with the median age of the *Aus*-specific insertions (see above), our results suggest that the split between the progenitors of *Indica* and *Aus* may have occurred between ~230,000 and ~540,000 years ago. However the median date of insertions that are common between *Japonica* and *Aus* (seventh group) appears to be older (96.4%, translating into a date of 1.4 Mya) than the median date of insertions that are common between *Japonica* and *Indica* (estimated above at 800,000 years ago). The much smaller sample of *Aus* varieties in the 1067 traditional cultivar's dataset (84 accessions) may explain this discrepency. Alternatively, it could originate from demographic history of the populations of the rice progenitor *O. rufipogon*. Further analyses may help clarifying this last point.

Our results, synthesized in Fig. 6b strongly suggest that *Indica*, *Japonica*, and *Aus/Boro* originated from three distinct domestication events. Unfortunately, the density of TIPs of all seven

categories was not sufficient to identify traces of introgression in the vicinity of domestication loci and therefore cannot solve the paradox of the presence of a single allele for these loci[34–36].

Discussion

Several softwares designed for TIPs detection are currently available[26,27]. However, all require as a first step to map all sequencing reads onto a reference genome sequence, using Burrows–Wheeler-based algorithms. This mapping step is computationally intensive and thus the computation time necessary to analyze a large dataset (i.e., >1000 genomes) is too long. The availability of large datasets is now only a reality for a few plant species such as rice and *A. thaliana*, but this will certainly change in the near future. In this report, we show that TRACKPOSON, the new software we developed for identifying TIPs, is efficient for the analysis of thousands of genomes. It produces few false positives, as evidenced by wet-lab validation (Supplementary Figure 1 and Supplementary Tables 1 and 2). False negatives remain an issue with low genome coverage data; however, the two-pass procedure (see Results and Methods sections) can considerably limit the risk of missing TIPs, as confirmed by the sequencing of one accession with long-read technology (Nanopore). However, unlike the softwares cited above, TRACKPOSON is not designed for the exhaustive characterization of all structural variations caused by transposition, but rather for the complete and efficient characterization of the transpositional activity of known families among thousands genomes.

The 53,262 TIPs we identified in the 3000 rice genomes for 32 retrotransposon families were found at low frequency among rice accessions, similarly to what was observed for *A. thaliana*[45,50]. Most insertions are private or shared by two accessions, which suggests that they occurred after rice domestication and therefore in agro. This implies that retrotransposition-driven genomic diversification is ongoing in rice fields. We looked for genetic factors—the impediment of which may be causative of such activation, but found no evidence of mutations in genes of known epigenetic pathways involved in transposition control. Instead, the majority of peaks were found where a TIP of the same TE family was identified. From this, we propose that transposition in agro may require both the presence of an active TE copy in the genome and an external stimulus that may modify the epigenetic status of such copy, although the nature of such stimulus in rice is unknown. In particular, whether biotic and/or abiotic stresses are involved like in the case of other species remains to be elucidated[47].

We used retrotransposon insertions as paleogenomic tools to investigate the origin of rice domestication and clearly showed that *Indica*, *Japonica*, and *Aus/Boro* genomes diverged significantly earlier than rice domestication during late neolithic ca. 10,000 years ago. This finding contradicts the single origin hypothesis based on the identification of a single "domesticated" allele in most domestication loci. Although we cannot resolve this paradox, the most parsimonious interpretation of existing data is that at least three domestications occurred from three distinct gene pools of the wild rice progenitor *O. rufipogon*. The distribution of *O. rufipogon* in Asia today does not contradict this hypothesis because the species is now found all across South Asia, South-East Asia, and East Asia. Moreover, it is possible to identify populations that are closely related to each of the two cultivated types *Japonica* and *Indica*[38]. The presence of only one domesticated allele at a key domestication locus must then result from the introgression of one domesticated form by the other, thus explaining the loss of one of the two original alleles[37]. In support of this hypothesis, the current distribution of all varietal groups in ssAsia shows that they have been extensively disseminated throughout the continent for a long time (Fig. 5b). Alternatively, a single allele may have been present in the wild progenitor and selected several times independently to give rise to *Japonica* and *Indica* rice[38].

The complete TIP data reported here is based on the analysis of 32 retrotransposon families. We chose a sample of families that is representative of the diversity found in the rice genome[16]. Other TE types, e.g., Miniature Inverted Transposable Element (MITEs), helitrons, and transposons should also be investigated in order to complete the characterization of the transpositional landscape of rice. The small size of many TE families of MITEs and Short Interspersed Element (SINEs) may however require to modify our detection method since the initial mapping step may be hindered by the small size of the element.

The advent of large genomic datasets opens new perspectives in the exploitation of genetic resources for most crops. For rice, the 3000 genomes are now being routinely used for gene discovery. However, most studies rely on the use of SNP data, while large InDels and in particular those caused by the activity of TEs remain mostly unexploited. In this study, we show that TEs are at the origin of a large extent of structural variations found genome-wide and arise at a rapid rate. Hence, these mobile elements contribute significantly to the genomic diversity of rice. We also show that most of this diversity originated after domestication, therefore in rice fields and wherever rice is grown, further suggesting that the whole rice gene pool may be a reservoir of structural variants. Whether such TE-driven genomic changes are phenotypically relevant and could therefore have been selected for as a source of adaptation for rice remains to be demonstrated, although evidence that TIPs are causal to phenotypes of agronomic interest have been reported for other crops. This dataset will allow others to explore the role of TIPs in rice diversification and breeding.

Methods

TRACKPOSON pipeline. The TRACKPOSON pipeline is freely available at http://gamay.univ-perp.fr/~Panaudlab/TRACKPOSON.tar.gz. It is designed to detect TIPs in large datasets (>1000 genomes), using an existing TE database. The first step consists in mapping paired reads of genomic data in fastq format onto indexed consensus sequence of each TE family. Bowtie2 (v. 2.2.0) in very-sensitive mode[51] is used in this first step. The Sam file thus obtained is parsed using the flag value as a criterion as follows: only the read pairs for which one read mapped against the TE were kept. The next step consists of mapping the unmapped paired read onto the rice reference genome sequence (Nipponbare rice genome IRGSP1.0) using blastn (v. 2.2.31+)[52] with an *e* value threshold of $1e{-}20$. Only unique blast hits were considered and converted into bed format. In parallel, the Nipponbare genome (IRGSP1.0) was split into 10 kb windows by using bedtools makewindows (v. 2.25.0). Each TE insertion was thus assigned to a 10-kb window. Because a minimum of five amplicons spanning the insertion is required as an initial detection threshold, once a TIP had been detected in at least one accession, then the third step of the procedure consists in a new scan of all remaining accessions with a lower threshold (2 amplicons). Finally, a full matrix of presence/absence of TE insertions was created by using a home-made R (v. 3.3.1) script, with the 3000 rice genomes in columns and all the TIPs (for the TE family studied) in lines. These data are freely available at http://gamay.univ-perp.fr/~Panaudlab/TRACKPOSON_Results.tar.gz.

The wall time for running the software on the 3000 genomes ranged from a few hours to several days on a cluster of 88 cores, depending on the level of repetition of the family, confirming that this new strategy considerably improved the speed of detection.

Mappability. As mentioned in the paragraph above, only reads mapping unequivocally at a unique position onto the reference genome sequence were kept for positioning the TIPs. This obviously reduced the mappable genome to only unique sequences, which are estimated to be ~60% of the rice genome[15]. Each 10 kb window was sliced into 100 bp sequences—the repeat level of which was estimated using blastn against the reference genome with the same parameters as used in the mapping step of TRACKPOSON. Results were concatenated over each window to produce a percentage of mappability. These data were plotted onto the genome (Fig. 2).

Validation TIPs with Nanopore sequencing. We resequenced one *Indica* rice variety—i.e., IRIS-313–11419 using the Nanopore long-read technology. The

library was prepared using the 1D kit according to the manufacturer's instructions. One R9.4 flowcell was used. After basecalling of long reads with Albacore Oxford Nanopore software (to convert fast5 files in fasta format), a blast database was created. We performed a blastn of the TE families against this Nanopore database, with *e* value 1*e*−50 and penalities for open and extend gap equal to 0.

Only the reads with a High Scoring Pair (HSPs). corresponding to maximum of 80% of their length were kept. With this filtering, we eliminated the reads corresponding only to a TE. All the reads were validated by hand by dotter and NCBI blast against the rice reference genome IRGSP1.0. For each read thus selected, 300 bp of sequences flanking the insertion were used as query for blastn search against the IRGSP 1.0 rice pseudomolecules with 1*e*−50 *e* value threshold. This allowed unambiguous mapping of the TE insertions.

Distance between TE insertion and gene estimation. Estimation was performed with bedtools (v. 2.25) between the Nipponbare gtf annotation file (locus_IRSP1–0_predicted.gtf) and the output of TRACKPOSON pipeline (i.e., the TIPs localization).

GWAS analysis. For each of the 32 TE families, the number of TE insertions was determined for each variety by summing the number of TE copies identified. Thus, TE copy number was treated as quantitative phenotype from TE mobilization along the varieties. GWAS for TE copy number was carried out with a 404 K coreSNP dataset (160 K after minor allele frequency>5% and missing data<20%) downloaded from the Rice SNP-Seek Database[46] and using a linear-mixed model in EMMAX[53]. This procedure takes the underlying population structure into account by including a kinship matrix as a random effect. After checking the values of FDR (performed with *Q*-value R package), a stringent threshold of $-\log10\ P = 6$ was set to declare a significant association. The GWAS interval was determined by taking into account the significant SNPs at each locus ($-\log10\ P \geq 6$).

For characterization of the GWAS intervals, the TE annotation file and the Locus file containing all gene models were downloaded from the MSU Rice Genome Annotation Project v 7. All genomic annotations overlapping with these intervals were considered as putative causal genes. Regarding overlaps with TE, the GWAS intervals were compared to annotated TEs in the reference genome or the non-reference TE insertions identified by TRACKPOSON. The TE insertion that overlaps with GWAS intervals of the same family should be causal factors.

Discriminent analysis. DAPC was performed with the adegenet package (dapc function) in R (v. 3.3.1).

Genomic paleontology. Complete retrotransposon insertions were characterized in three well assembled rice genomes (*Japonica* Nipponbare, *Indica* IR8, and *Aus/Boro* N22) for nine of the most repeated families (i.e., *Dagul*, *Dasheng*, *Rire2*, *Rire3*, *Hopi*, *Houba*, *RN215*, *Osr37*, and *Poprice*). For this, a home-made script perl (available upon demand) was used. Briefly, for each family, three sequences are first generated, i.e., the full element, the LTR sequence, and the portion of the internal sequence that corresponds to the RT domain (except for the LARD dasheng for which a non-complex portion of the internal sequence was used). The first step consists in looking at all the paralogs of the element using the RT domain as a query in a blast search against the reference genome, each hit being extended both upstream and downstream, and the resulting genomic sequence checked for the presence of both LTRs (using blastn with the LTR of the consensus sequence against the genomic region obtained above). After trimming the sequence, the identity between the two LTRs was estimated by splitting the element into two equal sequences and blasting one half against the other (only the LTRs produce alignments, from which a percentage of identity was obtained from the largest HSP).

Once all complete elements were identified in both genomes, their orthologous relationship was secured as follows: 300 bp of genomic sequence upstream of the element was extracted from the first reference genome sequence and used as query in a blastn search against the other genome. Only sequences producing a single hit with at least 90% of the query length were considered orthologous and therefore kept for further analyses. The presence of the element at an orthologous position in the other genome was checked by first extracting 15 kb of the sequence downstream of the 300 bp and blasting the resulting sequence against the consensus sequence of the TE family. Therefore, each insertion could be classified as *Japonica*-specific (present in Nipponbare and absent from IR8 and N22), *Indica*-specific (present in IR8 but absent from Nipponbare and N22), *Aus*-specific (present in N22 but absent from Nipponbare and IR8), common to all (present in the three genomes), or common to two of the three genomes (three classes). This classification was finally validated as follows: *Indica*-, *Japonica*-, and *Aus*-specific insertions were retained only if present in at least 60% of varieties from the corresponding group, and the common insertions should be present in at least 80% of the varieties of the groups.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The Nanopore sequencing data is available on NCBI under the BioProject ID PRJNA507708. The previously published 3000 rice genome raw sequencing data are available from GigaScience Database (https://doi.org/10.5524/200001)[23]. A reporting summary for this Article is available as a Supplementary Information file. The source data underlying Fig. 2, Fig. 3a–c, Fig. 5, Supplementary Figure 1, Supplementary Figure 2, and Supplementary Table 3 are provided as a Source Data file.

Received: 25 April 2018 Accepted: 5 December 2018
Published online: 03 January 2019

References

1. Kidwell, M. G. Transposable elements and the evolution of genome size in eukaryotes. *Genetica* **115**, 49–63 (2002).
2. Wendel, J. F., Jackson, S. A., Meyers, B. C. & Wing, R. A. Evolution of plant genome architecture. *Genome Biol.* **17**, 37 (2016).
3. Leitch I. J. and Leitch A. R. Genome size diversity and evolution in land plants. In *Plant Genome Diversity Volume 2. Physical Structure, Behaviour and Evolution of Plant Genomes* (eds Greilhuber, J., Dolezel, J. & Wendel, J. F.) 307–322 (Springer-Verlag, Wien, 2013).
4. Wicker, T. et al. A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8**, 973–982 (2007).
5. Piegu, B. et al. Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Res.* **16**, 1262–1269 (2006).
6. Lynch, V. J., Leclerc, R. D., May, G. & Wagner, G. P. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. *Nat. Genet.* **43**, 1154–1159 (2011).
7. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* **351**, 1083–1087 (2016).
8. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory activities of transposable elements: from conflicts to benefits. *Nat. Rev. Genet.* **18**, 71–86 (2017).
9. Jangam, D., Feschotte, C. & Betrán, E. Transposable element domestication as an adaptation to evolutionary conflicts. *Trends Genet.* **33**, 817–831 (2017).
10. van't Hof, A. E. et al. The industrial melanism mutation in British peppered moths is a transposable element. *Nature* **534**, 102–105 (2016).
11. Kobayashi, S., Goto-Yamamoto, N. & Hirochika, H. Retrotransposon-induced mutations in grape skin color. *Science* **304**, 982 (2004).
12. Butelli, E. et al. Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. *Plant Cell* **24**, 1242–1255 (2012).
13. Barrón, M. G., Fiston-Lavier, A. S., Petrov, D. A. & González, J. Population genomics of transposable elements in Drosophila. *Annu. Rev. Genet.* **48**, 561–581 (2014).
14. Rigal, M. & Mathieu, O. A "mille-feuille" of silencing: epigenetic control of transposable elements. *Biochim. Biophys. Acta* **1809**, 452–458 (2011).
15. International Rice Genome Sequencing Project. The map-based sequence of the rice genome. *Nature* **436**, 793–800 (2005).
16. Chaparro, C., Guyot, R., Zuccolo, A., Piégu, B. & Panaud, O. RetrOryza: a database of the rice LTR-retrotransposons. *Nucleic Acids Res.* **35**, D66–D70 (2007).
17. Copetti, D. et al. RiTE database: a resource database for genus-wide rice genomics and evolutionary biology. *BMC Genomics* **16**, 538 (2015).
18. Panaud, O. et al. Characterization of transposable elements in the genome of rice (*Oryza sativa* L.) using representational difference analysis (RDA). *Mol. Genet. Genomics* **268**, 113–121 (2002).
19. Jiang, N., Feschotte, C., Zhang, X. & Wessler, S. R. Using rice to understand the origin and amplification of miniature inverted repeat transposable elements (MITEs). *Curr. Opin. Plant Biol.* **7**, 115–119 (2004).
20. Vitte, C. & Panaud, O. LTR retrotransposons and flowering plant genome size: emergence of the increase/decrease model. *Cytogenet. Genome Res.* **110**, 91–107 (2005).
21. Piegu, B. et al. Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Res.* **16**, 1262–1269 (2006).
22. Schatz, M. C. et al. Whole genome de novo assemblies of three divergent strains of rice, *Oryza sativa*, document novel gene space of *aus* and *indica*. *Genome Biol.* **15**, 506 (2014).
23. Rice Genome Project. The 3,000 rice genomes project. *Gigascience* **3**, 7 (2014).
24. Handsaker, R. E., Kom, J. M., Nemesh, J. & McCarroll, S. A. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat. Genet.* **43**, 269–276 (2011).

25. Sabot, F. et al. Transpositional landscape of the rice genome revealed by paired-end mapping of high-throughput re-sequencing data. *Plant J.* **66**, 241–246 (2011).
26. Fan X., Abbott T. E., Larson D. & Chen K. BreakDancer - identification of genomic structural variation from paired-end read mapping. *Curr. Protoc. Bioinformatics* https://doi.org/10.1002/0471250953.bi1506s45 (2014).
27. Hénaff, E., Zapata, L., Casacuberta, J. M. & Ossowski, S. Jitterbug: somatic and germline transposon insertion detection at single-nucleotide resolution. *BMC Genomics* **16**, 768 (2015).
28. Vitte, C. & Panaud, O. Formation of solo-LTRs through unequal homologous recombination counterbalances amplifications of LTR retrotransposons in rice *Oryza sativa* L. *Mol. Biol. Evol.* **20**, 528–540 (2003).
29. Hirochika, H., Sugimoto, K., Otsuki, Y., Tsugawa, H. & Kanda, M. Retrotransposons of rice involved in mutations induced by tissue culture. *Proc. Natl Acad. Sci. USA* **93**, 7783–7788 (1996).
30. Komatsu, M., Shimamoto, K. & Kyozuka, J. Two-step regulation and continuous retrotransposition of the rice LINE-type retrotransposon Karma. *Plant Cell* **15**, 1934–1944 (2003).
31. Picault, N. et al. Identification of an active LTR retrotransposon in rice. *Plant J.* **58**, 754–765 (2009).
32. Vitte, C., Ishii, T., Lamy, F., Brar, D. & Panaud, O. Genomic paleontology provides evidence for two distinct origins of Asian rice (*Oryza sativa* L.). *Mol. Genet. Genomics* **272**, 504–511 (2004).
33. Londo, J. P., Chiang, Y. C., Hung, K. H., Chiang, T. Y. & Schaal, B. A. Phylogeography of Asian wild rice, *Oryza rufipogon*, reveals multiple independent domestications of cultivated rice, *Oryza sativa*. *Proc. Natl. Acad. Sci. USA* **103**, 9578–9583 (2006).
34. Li, C., Zhou, A. & Sang, T. Rice domestication by reducing shattering. *Science* **311**, 1936–1939 (2006).
35. Huang, X. et al. A map of rice genome variation reveals the origin of cultivated rice. *Nature* **490**, 497–501 (2012).
36. Gross, B. L. & Zhao, Z. Archaeological and genetic insights into the origins of domesticated rice. *Proc. Natl Acad. Sci. USA* **111**, 6190–6197 (2014).
37. Wang, W. et al. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* **557**, 43–49 (2018).
38. Civáň, P., Craig, H., Cox, C. J. & Brown, T. A. Three geographically separate domestications of Asian rice. *Nat. Plants* **1**, 15164 (2015).
39. Elbaidouri, M., Chaparro, C. & Panaud, O. Use of next generation sequencing (NGS) technologies for the genome-wide detection of transposition. *Methods Mol. Biol.* **1057**, 265–274 (2013).
40. Debladis, E., Llauro, C., Carpentier, M. C., Mirouze, M. & Panaud, O. Detection of active transposable elements in *Arabidopsis thaliana* using Oxford Nanopore Sequencing technology. *BMC Genomics* **18**, 537 (2017).
41. Li, W. & Freudenberg, J. Mappability and read length. *Front. Genet.* **5**, 381 (2014).
42. Kumekawa, N. et al. A new gypsy-type retrotransposon, RIRE7: preferential insertion into the tandem repeat sequence TrsD in pericentromeric heterochromatin regions of rice chromosomes. *Mol. Genet. Genomics* **265**, 480–488 (2001).
43. Wang, Y. et al. Euchromatin and pericentromeric heterochromatin: comparative composition in the tomato genome. *Genetics* **172**, 2529–2540 (2006).
44. Chen, X. & Zhou, D. X. Rice epigenomics and epigenetics: challenges and opportunities. 2013. *Curr. Opin. Plant Biol.* **16**, 164–169 (2013).
45. Quadrana, L. et al. The *Arabidopsis thaliana* mobilome and its impact at the species level. *Elife* **5**, e15716 (2016). pii.
46. Mansueto, L. et al. Rice SNP-seek database update: new SNPs, indels, and queries. *Nucleic Acids Res.* **45**, D1075–D1081 (2017).
47. Grandbastien, M. A. LTR retrotransposons, handy hitchhikers of plant regulation and stress response. *Biochim. Biophys. Acta* **1849**, 403–416 (2015).
48. SanMiguel, P., Gaut, B. S., Tikhonov, A., Nakajima, Y. & Bennetzen, J. L. The paleontology of intergene retrotransposons of maize. *Nat. Genet.* **20**, 43–45 (1998).
49. Ma, J. & Bennetzen, J. L. Rapid recent growth and divergence of rice nuclear genomes. *Proc. Natl Acad. Sci. USA* **101**, 12404–12410 (2004).
50. Stuart, T. et al. Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation. *Elife* **5**, e20777 (2016).
51. Langmead, B. & Salzberg, S. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
52. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2008).
53. Kang, H. M. et al. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).

Acknowledgements

This work was supported by a CNRS/Région Languedoc Roussillon research grant and by the University of Perpignan via domitia. This work was publicly funded through a CNRS/Région Languedoc Roussillon research grant, by the University of Perpignan via domitia and ANR (the French National Research Agency) under the "Investissements d'avenir" programme with the reference ANR-10-LABX-001–01 Labex Agro and coordinated by Agropolis Fondation. The work was also supported by an Academia Sinica Thematic Project AS-TP-107-L02 in Taiwan. The authors thank Marie Mirouze, Scott A. Jackson, Josep Casacuberta, and Joris Bertrand for their useful comments on the manuscript.

Author contributions

M.-C.C. designed and built the TRACKPOSON software, performed TIP detection, contributed to the other analyses, and wrote the manuscript. E.M. and R.A. conducted GWAS. F.-J.W. and H.-P.W. implemented the software in the computer facility at Academia Sinica (Taipei, Taiwan). E.L. and E.D. performed the wet-lab experiments and analyses for the validations of TIPs. C.L. performed the Nanopore sequencing. Y.-I.H. supervised the work at Academia Sinica. O.P. supervised the project and performed the genomic paleontology analyses.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-018-07974-5.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

6.3 Perspectives

TRACKPOSON est un outil pour détecter de nouvelles transpositions d'éléments transposables au sein d'un grand jeu de données. Il a été développé pour détecter les ET actifs, les plus polymorphes. Le grand avantage de TRACKPOSON est sa rapidité et sa portabilité : il est facilement adaptable à d'autres génomes. Pour une meilleure visibilié, celui-ci a été déposé sur Github. TRACKPOSON a déjà été utilisé pour détecter les polymorphismes d'insertions au sein de différentes populations naturelles d'*Arabidopsis thaliana* (collaboration Fabrice Roux, INRA Toulouse et Valéry Hinoux, LGDP), chez la carotte avec les MITEs (collaboration avec Alicja Macko-Podgórni, University of Agriculture in Krakow) et les 1002 génomes de levures. En revanche, mon outil détecte toutes les copies mais d'une seule famille d'ET seulement. Comme vu dans l'introduction, une très faible proportion de familles d'ET sont actives au sein des génomes. Ainsi pour avoir une image de l'activité transpositionnelle d'une population, il n'est pas nécessaire d'analyser toutes les familles d'ET : les plus récentes sont les plus informatives, car considérées comme les derniers ET actifs. Ainsi avec TRACKPOSON il est possible d'étudier la dynamique d'ET au niveau micro-évolutif.

Castanera et al. ont effectué une analyse similaire (Castanera *et al.* 2021) au sein du même jeux de données des 3000 génomes de riz mais avec l'outil PoPoolationTE2 (Kofler, 2016) . Les résultats concernant la dynamique des rétrotransposons sont similaires à ceux de TRACKPOSON : ce qui montre la robustesse de ce dernier. Mais contrairement à mon outil, PopoolationTE2 permet la détection des insertions de toutes les familles d'ET en même temps. Donc cela nécessite un grand temps de calcul et de stockage pour conserver les fichiers d'alignements (les auteurs de cet article ont pu avoir accès au centre de calcul de Barcelone).

Lors de ces analyses, nous nous sommes focalisés sur les rétrotransposons, car ceux-ci sont les plus importants au sein des génomes de plantes. De plus, au sein de mon équipe de recherche, nous avons une base de données expertisée pour ces éléments de classe I. Ainsi, en collaboration avec Raul Castanera et Josep Casacuberta du CRAG de Barcelone, il a été décidé de mener la même analyse mais en se focalisant sur les ET de classe II, et plus spécifiquement les MITEs. Il a été observé que la dynamique des MITEs est assez différente, comparée à celle des rétrotransposons à LTR (Figure 6.1).

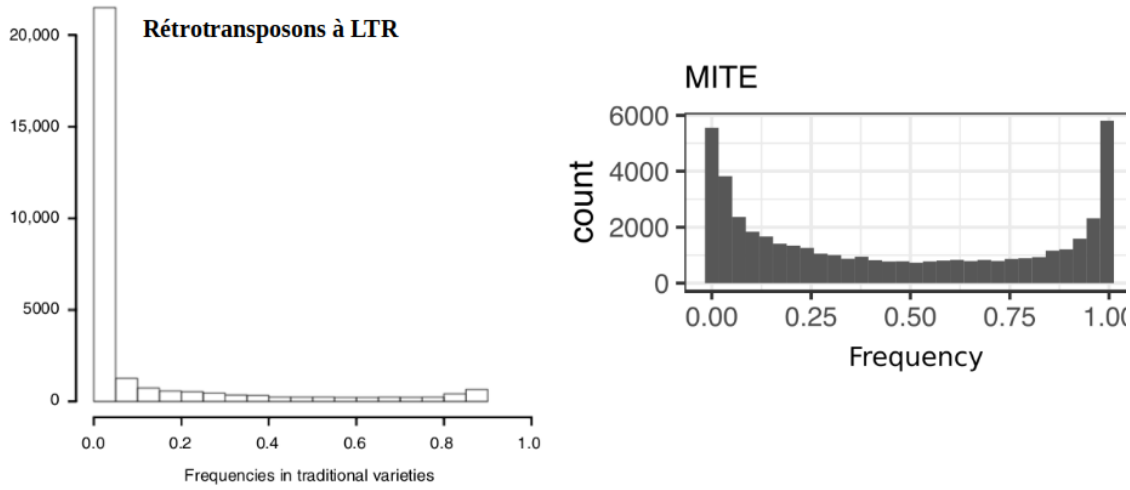


FIGURE 6.1 – **Distribution de la fréquence des TIP (*Transposable element Insertion Polymorphism*) au sein des variétés de riz issues des 3,000 génomes**. D'après CARPENTIER *et al.* 2019 et CASTANERA *et al.* 2021 respectivement. Pour les rétrotransposons, on observe que les TIP (*Transposable element Insertion Polymorphism*), en majorité, sont présents à très faible fréquence au sein des variétés traditionnelles. Pour les MITE, on observe qu'on a également une grande proportion de TIP présents chez seulement quelques variétés de riz. Par contre, certaines insertions sont présentes chez toutes les variétés. Ces dernières sont donc conservées entre les différentes variétés et suggère une sélection positive.

Comme ce que nous avons observé, de nombreuses copies de MITEs sont trouvées en faible fréquence au sein des 3,000 génomes : ce qui suggère une activité récente ou une élimination efficace. Par contre, certaines copies de MITE sont retrouvées au sein de toutes les variétés : elles sont fixées au sein de la population. Contrairement aux rétrotransposons à LTR, la distribution de la fréquence d'insertion n'est pas en forme de « L » mais plus en forme de « U ». Cela suggère que ces insertions à haute fréquence sont sous sélection positive, jouant un rôle important dans l'évolution du génome du riz (CASTANERA *et al.* 2021).

CHAPITRE 7

Utilisation de la technologie de séquençage longues lectures pour les analyses structurales des génomes

7.1 Contexte

Sauf cas exceptionnel, il est rare de visualiser phénotypiquement l'impact d'une insertion d'élément transposable : d'avoir un nouveau phénotype dû à l'insertion d'un ET, qui sera considéré comme variation causale de ce caractère spécifique. C'est le cas pour les insertions d'ET qui modifient la couleur du fruit ou de la plante (décrit dans l'introduction) comme chez la vigne, l'orange ou même l'orchidée (Hsu *et al.* 2020).

Au sein des 3,000 génomes de riz cultivé, une fois l'inventaire des polymorphismes liés à l'activité des rétrotransposons réalisé, se pose la question de l'impact fonctionnel que peuvent avoir ces polymorphismes, et donc leur contribution à la diversité génétique et phénotypique de cette plante cultivée.

Roland Akakpo, post-doctorant au laboratoire de 2018 à 2019, grâce à son expertise des approches de génomique d'association (*GWAS*), a ainsi développé un nouveau concept - le TE-GWAS - permettant de rechercher des associations entre une insertion d'ET et un phénotype. En exploitant la base de données de phénotypage de l'IRRI (*International Rice Research Institute*) pour les 3,000 génomes de riz, plusieurs phénotypes ont été analysés. Par la structure très distincte entre les variétés *Indica* et *Japonica*, il a été décidé d'effectuer ces analyses d'association sur ces deux populations de façon séparée. Pour améliorer la puissance des tests statistiques liés à ces études de GWAS, nous nous sommes focalisés, plus particulièrement, sur les variétés *Indica*, du fait de leur nombre plus important (N=1743) comparé aux variétés *Japonica* (N=839).

Nous avons pu identifier ainsi une association significative entre une insertion de

l'élément *rn215-125* et la largeur du grain. A noter que sur les 1743 variétés *Indica*, seulement 1132 ont été caractérisées pour ce trait. Les variétés de riz possédant cette insertion ont des grains significativement plus larges que celles qui ne l'ont pas (Figure 7.1). Ma contribution à ce travail a été de fournir toutes les données génomiques nécessaires aux analyses GWAS. Ce sont les sorties de mon pipeline TRACKPOSON.

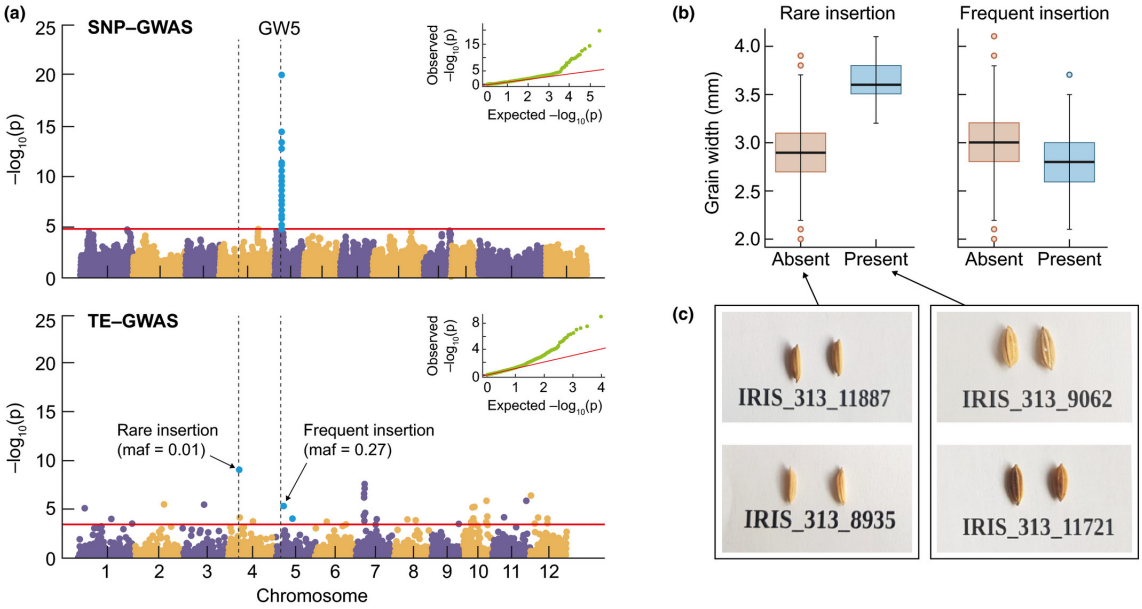


FIGURE 7.1 – **Étude d'association (GWAS) pour la largeur du grain**. Issue de **akakpo_impact_2020**. **(a)** Résultats GWAS (Manhattan et tracé quantile-quantile) détectés en utilisant GWAS classique avec les SNP et TE-GWAS pour la largeur du grain. Les points au-dessus de la ligne rouge (FDR=0.05) sont des candidats pour les insertions/SNP. Les points verts représentent les associations les plus significatives. Le panneau supérieur montre le SNP-GWAS en utilisant les SNP avec des fréquences alléliques mineures (MAF) > 0,1. Nous avons identifié un SNP candidat qui est probablement situé en aval du gène GW5, un gène majeur pour la forme du grain chez le riz (Weng et al., 2008).
Le panneau inférieur gauche montre notre résultat TE-GWAS en utilisant des TIPs avec des fréquences mineures (MF) > 0,01. Bien qu'il ne soit pas aussi saturé que la carte SNP, l'ensemble de données TIP a permis la détection d'associations significatives (points verts) à deux loci. L'un se trouve dans la région GW5 et a été détecté par une "ancienne" insertion du rétrotransposon à LTR *rn215-225*, trouvé dans 306 accessions. L'autre sur le chromosome 4, une insertion "récente" de la même famille, est trouvé dans seulement 12 accessions.
(b) Box plot de la largeur du grain pour les insertions candidates anciennes et récentes. Nous observons que les variétés qui possèdent l'insertion récente présentent un grain large (largeur moyenne du grain d'environ 3,4 mm) par rapport à celles qui ne possèdent pas l'insertion (largeur moyenne du grain d'environ 2,8 mm), ce qui suggère que cette insertion a un effet important sur la largeur du grain.
(c) Grains de riz larges (125831 et 127372) et étroits (127034 et 127434)

Ce travail a ainsi fait l'objet d'une publication dans New Phytologist en avril 2020.

Review

New Phytologist

Tansley insight

The impact of transposable elements on the structure, evolution and function of the rice genome

Author for correspondence:
Olivier Panaud
Tel: +33 646460372
Email: panaud@univ-perp.fr

Received: *2 September 2019*
Accepted: *5 November 2019*

Roland Akakpo[1], Marie-Christine Carpentier[1], Yue Ie Hsing[2] and Olivier Panaud[1,3]

[1]Laboratoire Génome et Développement des Plantes, UMR 5096 CNRS/UPVD, Université de Perpignan, Via Domitia, 52 Avenue Paul Alduy, 66860 Perpignan Cedex, France; [2]Institute of Plant and Microbial Biology, Acadeia Sinica, 128, Section 2, Yien-chu-yuan Road, Nankang 115, Taipei, Taiwan; [3]Institut Universitaire de France, 1 Rue Descartes, 75231 Paris Cedex 05, France

Contents

New Phytologist (2020) **226:** 44–49
doi: 10.1111/nph.16356

Key words: functional genomics, model crop species, rice, transposable elements (TEs).

Summary

Transposable elements (TEs) are ubiquitous in plants and are the primary genomic component of the majority of taxa. Knowledge of their impact on the structure, function and evolution of plant genomes is therefore a priority in the field of genomics. Rice, as one of the most prevalent crops for food security worldwide, has been subjected to intense research efforts over recent decades. Consequently, a considerable amount of genomic resources has been generated and made freely available to the scientific community. These can be exploited both to improve our understanding of some basic aspects of genome biology of this species and to develop new concepts for crop improvement. In this review, we describe the current knowledge on how TEs have shaped rice chromosomes and propose a new strategy based on a genome-wide association study (GWAS) to address the important question of their functional impact on this crop.

I. Introduction

Transposable elements (TEs) are ubiquitous components of eukaryotic genomes. In flowering plants, their contribution to genome size variation has been documented in many studies, and it is now clearly established that transposition is the main factor responsible for such variation, besides polyploidy (Bennetzen & Wang, 2014). Fig. 1 shows the distribution of genome size for 6000 plant species (https://cvalues.science.kew.org/). Considering that the gene space in diploid genomes occupies *c.* 100–200 Mbp, the large extent of variation in this distribution, the peak observed at *c.* 700 Mbp, and the median at *c.* 2 Gbp suggest that the majority of plant genomes are composed mostly of TEs or TE-related sequences. This has been confirmed by the many plant genome sequencing projects completed so far (see Chen *et al.*, 2018 for review). For instance, while < 10% of the small genome of the model species Arabidopsis is composed of TEs (AGI, 2000), TEs make up > 90% of the large genome of hexaploid wheat (Wicker

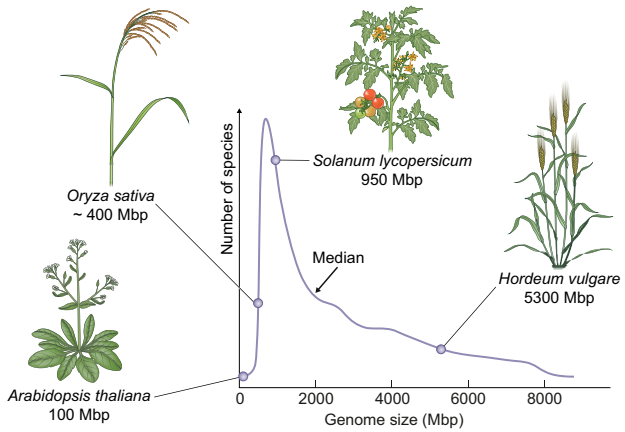


Fig. 1 Distribution of genome size for 6000 plant species. Data obtained from the Plant C-value database (https://cvalues.science.kew.org/). [Correction after first publication 8 January 2020; Fig. 1 has been amended.]

et al., 2018). The dynamics of the process through which TEs actually shape plant chromosomes are not so well understood. A comparative study among several plant genomes showed that most TE-related sequences arose from recent insertions (within the last 2 Myr, El Baidouri & Panaud, 2013) suggesting that older ones were eliminated from the genomes and therefore that there should exist a force counterbalancing TE-driven genomic expansions, which we formulated earlier in an 'increase/decrease' model (Vitte & Panaud, 2005) and which was recently addressed more specifically through mathematical models (Dai *et al.*, 2018). The dynamics of such processes (i.e. that of genomic amplification vs TE elimination) are yet to be fully understood.

In addition, the ubiquitous nature of TEs in plant genomes raises the overarching question of their biological impact and how transposition at large could contribute to plant biodiversity. TEs have long been considered mutagenic agents that impede gene function upon insertion into coding sequences, as exemplified by their early discovery as the causal agent of pigmentation loss in maize kernel by McClintock (1953). This negative view has since been challenged by several reports showing that transposition may in fact benefit an organism by regulating gene expression either by providing alternative promoters or novel cis-acting regulatory sites, or by acting as epigenetic mediators (Mirouze & Vitte, 2014; Hirsch & Springer, 2017). Moreover, two reports in mammals implicate TE amplification in the emergence of important biological novelty (i.e. placental pregnancy (Lynch *et al.*, 2011) and innate immunity (Chuong *et al.*, 2016)) indicating that transposition can be an important driver of eukaryote evolution through the rewiring of gene networks resulting in novel traits. In crops, there are few examples for which TEs are implicated in the variation of agronomic traits, such as fruit shape in tomato (Xiao *et al.*, 2008), and red pigmentation in apple (Zhang *et al.*, 2019). However, the extent to which TE-driven genetic variation could be exploited agronomically remains to be established for all crops.

Rice is the staple food for billions of people among the world's poorest populations. This cereal has therefore been subjected to intense research efforts over recent decades in order to secure food production world-wide. Consequently, a considerable amount of resources and knowledge have been generated for this species for all aspects of the biology of the plant. In particular, a high quality physical-map based genomic sequence of the rice variety 'Nipponbare' was the first crop genome to be released (IRGSP, 2005). As of today, high quality genome assemblies have been made available for five additional rice varieties (Song *et al.*, 2018). The release of the genomic sequence for 3000 rice genomes was an important milestone for rice research, giving access to the diversity of a crop at an unprecedented scale and providing a new approach for accelerating gene discovery through association studies (Rice genome project, 2014). Finally, the release of 10 genome assemblies for several wild relatives of rice allows the study of the evolution of a plant genome over 15 Myr (Stein *et al.*, 2018). In addition to providing rice biologists with new, efficient tools with which to develop future varieties, these resources have allowed us to decipher some basic aspects of TE-driven genome dynamics in rice, from long-term evolutionary aspects to their contribution to the genomic diversity of the crop following its domestication.

In this article, we first provide a synthetic view of these advances and then propose a new strategy to investigate the biological impact of TEs in rice, taking advantage of the genomic resources available for this model crop species.

II. Rice transposable elements

As in any other plant species, the rice genome harbours most known types of TEs as defined by Wicker *et al.* (2007) – for example, LTR-retrotransposons (Hirochika *et al.*, 1996; Chaparro *et al.*, 2007), transposons (Panaud *et al.*, 2002), miniature inverted-repeat transposable elements (MITEs) (Jiang *et al.*, 2004), long interspersed elements (Komatsu *et al.*, 2003), short interspersed elements (Tsuchimoto *et al.*, 2008) and terminal-repeat retrotransposon in miniatures (Gao *et al.*, 2016). Altogether, TE-related sequences make up *c.* 40% of the rice genome (Jiang & Panaud, 2013). Interestingly, this contrasts with the scarcity of known active TEs in rice, where 'activity' is defined as their mobility under clearly defined and repeatable experimental procedures. These are three retrotransposons – *Tos17* (Hirochika *et al.*, 1996), *Lullaby* (Picault *et al.*, 2009) and *Karma* (Komatsu *et al.*, 2003) – that can be activated during *callus vitro* culture. Additionally, the transposon *nDart* can be activated when rice plants are treated with methylation inhibitor (Eun *et al.*, 2012), and *mPing* can be activated using laser irradiation (S. Li *et al.*, 2017). To date, there is no other TE family among the hundreds that the rice genome harbours for which such direct evidence of activity has been established. However, *Oryza* genomes harbour traces of strong transpositional activity for many families (Copetti *et al.*, 2015). This paradox can only be solved when the conditions of the activation of transposition *in natura* are fully elucidated.

III. TE activity in *Oryza* lineage

The first comparative genomic analyses in the genus, focusing on several key loci such as *ADH* (Ammiraju *et al.*, 2007), *Hd1* (Sanyal *et al.*, 2010) and *Monoculm 1* (Lu *et al.*, 2009) revealed a good

New Phytologist

Box 1 The rationale for a transposable element (TE)-genome-wide association study (GWAS).

Since the late neolithic and development of agriculture some 10,000 years ago, farmers have moved many crops outside their centers of origin to new and often extreme environments.

Only crops that could adapt to these adverse conditions were retained in production.

In this regard, germplasm collections and in particular traditional varieties represent a reservoir (mostly untapped) of genetic factors involved in adaptation.

↓

TEs are known to be activated upon stress. Our model posits that transposition was triggered during the early phase of crop dissemination due to the stresses of domestication and adaptation to new environments.

Some TEs (e.g. the LTR-retrotransposons) harbour promoters that can be stress-inducible.

On the other hand, other TEs, such as MITEs, have been found to frequently contain transcription factor binding sites, which in some cases can wire new genes into stress-regulatory networks.

Thus, while grown in new and adverse conditions, crops accumulated new copies of TEs, thereby spreading new stress-inducible promoters throughout their genomes.

↓

In some instances, the genomic amplification of new regulatory sequences **may have lead to the re-programming of gene networks, resulting in new phenotypes.**

We anticipate that some TE insertions are causative of adaptive traits and may not be tagged by SNPs, due to recent movement, and, thus, can **only be detected using the TIPs** themselves as markers.

conservation of synteny among *Oryza* species, but with very limited correspondence in intergenic regions. A close examination of these loci confirmed that these regions were mostly comprised of TEs. These early reports therefore suggested that the rate of TE-driven genomic turnover was high, with effects observable at a moderate evolutionary timescale (< 3 Myr). Such turnover could only arise from the combined effect of successive waves of transposition that would quickly be eliminated from the genome through deletion or recombination, as posited by the increase/decrease model mentioned in the Introduction section. This was further illustrated by several genomic studies: Piegu *et al.* (2006) and Ammiraju *et al.* (2007) indeed showed that large and rapid genome size increases in *Oryza australiensis* and *Oryza granulata*, respectively, could be accounted for by the activity of LTR-retrotransposons. Interestingly, such dramatic genomic invasions occurred following speciation in both lineages, which provides indirect evidence that TEs may contribute to a large extent to lineage-specific genomic differentiation. This interpretation was further supported by the comparative analysis of eight A-genome *Oryza* species showing differential transpositional activity of several families among lineages within the last million years (Zhang & Gao, 2017). More recently, the sequencing and assembly of the genome of 10 relatives of rice (Stein *et al.*, 2018) provided a better view of the evolutionary fate of TE-related sequences: a comparative survey of orthologous insertions of LTR-retrotransposons among the species of the A-genome type (i.e. the most closely related to cultivated rice) allowed estimation of the elimination rate of TE-related sequences through deletions. This estimation translated into a half-life of 1.2 Myr, which is much faster than what had been established earlier for *Drosophila* (14 Myr) and mammals (> 800 Myr). Together, these studies provide an explanation for earlier observations by providing an estimation of the parameters of the increase/decrease model. One consequence of this is that TE-driven genomic turnover in the *Oryza* lineage occurs at such high rate that it may lead to genomic diversification within species. When it comes to rice, this may be of primary importance if one considers that TE-associated structural variation may have a biological impact that could be exploited if at least part of it could be agronomically favorable, as we will show in the fifth section of this article.

IV. TE activity in cultivated rice

As mentioned above, only a few TE families were shown to be transpositionally active in cultivated rice. However, earlier studies clearly showed that TEs could contribute to the genetic diversity of the crop; for example, the mPing MITE was found to be polymorphic between 'Indica' and 'Japonica' varieties (Jiang *et al.*, 2003). The release of genomic sequences of 3000 rice varieties (Rice Genome Project, 2014) provided a unique opportunity to tentatively estimate at what level TEs actually contribute to the genomic diversity of the crop in the recent past. The identification of TE insertion polymorphisms (TIPs) from large datasets requires the development of suitable software that is fast and has a low false discovery rate (FDR). Several methods have been recently developed for this purpose. RelocaTE2 (Chen *et al.*, 2017) is based on parsing mapping files of paired-read Illumina sequences against a reference genome. It was used recently to track the transpositional activity of the mPing MITE in the 3000 rice genomes (Chen *et al.*, 2019). Carpentier *et al.* (2019) developed a new software package, Trackposon, that was used to identify retrotransposon TIPs among the 3000 rice genomes (Rice genome project, 2014). This method is a two-step procedure consisting of first mapping the Illumina reads onto a consensus TE sequence and then mapping the un-mapped paired reads onto the reference rice genome. This method was found to be very fast, reliable and particularly well suited to large datasets such as the 3000 rice genomes. The authors showed that the three retrotransposon families that are known to be active *in vitro* (i.e. *Tos17*, *Lullaby* and *Karma*) are polymorphic in the cultivated gene pool, thus confirming that these TEs may in fact be active *in planta* and *in agro*. Surprisingly, the 32 TE families that were investigated in that study are also polymorphic. This may be a result of lineage sorting of polymorphisms that existed prior to domestication in the wild populations of rice progenitors, or of the fact that these families have been transpositionally active since domestication. In fact, the latter possibility is more parsimonious since most of the TIPs

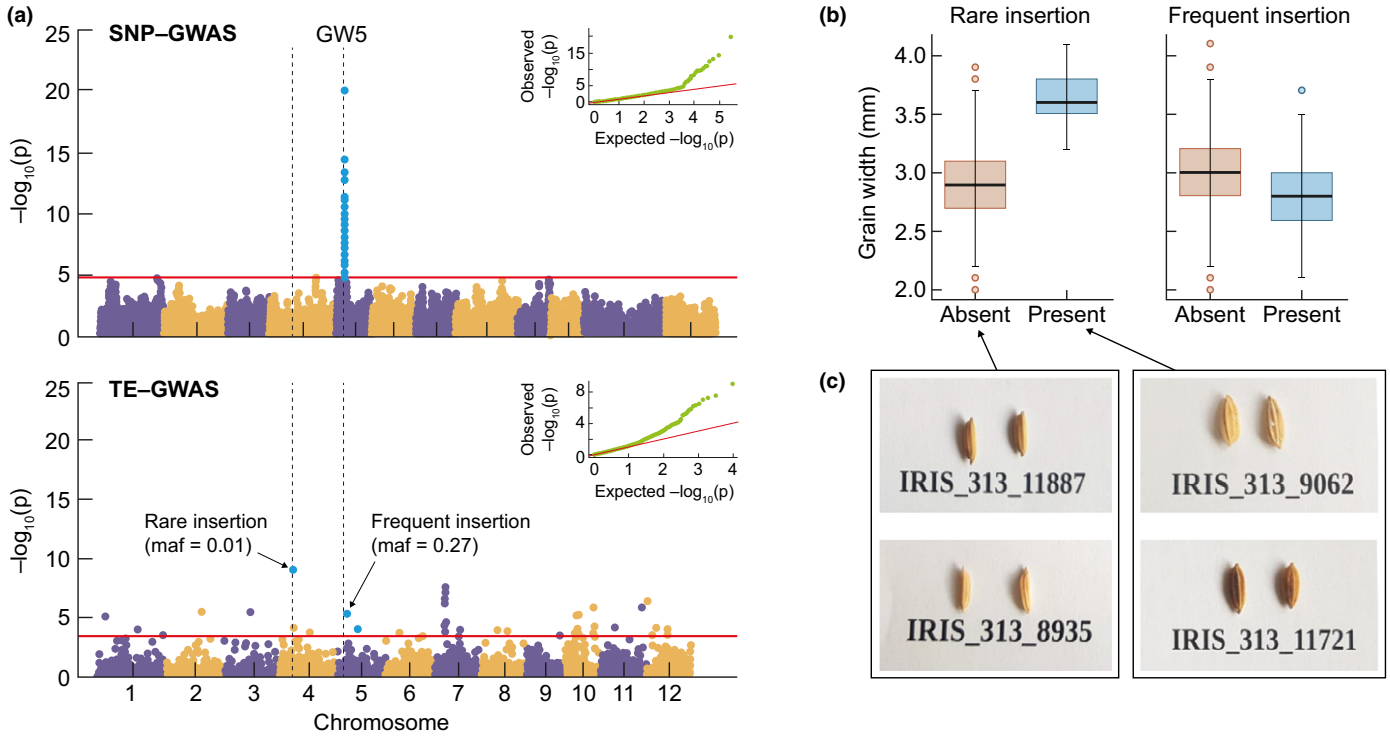


Fig. 2 Transposable element (TE)-genome-wide association study (GWAS) for grain width. (a) GWAS results (Manhattan and quantile–quantile plot) detected using SNP-GWAS and TE-GWAS for grain width. The solid red line represents FDR = 0.05. The dots above the red line are candidates for insertions/SNPs. The green dots represent the most significant associations. The top panel in (a) shows the SNP-GWAS using SNPs with minor allele frequencies (MAF) > 0.1. We identified a candidate SNP that is probably located downstream of the *GW5* gene, a major gene for grain shape in rice (Weng *et al.*, 2008). The bottom left panel in (a) shows our TE-GWAS result using TIPs with minor frequencies (MF) > 0.01. Although it is not as saturated as the SNP map, the TIP dataset enabled the detection of significant associations (green dots) at two loci. One is in the *GW5* region and was detected by an 'old' insertion of the LTR retrotransposon *rn215-225*, found in 306 accessions. The other one on chromosome 4, a 'recent' insertion of the same family, is found in only 12 accessions. (b) Box plot of grain width for the old and recent candidate insertions. The boxes represent the grain width variation between the first and third quartiles. The two whiskers represent the variation between the 'minimum' (Q1 – 1.5(interquartile range(IQR))) and the 'maximum' (Q3 + 1.5(IQR)). Dots represent outliers. We observe that the varieties that hold the recent insertion exhibit a large grain (mean grain width *c.* 3.4 mm) compared to those without the insertion (mean grain width *c.* 2.8 mm), suggesting that this insertion has a strong effect on grain width. (c) Large (IRIS_313_9062; IRIS_313_11721) and narrow (IRIS_313_8935; IRIS_313_11887) rice grains.

identified are shared by very few rice varieties, most being private. More recently, Fuentes *et al.* (2019) conducted an exhaustive search of structural variations (SVs) among the 3000 rice genomes. For this, the authors benchmarked available software programs, combining several of them into a single pipeline that yielded over 63 million SVs of various origins, 8.7% of which were found to be TE-related. One conclusion of these studies is that transposition did indeed contribute to the diversity of cultivated rice. The next task is to investigate the putative biological impact of these polymorphisms.

V. TE-GWAS – a new strategy to unravel the functional impact of TEs at species level

Until recently, gene discovery in crops relied on the use of map-based cloning (Tanksley *et al.*, 1995). However, the advent of multi-genome sequencing projects has led to the development of new strategies based on genome-wide association surveys (Atwell *et al.*, 2010). In rice, the availability of such resources has fueled many such projects (Si *et al.*, 2016; X. Li *et al.*, 2017; Sales *et al.*, 2017), demonstrating the efficiency of association studies in the crop. Given the availability of TIP data from the 3000 rice genomes for several TE families (Carpentier *et al.*, 2019), we propose a new strategy based on a genome-wide association survey in an attempt to unravel TE-associated genetic factors involved in agronomic traits in rice. This strategy is built on the rationale described in Box 1.

An important aspect of the TE-GWAS procedure is the risk of false discovery because of low allele frequency. The efficiency of GWAS is such that it is usually able to detect associations between traits and alleles with a minimum allele frequency (MAF) of at least 5% (Willard, 2013). However, the majority of TIPs in rice are found at a much lower frequency (Carpentier *et al.*, 2019). Consequently, we had to adapt our method to this peculiarity of TE insertions. First, we applied a correction for genetic structure (Thornsberry *et al.*, 2001). Next, we implemented four validation criteria in order to limit the risk of false positive detection: application of a stringent threshold of an FDR < 1% for significance, followed by the selection of the top 10% most significant candidate insertions; selection of insertions in genic regions (excluding TIPs in TE-rich, pericentromeric regions, for example); selection of insertions with

New Phytologist

significant effects on the trait, contrasting the phenotypes of the varieties without the insertion against those with the insertion; selection of recent insertions, discarding those in the same regions as the ones detected with SNP-GWAS.

VI. A case study: GWAS of LTR-retrotransposon insertions associated with grain width in rice

We implemented TE-GWAS in rice using TIPs from the *RN_215* LTR-retrotransposon family among 1132 accessions of *Oryza sativa* 'Indica'. The association study was performed using grain width as the phenotype (http://snp-seek.irri.org/). The results are shown in Fig. 2. Interestingly, a major quantitative trait locus (QTL; GW5) that was previously characterized through map-based cloning (Weng *et al.*, 2008) can be identified through classical GWAS (i.e. using SNP data as genotype; Fig. 2a). The use of TE-GWAS allowed us to identify a significant negative correlation between TE insertion and grain width (Fig. 2a,b) on chromosome 5 at the *GW5* locus. This insertion is found at high frequency (0.27) in the rice population and is probably in linkage disequilibrium with the genetic factor that causes the trait variation. In this regard, the TIP on chromosome 5 can be used as a genetic marker, like the other SNPs from the same region that show a significant association with the trait. More interestingly, another insertion found on chromosome 4 is highly significant (Fig. 2a) and is positively correlated with grain width (Fig. 2b). There is no association between any SNP from the same region and the trait, suggesting that the insertion is recent, which is supported by the low frequency of the insertion (0.01). Although this result requires confirmation through wet-lab experiments (e.g. an expression study of nearby genes throughout panicle development and CRISPR-Cas9 removal of the *RN_205* insertion from large seeded accessions), it nevertheless suggests that recent TE insertions may have led, to some extent, to the diversification of agronomic traits in rice.

VII. Conclusion

The availability of genomic resources for large germplasm collections opens a new perspective that can help us to understand some basic aspects of plant genome dynamics, as well as for the development of new strategies for gene discovery and crop improvement. Transposable elements, although ubiquitous in eukaryotes, have long been overlooked as important factors in genetic variation that could be beneficial for crops. The development of new strategies of genome-wide association surveys may help unravel some of the hidden heritability for major agronomic traits. In this context, rice, with unprecedented resources, will pave the way to the development of such new strategies.

Acknowledgements

This work was supported by Institut Universitaire de France, Agropolis Fondation (grant no. CFP 2015-02) and Academia Sinica thematic project AS-TP-107-L02. This study was conducted within the framework of the Laboratoires d'Excellences (LABEX) TULIP (ANR-10-LABX-41).

ORCID

Olivier Panaud https://orcid.org/0000-0002-9292-503X

References

AGI. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**: 796–815.

Ammiraju JSS, Zuccolo A, Yu Y, Song X, Piegu B, Chevalier F, Walling JG, Ma J, Talag J, Brar DS *et al.* 2007. Evolutionary dynamics of an ancient retrotransposon family provides insights into evolution of genome size in the genus *Oryza*. *The Plant Journal* **52**: 342–351.

Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y, Meng D, Platt A, Tarone AM, Hu TT *et al.* 2010. Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* **465**: 627–631.

Bennetzen JL, Wang H. 2014. The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annual Review of Plant Biology* **65**: 505–530.

Carpentier M-C, Manfroi E, Wei F-J, Wu H-P, Lasserre E, Llauro C, Debladis E, Akakpo R, Hsing Y-I, Panaud O. 2019. Retrotranspositional landscape of Asian rice revealed by 3000 genomes. *Nature Communications* **10**: 24.

Chaparro C, Guyot R, Zuccolo A, Piégu B, Panaud O. 2007. RetrOryza: a database of the rice LTR-retrotransposons. *Nucleic Acids Research* **35**: D66–D70.

Chen F, Dong W, Zhang J, Guo X, Chen J, Wang Z, Lin Z, Tang H, Zhang L. 2018. The sequenced angiosperm genomes and genome databases. *Frontiers in Plant Science* **9**: 418.

Chen J, Lu L, Benjamin J, Diaz S, Hancock CN, Stajich JE, Wessler SR. 2019. Tracking the origin of two genetic components associated with transposable element bursts in domesticated rice. *Nature Communications* **10**: 641.

Chen J, Wrightsman TR, Wessler SR, Stajich JE. 2017. RelocaTE2: a high resolution transposable element insertion site mapping tool for population resequencing. *PeerJ* **5**: e2942

Chuong EB, Elde NC, Feschotte C. 2016. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* **351**: 1083–1087.

Copetti D, Zhang J, El Baidouri M, Gao D, Wang J, Barghini E, Cossu RM, Angelova A, Maldonado LCE, Roffler S *et al.* 2015. RiTE database: a resource database for genus-wide rice genomics and evolutionary biology. *BMC Genomics* **16**: 538.

Dai X, Wang H, Zhou H, Wang L, Dvořák J, Bennetzen JL, Müller H-G. 2018. Birth and death of LTR-retrotransposons in *Aegilops tauschii*. *Genetics* **210**: 1039–1051.

El Baidouri M, Panaud O. 2013. Comparative genomic paleontology across plant kingdom reveals the dynamics of TE-driven genome evolution. *Genome Biology and Evolution* **5**: 954–965.

Eun C-H, Takagi K, Park K-I, Maekawa M, Iida S, Tsugane K. 2012. Activation and epigenetic regulation of DNA transposon nDart1 in rice. *Plant & Cell Physiology* **53**: 857–868.

Fuentes RR, Chebotarov D, Duitama J, Smith S, De la Hoz JF, Mohiyuddin M, Wing RA, McNally KL, Tatarinova T, Grigoriev A *et al.*. 2019. Structural variants in 3000 rice genomes. *Genome Research* **29**: 870–880.

Gao D, Li Y, Kim KD, Abernathy B, Jackson SA. 2016. Landscape and evolutionary dynamics of terminal repeat retrotransposons in miniature in plant genomes. *Genome Biology* **17**: 7.

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, USA* **93**: 7783–7788.

Hirsch CD, Springer NM. 2017. Transposable element influences on gene expression in plants. *Biochimica et Biophysica Acta (BBA) – Gene Regulatory Mechanisms* **1860**: 157–165.

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

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, 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.

Jiang N, Panaud O.2013. Transposable element dynamics in rice and its wild relatives. In: Zhang Q, Wing RA, eds. *Plant genetics and genomics: crops and models. Genetics and genomics of rice*. New York, NY, USA: Springer, 55–69.

Komatsu M, Shimamoto K, Kyozuka J. 2003. Two-step regulation and continuous retrotransposition of the rice LINE-type retrotransposon Karma. *Plant Cell* **15**: 1934–1944.

Li S, Xia Q, Wang F, Yu X, Ma J, Kou H, Lin X, Gao X, Liu B. 2017. Laser irradiation-induced DNA methylation changes are heritable and accompanied with transpositional activation of mPing in rice. *Frontiers in Plant Science* **8**: 363.

Li X, Guo Z, Lv Y, Cen X, Ding X, Wu H, Li X, Huang J, Xiong L. 2017. Genetic control of the root system in rice under normal and drought stress conditions by genome-wide association study. *PLoS Genetics* **13**: e1006889.

Lu F, Ammiraju JSS, Sanyal A, Zhang S, Song R, Chen J, Li G, Sui Y, Song X, Cheng Z *et al.* 2009. Comparative sequence analysis of MONOCULM1-orthologous regions in 14 Oryza genomes. *Proceedings of the National Academy of Sciences, USA* **106**: 2071–2076.

Lynch VJ, Leclerc RD, May G, Wagner GP. 2011. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. *Nature Genetics* **43**: 1154–1159.

McClintock B. 1953. Induction of instability at selected loci in maize. *Genetics* **38**: 579–599.

Mirouze M, Vitte C. 2014. Transposable elements, a treasure trove to decipher epigenetic variation: insights from Arabidopsis and crop epigenomes. *Journal of Experimental Botany* **65**: 2801–2812.

Panaud O, Vitte C, Hivert J, Muzlak S, Talag J, Brar D, Sarr A. 2002. Characterization of transposable elements in the genome of rice (*Oryza sativa* L.) using Representational Difference Analysis (RDA). *Molecular Genetics and Genomics* **268**: 113–121.

Picault N, Chaparro C, Piegu 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. *The Plant Journal* **58**: 754–765.

Piegu B, Guyot R, Picault N, Roulin A, Saniyal A, Kim H, Collura K, Brar DS, Jackson S, Wing RA *et al.* 2006. Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Research* **16**: 1262–1269.

Rice Genome Project. 2014. The 3,000 rice genomes project. *GigaScience* **3**: 7.

Sales E, Viruel J, Domingo C, Marqués L. 2017. Genome wide association analysis of cold tolerance at germination in temperate japonica rice (*Oryza sativa* L.) varieties. *PLoS ONE* **12**: e0183416.

Sanyal A, Ammiraju JSS, Lu F, Yu Y, Rambo T, Currie J, Kollura K, Kim H-R, Chen J, Ma J *et al.* 2010. Orthologous comparisons of the Hd1 region across genera reveal Hd1 gene lability within diploid *Oryza* species and disruptions to microsynteny in Sorghum. *Molecular Biology and Evolution* **27**: 2487–2506.

Si L, Chen J, Huang X, Gong H, Luo J, Hou Q, Zhou T, Lu T, Zhu J, Shangguan Y *et al.* 2016. *OsSPL13* controls grain size in cultivated rice. *Nature Genetics* **48**: 447–456.

Song S, Tian D, Zhang Z, Hu S, Yu J. 2018. Rice genomics: over the past two decades and into the future. *Genomics, Proteomics & Bioinformatics* **16**: 397–404.

Stein JC, Yu Y, Copetti D, Zwickl DJ, Zhang L, Zhang C, Chougule K, Gao D, Iwata A, Goicoechea JL *et al.* 2018. Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus *Oryza*. *Nature Genetics* **50**: 285–296.

Tanksley SD, Ganal MW, Martin GB. 1995. Chromosome landing: a paradigm for map-based gene cloning in plants with large genomes. *Trends in Genetics* **11**: 63–68.

Thornsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen D, Buckler ES. 2001. Dwarf8 polymorphisms associate with variation in flowering time. *Nature Genetics* **28**: 286–289.

Tsuchimoto S, Hirao Y, Ohtsubo E, Ohtsubo H. 2008. New SINE families from rice, OsSN, with poly(A) at the 3′ ends. *Genes & Genetic Systems* **83**: 227–236.

Vitte C, Panaud O. 2005. LTR retrotransposons and flowering plant genome size: emergence of the increase/decrease model. *Cytogenetic and Genome Research* **110**: 91–107.

Weng J, Gu S, Wan X, Gao H, Guo T, Su N, Lei C, Zhang X, Cheng Z, Guo X *et al.* 2008. Isolation and initial characterization of GW5, a major QTL associated with rice grain width and weight. *Cell Research* **18**: 1199–1209.

Wicker T, Gundlach H, Spannagl M, Uauy C, Borrill P, Ramírez-González RH, De Oliveira R, Mayer KFX, Paux E, Choulet F *et al.* 2018. Impact of transposable elements on genome structure and evolution in bread wheat. *Genome Biology* **19**: 103.

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**: nrg2165.

Willard HF. 2013. Chapter 1 – The human genome: a window on human genetics, biology, and medicine. In: Ginsburg GS, Willard HF, eds. *Genomic and personalized medicine* (*2nd edn*). London, UK: Elsevier/Academic Press, 4–27.

Xiao H, Jiang N, Schaffner E, Stockinger EJ, van der Knaap E. 2008. A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. *Science* **319**: 1527–1530.

Zhang L, Hu J, Han X, Li J, Gao Y, Richards CM, Zhang C, Tian Y, Liu G, Gul H *et al.* 2019. A high-quality apple genome assembly reveals the association of a retrotransposon and red fruit colour. *Nature Communications* **10**: 1494.

Zhang QJ, Gao LZ. 2017. Rapid and recent evolution of LTR retrotransposons drives rice genome evolution during the speciation of AA-genome *Oryza* species. *G3 (Bethesda)* **7**:1875–1885.

Très récemment, d'autres études similaires d'association d'ET-phénotype (TE-GWAS) ont été menées. Au sein de 602 variétés de tomates, il a été mis en évidence l'impact de 40 insertions d'ET sur la diversité phénotypique de la tomate (Domínguez *et al.* 2020).

Toutes ces analyses sont des analyses *in silico*, il est donc nécessaire et indispensable de valider nos résultats. Traditionnellement, une validation par *wet lab* est préconisée, telle que des designs d'amorces autour de l'insertion d'ET candidate suivi d'une amplification par PCR (Hsu *et al.* 2020). Nous avons décidé d'utiliser une approche alternative : la 3ème génération de technologie de séquençage, le séquençage longues lectures Nanopore. La validation structurale du projet via cette nouvelle génération de technologie de séquençage est la base pour envisager les validations *wet-lab*. Grâce à la grande longueur de ces lectures, il est possible de mettre en évidence les lectures contenant l'insertion d'un ET donné et donc de déterminer la position exacte de l'insertion au sein du génome de référence (Debladis *et al.* 2017). Il sera donc nécessaire de reséquencer avec cette nouvelle technologie, les génomes des variétés *Indica* qui présentent une taille de grains différente.

7.2 Analyse de la région d'intérêt

7.2.1 Homologie entre *Japonica* et *Indica*

L'association significative trouvée entre l'insertion de l'élément transposable *rn215-125* et la taille des grains chez les variétés *Indica*, est localisé au niveau du chromosome 4 entre la position 7,030,000 et 7,040,000 pb. Cette fenêtre de 10kb correspond à la sortie de mon outil TRACKPOSON, ainsi celle-ci se situe au sein de *Nipponbare*, le génome de référence *Japonica*.

En premier lieu, il faut donc détecter la région orthologue au niveau du génome de référence *Indica, IR64*. Pour cela, une région de 100kb de chaque côté de la fenêtre d'intérêt a été extraite au sein du génome *Nipponbare*. Puis un alignement par blast sur le génome *Indica* de référence *IR64* a été effectué pour trouver les positions orthologues sur le chromosome 4. Pour visualiser l'homologie entre ces 2 régions, un alignement par dotter a été réalisé (Figure 7.2).

Sur cette représentation, si les séquences sont identiques une diagonale sera tracée. Si l'identité entre les séquences diminue, alors celle-ci sera morcelée, non continue. Ainsi, par cette méthode (*dotter*), il est possible de connaître la ressemblance entre les séquences et également s'il y a eu des variations de structure, comme des évènements d'insertions et/ou délétions.

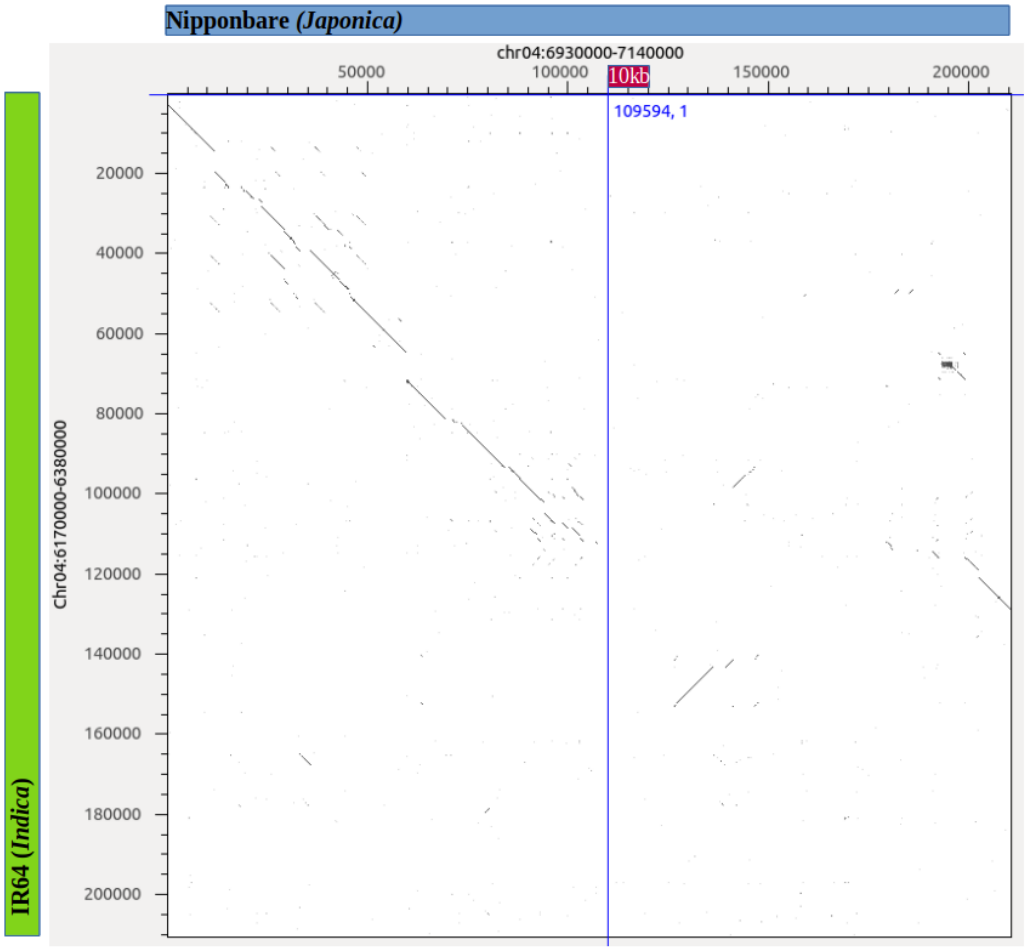


FIGURE 7.2 – **Alignement des régions orthologues entre *Nipponbare (Japonica) et IR64 (Indica)*.** En horizontal est représentée la région de 210kb (100kb autour de la fenêtre de 10kb de TRACKPOSON) issue du génome de *Nipponbare*. En verticale est représentée la région (de 210kb) orthologue d'*IR64*. On observe aucune homologie entre les deux régions.

Mais à notre grande surprise, il y a très peu, voir pas de similarité au niveau de cette région au sein de *IR64*. Ainsi on peut dire que cette région n'est pas homologue entre les deux variétés et que celle-ci est absente de l'*Indica* de référence. Cela peut donc suggérer une introgression de cette région issue de *Nipponbare (Japonica)* au sein des variétés *Indica* analysées en TE-GWAS.

7.2.2 Caractérisation génomique de la région

Nous avons donc décidé d'analyser en profondeur cette région du chromosome 4 autour de l'insertion de l'ET *rn215-125*, présente chez *Nipponbare*.
Il s'avère que cette région contient 4 paralogues de la famille multigénique de glucosyl transferase (GST, Os04g0204100). Cette famille GST est présente seulement au niveau du chromosome 4 : aucune autre copie de ces gènes n'a été identifié (par blast) sur les autres chromosomes de *Nipponbare* (Figure 7.3).

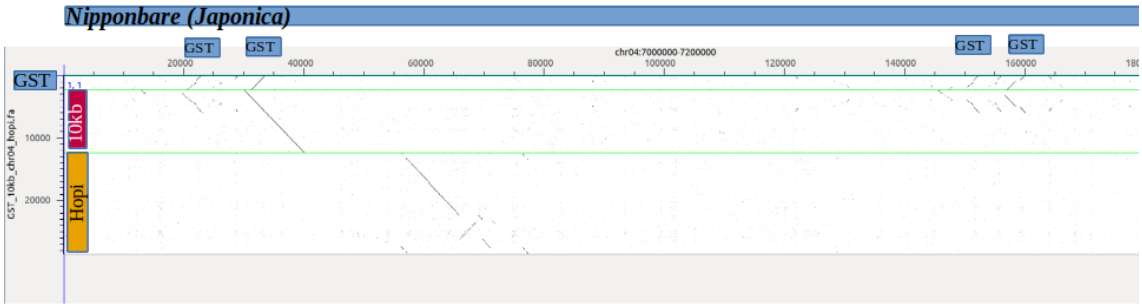


FIGURE 7.3 – **Alignement du gène de glucosyl transferase et de l'ET *Hopi* contre *Nipponbare (Japonica)*.** En vertical la région du chromosome 4 de *Nipponbare* et en verticale les séquences génomiques d'intêret (*GST, Hopi* et la région de 10kb). Dans la région d'intêret de *Nipponbare*, on observe 4 copies du gènes de glucosyl transferase et une insertion de l'ET *Hopi*.

De plus, nous avons observé une insertion d'*Hopi* au sein de cette région spécifique. En collaboration avec l'équipe de Yue-Ie Hshing de l'Academia Sinica à Taiwan, cette insertion a été validée par PCR, ce qui a permis de préciser le positionnement de l'insertion d'*Hopi* : celle-ci est positionnée au niveau 3'UTR du gène de GST (Figure 7.4).

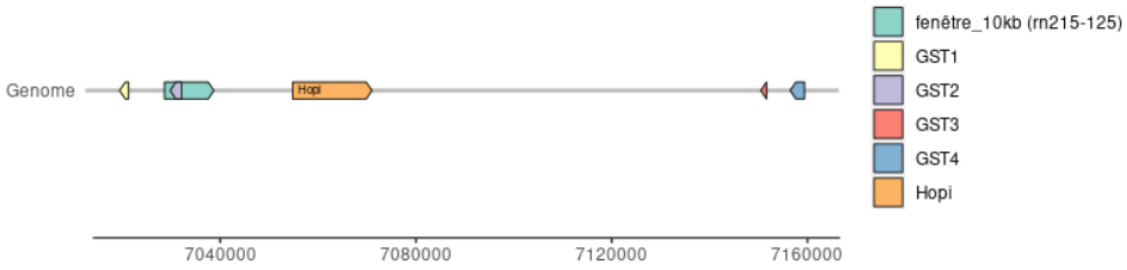


FIGURE 7.4 – **Schéma de la région d'intêret du chromosome 4 chez *Nipponbare***

A noter que cette insertion n'a pas été trouvée lors de l'analyse par TRACKPOSON. Mon pipeline ne peut identifier que les insertions d'ET au sein des régions uniques. Or,

dans la région d'intérêt par la présence des paralogues de GST, ceux-ci sont considérés comme des régions répétées et il est donc impossible pour TRACKPOSON de déterminer des insertions d'ET au niveau de cette région spécifique.

Ainsi, nous pouvons dire que cette région du chromosome 4 est une région dynamique et assez complexe.

7.2.3 Histoire transpositionnelle de la région candidate

La découverte de l'insertion de deux familles d'ET différents au sein de cette région nous amène à nous questionner sur l'histoire transpositionnelle de celle-ci. L'insertion d'*Hopi* est-elle arrivée en même temps que celle de *rn215-125*, ou celle-ci est arrivée avant, après ? Mais surtout est-elle causale du phénotype, sachant qu'elle est située dans le gène de glucosyl transferase.

Pour répondre à ces questions, j'ai ainsi modifié mon pipeline TRACKPOSON pour pouvoir identifier l'insertion d'*Hopi* au sein des paralogues de GST : je l'ai nommé TRACKPOSON inversé (Figure 7.5).

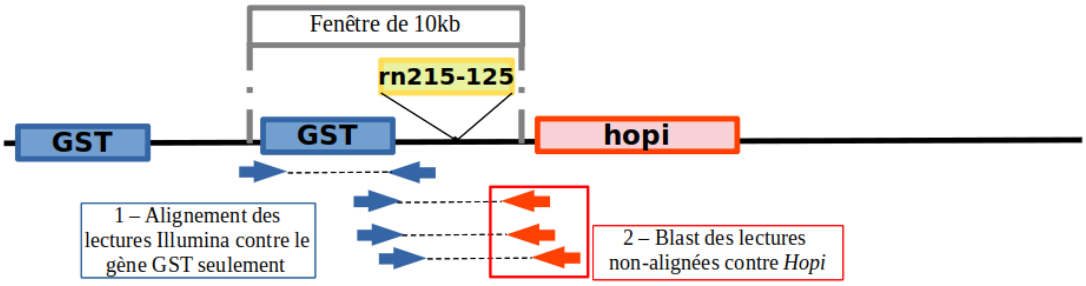


FIGURE 7.5 – **Méthode du TRACKPOSON inversé.** Cette méthode permet d'identifier les insertions d'ET (ici *Hopi*) au sein d'une région répétée définie.

Contrairement à la version classique qui aligne les lectures pairées contre une séquence d'ET, cette nouvelle version va dans un premier temps aligner les lectures contre la séquence consensus du gène de glucosyl transferase (GST).

Dans un second temps, comme pour la version classique, au sein des paires de lectures alignées/non-alignées, les lectures non-alignées sont récupérées. A la fin, un blast de ces séquences non-alignées contre la séquence consensus de l'ET *Hopi* est effectué. S'il y a un hit blast, l'insertion d'*Hopi* aux alentours du gène de GST dans cette région spécifique du chromosome 4 est déclarée.

En collaboration avec l'équipe YI. Hshing, j'ai ainsi appliqué mon nouveau pipeline sur les 3000 génomes de riz. Sur cette population, seulement 68 variétés ont l'insertion d'*Hopi* au sein du gène de GST. Sur ces 68 variétés, 47 (69 %) sont des variétés *Indica*.

Au sein des variétés *Indica*, analysées en TE-GWAS (N=11 322), 31 ont cette insertion spécifique d'*Hopi*. En croisant les résultats de TRACKPOSON (version normale) pour

rn215-125 et ceux du TRACKPOSON inversé pour *Hopi*, on remarque que sur ces 31 variétés avec l'insertion d'*Hopi*, 17 ont aussi une insertion de *rn215-125* au niveau de la région du chromosome 4. De plus, on remarque qu'il n'y a aucune variété qui a seulement l'insertion de *rn215-125*.

Ainsi par ces deux approches, on peut déclarer que l'insertion d'*Hopi* est antérieure à celle de *rn215-125* au niveau de cette région spécifique. De plus, les variétés qui ont ces deux insertions ont des grains significativement plus larges que les variétés avec une insertion d'*Hopi* seulement ou pas d'insertion (Figure 7.6).

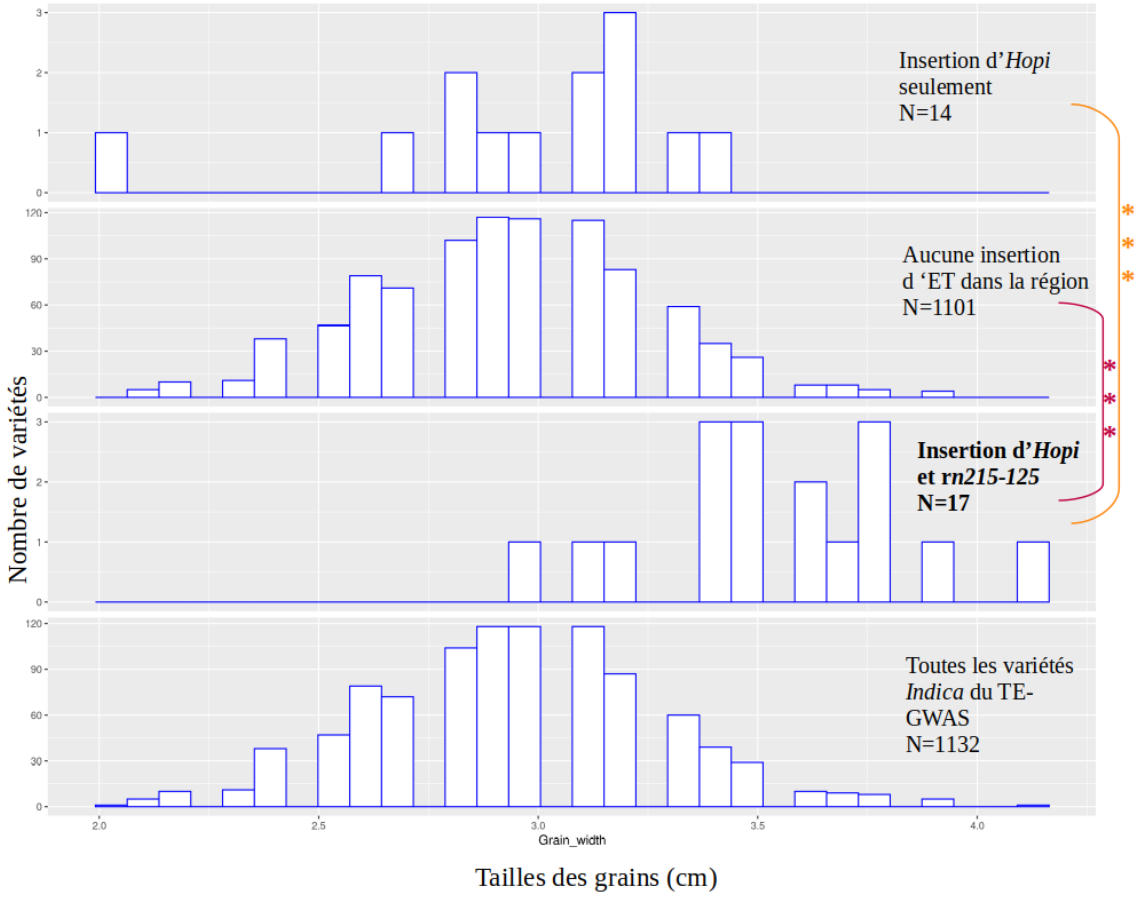


FIGURE 7.6 – **Distribution de la taille des grains pour les variétés *Indica* en fonction de la présence d'insertions de *rn215-125 et/ou Hopi*** . Les variétés ayant les 2 insertions d'ET (*rn215-125 et Hopi*) ont une largeur de grains significativement plus importante que celle sans insertion d'ET ou avec seulement une insertion d'*Hopi* (Wilcoxon test pvalue 1.6e-9 et 1.2e-4 respectivement).

Ainsi cela valide les résultats trouvés par l'analyse de TE-GWAS, faisant de cette région spécifique du chromosome 4 une région candidate pour l'étude de l'impact fonctionnel des ET.

A la suite de la validation des résultats de TE-GWAS, il reste encore une question à élucider : quelle est l'origine de cette région candidate ?

N'ayant pas d'homologie avec le génome de référence *Indica*, on peut penser que cette région est le fruit d'une introgression de *Japonica* au sein des variétés *Indica*.

7.3 Caractérisation d'une introgression

7.3.1 Séquençage Nanopore des variétés *Indica*

Pour confirmer (ou infirmer) cette hypothèse, les génomes de 5 variétés *Indica* à larges grains et portant l'insertion de *rn215-125* ont été séquencées avec la technologie Nanopore par le séquenceur MinIon.

Variétés	Taille des grains(mm)	Nb de Go	Origine
127 372	38	4,3	Thailande
127 582	38	1,7	Birmanie
128 076	39	2,6	Laos
128 087	**40**	**22**	Laos
128 104	38	8,3	Laos

Début 2021, mon équipe a acquis un ordinateur avec une carte GPU (Unité de Traitement Graphique) pour améliorer la détection des bases (*basecalling*) qui demandait un grand nombre d'heures de calcul sur un ordinateur CPU (Unité Centrale de Traitement). D'après nos estimations pour un jeu de données de 2 Go, la détection des bases aurait pris presqu'un mois complet de calcul. Ainsi avec le système GPU, les séquences issues des variétés ci-dessus ont été *basecallées* avec la dernière version du programme de Nanopore, appelé Guppy en mode *"High Accurancy"* (HAC). A noter que ce *basecalling* dure à peine quelques heures pour les plus gros jeux de données produits, montrant ainsi l'efficacité du système GPU (par rapport au CPU).

Pour la suite de cette partie, je me suis focalisée sur la variété 128 087, car celle-ci avait les plus gros grains et également le plus grand nombre de séquences produites.

7.3.2 Origine de l'introgression : *Japonica*?

Pour détecter l'origine de l'introgression, il est nécessaire d'analyser les régions flanquantes de la région de 10kb du chromosome 4 et d'identifier leur homologie avec des génomes de référence.

Ainsi, au préalable un assemblage *de novo* avec le logiciel Flye à partir des longues lectures Nanopore de la variété 128 087 a été effectué.

Statistiques	
Longueur totale	382,904,989 pb
Nombres de contigs	2,152
N50	752,396 pb
Plus long contig	3,419,478 pb
Couverture moyenne	21

La région candidate de 10kb issue de *Nipponbare* a été alignée contre les contigs assemblés, permettant ainsi de mettre en évidence le contig homologue de notre région candidate. C'est le contig_487 qui fait 529,685 pb. Ce contig_487 a été ensuite aligné avec la région de *Nipponbare* pour mettre en évidence l'homologie entre les 2 variétés et donc l'introgression (Figure 7.7). La région de *Nipponbare* correspondant aux régions flanquantes de 500kb autour de la fenêtre de 10kb : ainsi cette région fait au total 1010kb.

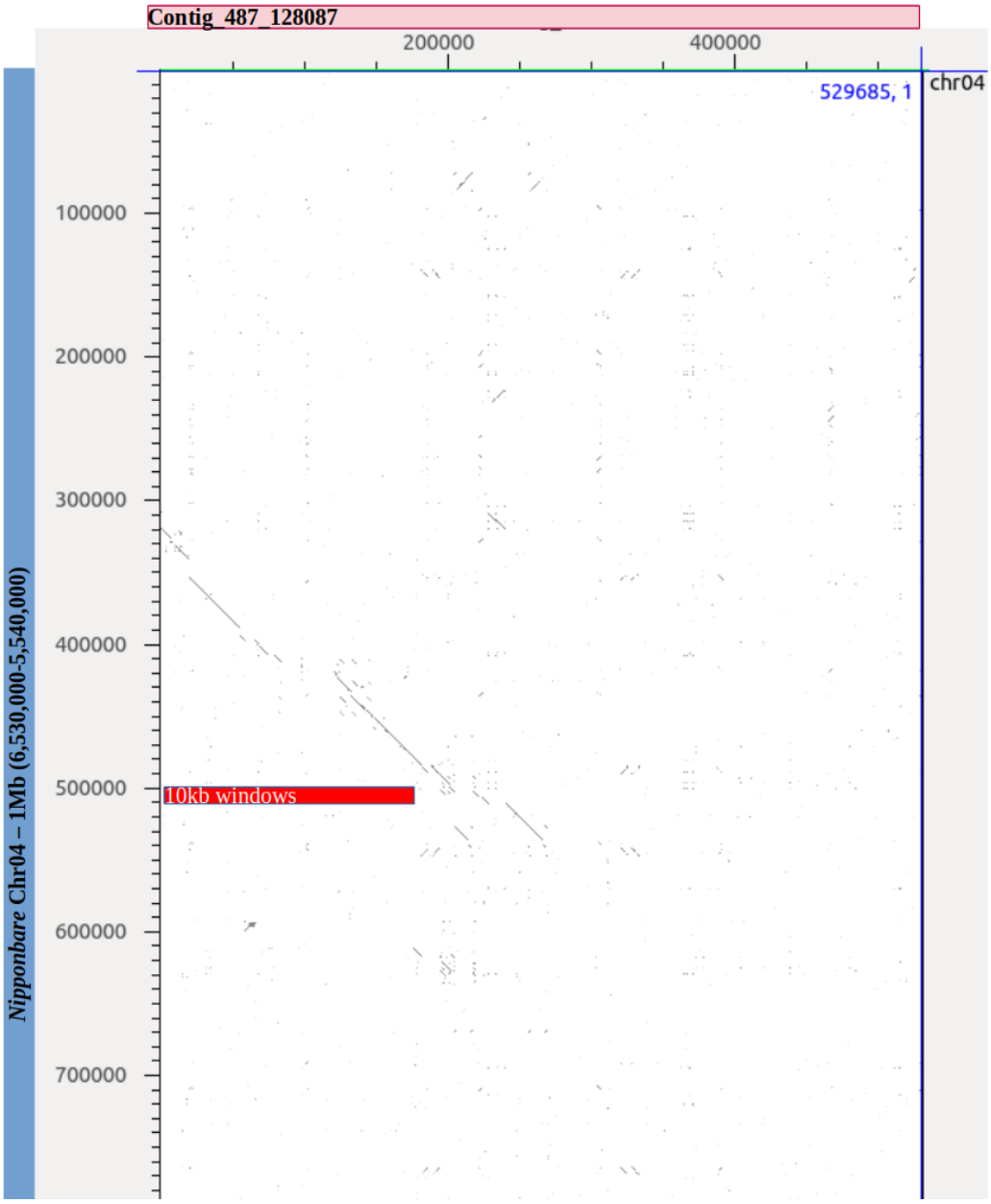


FIGURE 7.7 – **Alignement du contig_487 contre la région du chromosome 4**. En horizontal, le contig de la variété 128087 et en vertical la région du chromosome 4 autour de la fenêtre de 10kb issue de *Nipponbare (Japonica)*. Il y a très peu de similarité entre les deux séquences.

Contrairement à notre hypothèse de départ, la similarité entre le contig_487 et la région du chromosome 4 de *Nipponbare* est très faible. On a un profil identique à la comparaison entre les régions de *Nipponbare* et d'*IR64* (Figure 7.2).

Ainsi, on peut éliminer l'origine de *Japonica* pour l'introgression de cette région.

7.3.3 Comparaison avec les 12 génomes *Platinium*

En 2019, les génomes de 12 variétés de riz cultivé on été séquencés par la technologie PacBio et assemblés avec une grande qualité de résolution (Y. Zhou *et al.* 2020). Ces 12 génomes sont le reflet de la diversité des groupes variétaux au sein du genre *O.sativa* (Figure 7.8).

Pour chacun de ces 12 génomes, les positions orthologues de la fenêtre de 10kb du chromosome 4 chez *Nipponbare* ont été récupérées. Sur les 12 génomes, seulement 6 ont un résultat significatif en blast.

Nom de la variété	Groupes variétal	Longueur du hit blast (nt)
Arc 10497	*Aromatic*	7813
Cha Meo	*Japonica*	10000
Larha Mugad	*Indica*	7812
Khao Yau Guang	*Indica*	7812
pr106	*Indica*	7812
Ketan Nangka	*Japonica*	10000

Pour chacune de ces six variétés, une région d'environ 1Mb a été extraite : 500kb de part et d'autre de la fenêtre de 10kb. Ainsi un alignement du contig_487 issu du génome de la variété séquencée en Nanopore (128087) a été effectué contre ces 6 régions. La visualisation de la similarité entre ces séquences a été faite via dotter (Figure 7.9).

Aucune réelle homologie n'est observée : on a toujours le même profil avec une importante accumulation d'INDEL entre le contig_487 assemblé et les 6 régions issues des riz cultivés assemblés .

L'introgression de la région au sein de la variété *Indica* étudiée est bien présente mais l'origine de celle-ci reste un mystère : elle ne provient ni des variétés *Japonica* (Figure 7.7), ni des variétés *Indica* (Figure 7.9).

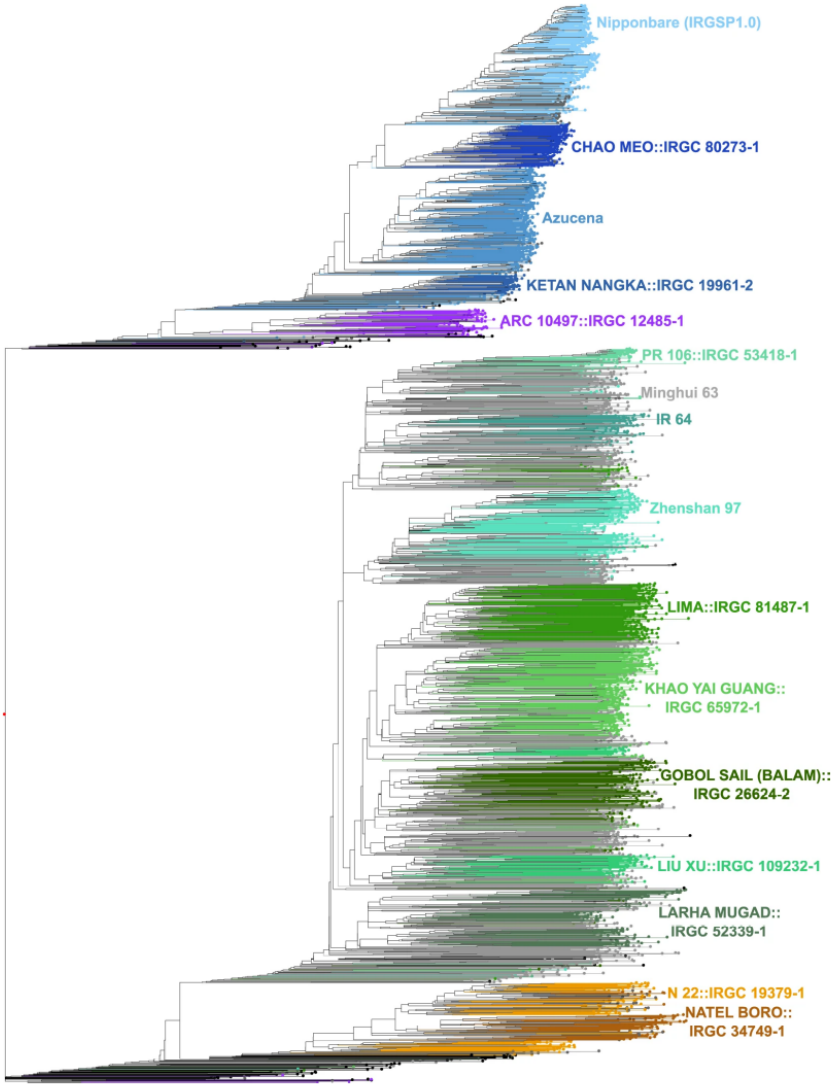


FIGURE 7.8 – **Arbre phénétique des variétés de riz cultivé séquencés en PacBio**. Figure 1 issue de la publication Y. ZHOU *et al.* 2020. Les groupes sont colorés selon l'assignation de l'analyse d'Admixture. Les variétés *Indica* sont en vert, en bleu sont colorées les varités *Japonica* et en marron les variétés *Aus/Boro*

Mais est-ce que cette variété est bien une variété *Indica* ?
Pour répondre à cette question, une autre région du génome a été sélectionnée : les mêmes positions que la région candidate (7,030,000-7,040,000) mais sur le chromosome 5. Le même protocole que pour la région du chromosome 4 a été appliqué : alignement par blast et récupération des coordonnées orthologues puis visualisation de l'homologie via dotter (Figure 7.10).

On observe une plus grande similarité de séquence au niveau de la région du chromosome 5 entre la variété Nanopore 128087 et les variétés *Indica*. Il y a moins d'INDEL

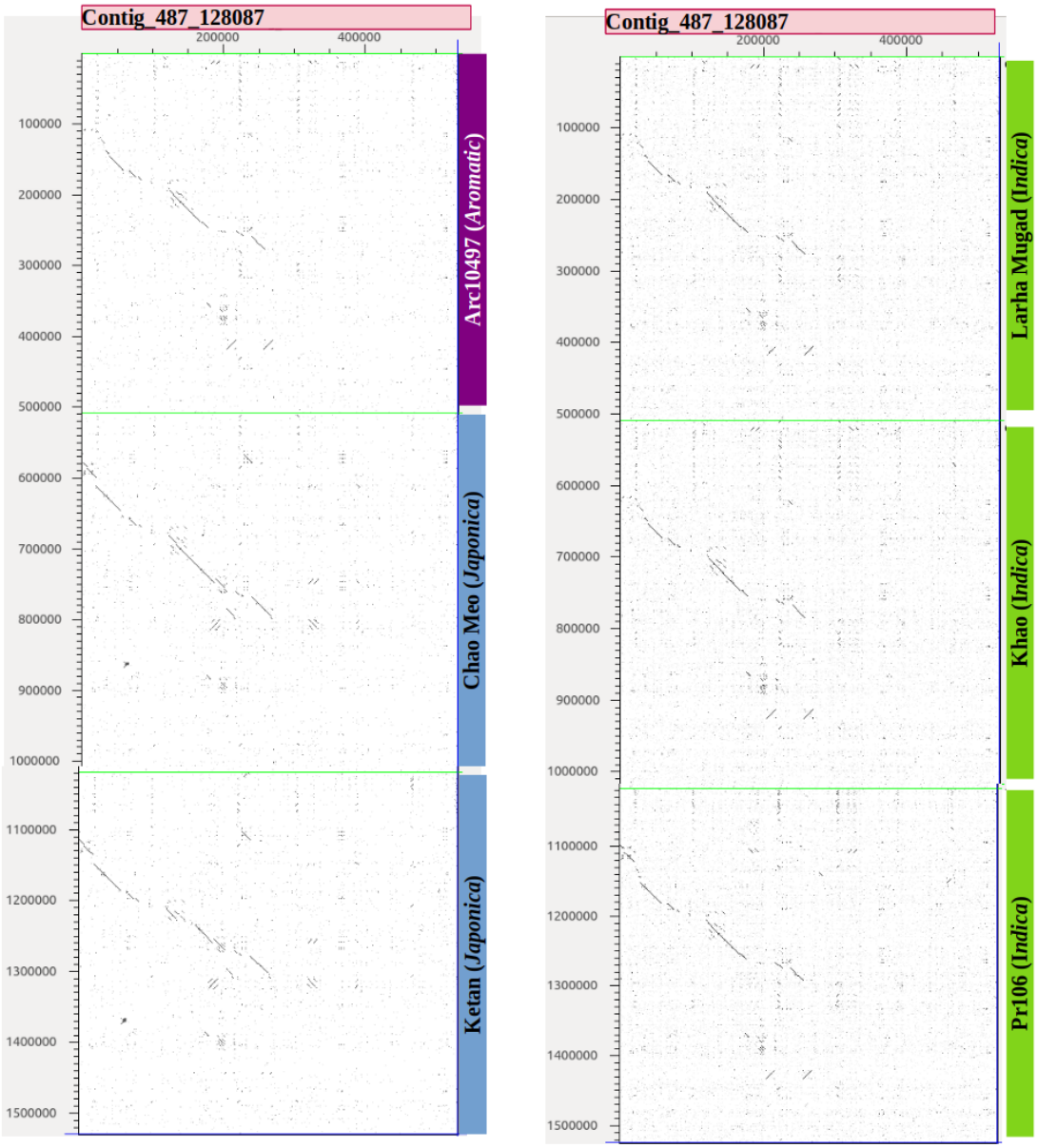


FIGURE 7.9 – **Alignement du contig_487 contre les régions des génomes Platinium**. En horizontal le contig de la variété 128087, en vertical les régions homologues de 510kb issues des 6 génomes assemblés en qualité *Platinium*. Aucune réelle homologie n'a été identifiée entre 128087 et les séquences issues des différentes variétés de riz cultivé.

comparé à l'alignement contre la région issue de Nipponbare (*Japonica*). De plus entre les deux variétés *Indica*, il semble il a voir une meilleure similarité entre le contig de la variété 128087 et la région de *Khao Yau Guang*. On observe une diagonale plus continue sur le dotter (Figure 7.10). Les variétés *Larha Mugad* et *Khao Yau Guang* sont des variétés *Indica* mais phylogénétiquement, elles appartiennent à des sous-groupes variétaux différents : *Indica2* et *Indica3*, respectivement (Figure 7.8).

Ainsi cela confirme le fait que la variété 128087 qui a été séquencée par Nanopore

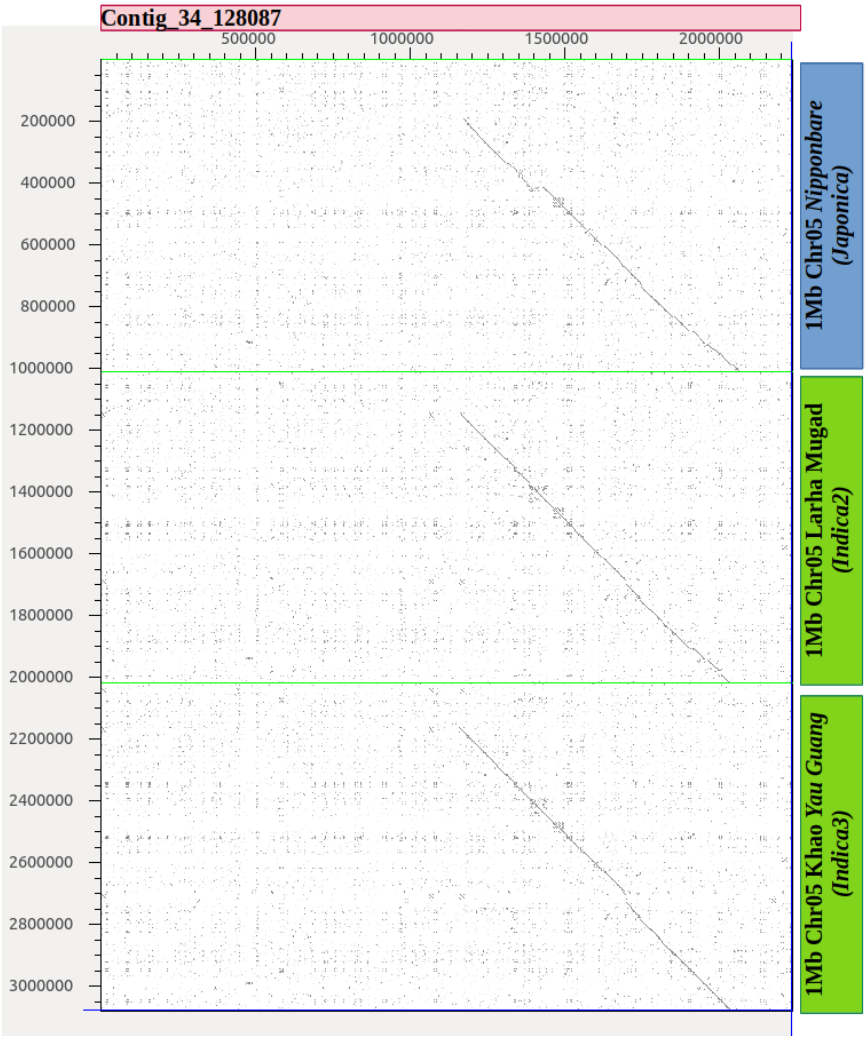


FIGURE 7.10 – **Alignement du contig_34 contre les régions du chromosome 5**. Alignement du contig_34 de la variété 128087 (en horizontal) contre la région homologue de 1Mb du chromosome 5 issue de Nipponbare (*Japonica*), celle de 1Mb de Larha Mugad (*Indica*) et de Khao Yau Guang (*Indica*). On observe un meilleur alignement entre le contig assemblé et les variétés *Indica*. On peut ainsi dire que la variété 128087 est bien une variété *Indica*.

est bel et bien une variété *Indica*. Et on peut même supposer que celle-ci appartiendrait au sous-groupe *Indica3*. Le doute sur la nature de la variété étant levé, l'origine de l'introgression de la région d'intérêt du chromosome 4 reste encore inconnue. Ayant peu de similarité avec les différentes variétés cultivées, peut être que cette région proviendrait de l'ancêtre commun du riz cultivé, *O. rufipogon*.

7.3.4 Origine de l'introgression : *O. rufipogon* ?

Au vu des différents alignements précédents contre les différents groupes variétaux de riz cultivé, il semblerait que la région d'introgression soit plus grande que celle qu'on

imaginait. On n'observe aucune homologie sur la droite des différents dotters produits, laissant supposer que l'introgression étudiée est donc plus grande que 500 kb.

En prenant en compte ces observations, tous les contigs assemblés de la variété Nanopore 128087 ont été alignés par minimap2 contre le chromosome 4 de la variété *Indica* assemblée *Khao Yau Guang*. Une région de 1,7Mb autour de la fenêtre de 10kb de TRACKPOSON a été définie, correspondant ainsi aux coordonnées suivantes 5,600,000-7,300,000. Les contigs de la variétés 128087 s'alignant sur cette région ont été récupérés.

Ces derniers ont ensuite été alignés avec minimap2 contre l'assemblage de la variété *O.rufipogon* disponible dans les bases de données génomiques. A noter que cet assemblage n'est pas sous forme de pseudo-molécules mais sous forme de contigs, qui sont au nombre de 2 582. Les contigs ayant des hits significatifs ont été conservés.

Puis une visualisation d'homologie entre les contigs de 128087 (comprenant le contig_487), la région de *Khao Yau Guang* et les contigs de *O.rufipogon* a été effectué (Figure 7.11).

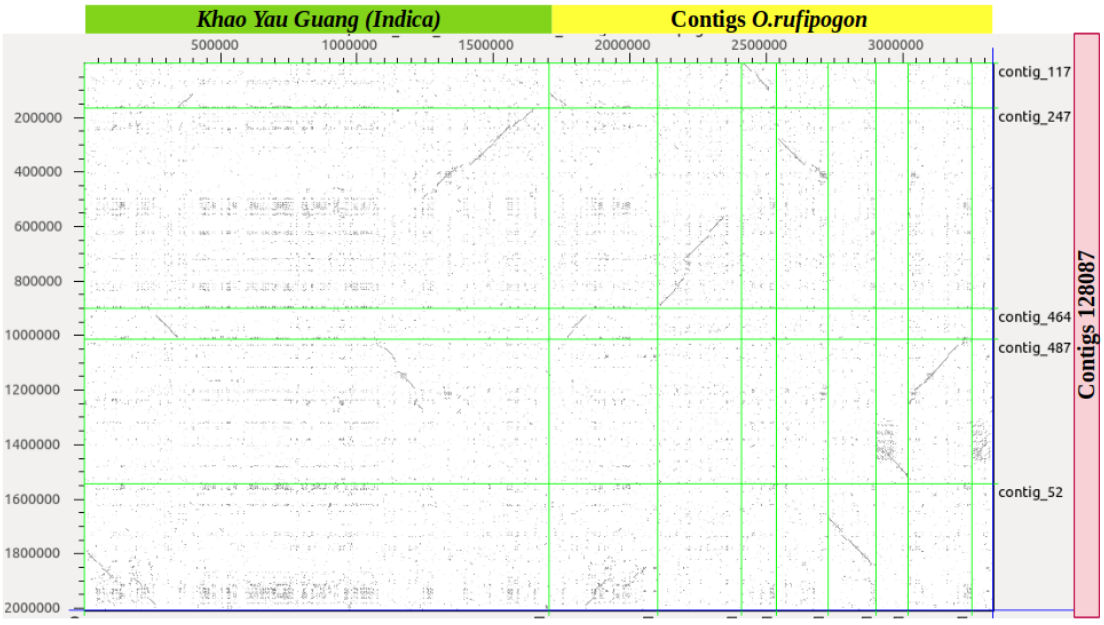


FIGURE 7.11 – **Alignement des contigs de la variété 128 087 contre les contigs de *O.rufipogon* et *Khai Yau Guang***. Alignement des contigs de la variété 128087 (en vertical) contre la région homologue de 1,7Mb de Khai Yau Guang (*Indica*) et les différents contigs issus de *O.rufipogon* (en vertical). On observe une plus grande identité de séquence entre les contigs assemblés de 128087 et les contigs de *O.rufipogon*. L'introgression semble donc provenir de cette variété sauvage et sa taille estimée serait de 800kb.

On observe une bien meilleure similarité de séquences entre le contig_487 de la variété 128087 avec *O.rufipogon*, comparée à tous les alignements effectués précédemment.

De plus, on observe bien les points de rupture d'homologie à gauche de l'introgression pour le contig_117 et à droite pour les contigs_247 et 487. Ainsi, on peut estimer la taille de cette introgression qui serait d'environ 800kb (Figure 7.12).

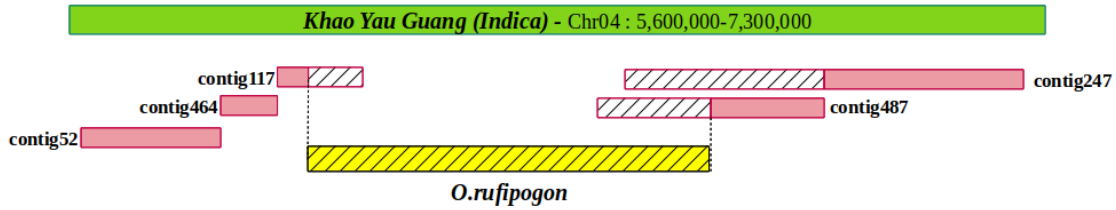


FIGURE 7.12 – **Schéma de l'alignement des contigs assemblés de 128087 contre la région du chromosome 4 de *Khao Yau Guang***. En vert est représentée la région du chromosome 4 de la variété *Indica, Khao Yau Guang*. En rose sont représentées les parties des contigs de 128087 qui s'alignent contre le génome de référence *Indica*. En hachuré sont représentées les parties des contigs de 128087 qui s'alignent contre les contigs de *O.rufipogon*. Par cet alignement, on peut estimer la taille de l'introgression issue de *O.rufipogon*, qui est d'environ 800kb.

Cette région complexe d'introgression semble provenir du riz *O.rufipogon*, ancêtre commun du riz cultivé *Japonica* et *Indica*. Ainsi, on observe une plus grande identité de séquences (ie diagonale plus continue) entre les contigs de la variété 128087 et la variété d'*O.rufipogon* (comparé à la variété *Indica, Khai Yau Guang*). Néanmois l'identité n'est pas parfaite : cela laisse supposer que la variété donneuse est bien une variété *O.rufipogon* mais que celle-ci n'est pas celle référencée dans les bases de données. Elle doit être sûrement proche phylogénétiquement de cette dernière.

7.4 Conclusions et Perspectives

En conclusion, avec l'analyse des données de 2ème (Illumina) et 3ème (Nanopore) génération de séquençage, nous avons pu mettre en évidence (confirmer) la présence d'une insertion d'un élément transposable, *rn215-125* et sa corrélation sur la largeur des grains au sein des variétés *Indica*. Cette insertion est située au niveau d'une région complexe du génome d'*Indica*. Il semblerait que cette région provienne d'une introgression d'une taille estimée de 800kb. L'origine de l'introgression pourrait être le riz ancestral, *O.rufipogon*.

Ce projet est actuellement en cours d'analyse avec des perspectives plus ou moins à long terme.

Dans un premier temps, il faudrait valider les extrémités de cette introgression en utilisant les données brutes Nanopore issues des variétés à gros grains. De plus, il est prévu d'analyser des séquences Nanopore d'une variété à petits grains (la variété 127244 proche phylogénétiquement de 128087) pour mettre en évidence l'absence d'introgression au niveau de la région du chromosome 4.

Il faudra également valider l'origine de cette introgression et trouver l'accession la plus proche de la plante "donneuse" *O.rupipogon* par analyse de séquences Illumina présentes dans les bases de données.

En paralèle, une amélioration de la qualité de l'assemblage de la variété 128087 est envisagée en utilisant l'outil de méta-assemblage développé au sein de notre équipe SASAR. Ainsi, différents outils d'assemblage *de novo*, tels que Canu (Koren *et al.* 2017), SMARTdenovo et wtdbg2 (Ruan *et al.* 2020) sont actuellement en train d'être testés.

Suite à l'amélioration de l'assemblage de la variété 128087, une analyse fonctionnelle du contig contenant l'introgression sera effectuée avec une annotation précise de cette région (au niveau des gènes et éléments transposables). De plus, une fouille dans les bases de données transcriptomiques disponibles de riz (Xia *et al.* 2017) et la *Rice Annotation Database, RAP-DB* permettront de connaître les conditions d'expression des gènes associés.

Après validation *in silico* de cette région, il est envisagé ensuite de passer à la validation fonctionnelle de l'insertion de *rn215-125*. Pour cela, une étude génétique sera conduite en effectuant un croisement entre la variété à gros grains 128087 et la variété proche à petits grains 127244. Si au sein de la seconde génération de plants de riz issus de ce croisement, on observe une diminution de la taille des grains, cela confirmerait notre hypothèse sur l'implication de la région de 800kb.

Dans un futur à plus long terme, si toutes les analyses précédentes s'avèrent concluantes et concordent avec notre hypothèse, une validation en *wet-lab* sera envisagée avec l'uti-

lisation du système CRISPR-Cas9 en ciblant par exemple l'insertion de *rn215-125* (Liu *et al.* 2018). Ceci se ferait en collaboration avec l'équipe de Christophe Perrin à Montpellier. Si après délétion de cette insertion spécifique, la largeur des grains des variétés diminue, cela confirmerait sans équivoque notre hypothèse.

Discussion générale

Grâce aux avancées des technologies de séquençage ces dernières décennies, le nombre de génomes séquencés est en constante augmentation. Ceci a permis une évolution majeure dans le domaine de la génomique comparative que ce soit pour des études macro-évolutives (cf les grands projets de séquençage comme "The Darwin Tree of Life", "The Vertebrate Genomes" ou encore "The Earth BioGenome Project" qui ont pour objectif de séquencer "toutes" les espèces eucaryotes dans les 10 prochaines années) ou micro-évolutives graĉe aux nombreux projets de séquençage de populations naturelles ou domestiquées (cf les 1001 génomes d'*Arabidopsis*, les 3,000 génomes de riz cultivé, les 1002 génomes de levure et les 277 variétés de maïs).

Il est maintenant admis que les éléments transposables (ET) sont des entités qui contribuent à la dynamique des génomes. Du fait de leur forte proportion au sein des génomes des plantes, il est indispensable de connaitre leur "Biologie", c'est à dire leur dynamique dans les génomes pour comprendre leur impact sur l'adaptation et donc l'évolution de ces populations (Bourgeois *et al.* 2021).

La détection des éléments transposables au sein des génomes par la technologie Illumina reste un veritable défi bio-informatique. De nombreux outils ont été développés avant le début de ma thèse (comme décrit dans l'introduction au chapitre 3) pour détecter au mieux ces insertions d'éléments transposables à partir de séquences courtes. Mais avec de tels grands jeux de données produites, il est nécessaire de développer des logiciels de détection permettant de limiter les temps de calcul des analyses. En effet, on s'est vite rendu compte que les logiciels disponibles, certes efficaces pour l'analyse de quelques génomes, étaient bien trop consommateurs de calcul et donc inadaptés pour les grands jeux de données.

Au cours de ma thèse, j'ai ainsi développé le pipeline TRACKPOSON qui répond à cette problématique, à savoir l'analyse de 3,000 génomes de riz cultivés *O.sativa*. Ce pipeline aligne les lectures contre l'ET (et non le génome de référence), ce qui représente une faible proportion des lectures totales, permettant donc une nette diminution du temps de calcul et donc l'analyse d'un grand nombre de génomes en un temps d'analyse "raisonnable" sur un cluster de calcul (ici, celui de l'Academia Sinica à Taiwan). La limitation de cette méthode est que l'on ne peut analyser qu'une seule famille d'ET à la fois, mais cette limite s'avère peu contraignante puisque, pour un génome donné, seules quelques familles d'ETs sont actives.

Nos résultats montrent qu'un grand nombre d'insertions d'ET sont présentes à une fréquence très faible au sein du riz cultivé asiatique, suggérant une transposition active récente *in agro* sans néanmoins éliminer l'hyptohèse alternative (et non exclusive) d'une élimination très rapide par sélection.

Les données obtenues grâce au nouveau logiciel que j'ai développé nous permettent désormais d'envisager l'étude de l'impact fonctionnel des ET à partir d'un grand jeu

de données par des approches d'association (GWAS). En collaboration avec un post-doctorant de l'équipe, nous avons mis au point une nouvelle stratégie, que nous appelons *TE-GWAS*, qui consiste à associer des phénotypes, non plus avec des données de SNPs, mais de TIPs (*Transposable element Insertion Polymorphisms*) : nous avons ainsi pu mettre en évidence l'impact d'une insertion d'ET sur la largeur des grains des variétés *Indica*.

Grâce à la disponibilité de la technologie de séquençage longues lectures au laboratoire, j'ai pu dans le cadre de ma thèse mener une réflexion sur l'utilisation de ce type de données pour la caractérisation et surtout la validation de ces variations structurales. Dans ce cas précis, il s'agissait de caractériser la région de l'insertion de l'ET dont les résultats TE-GWAS suggèrent qu'elle est étroitement associée à un facteur génétique impliqué dans la largeur du grain. Le chapitre 7 de cette thèse décrit ces analyses.

Plus généralement, les nouvelles technologies de séquençage longues lectures (PacBio, Nanopore) ouvrent un champ d'investigation jusque là inaccessible avec les NGS (Illumina), à savoir l'étude des variations structurales du génome et en particulier celles liées à la transposition des ET. En effet par leur longueur moyenne (qui est est en perpétuelle augmentation, avec les nouveaux développements) de 20kb, il est maintenant possible d'avoir accès à de nouvelles régions complexes du génome. Ces régions complexes impliquent très souvent des élements transposables comme les insertions multiples d'ET au sein d'autre ET (insertions nichées). De plus, les longues lectures nous permettent d'affiner la détection des ET au sein des génomes reséquencés en augmentant la sensibilité des méthodes utilisées, comme par exemple en permettant une meilleure détection des TSD (Target-Site Duplications) au sein des longues lectures et également de discriminer les copies paralogues d'ET entre elles. De plus, avec les nouvelles améliorations au niveau du *basecalling* des longues lectures Nanopore, il maintenant possible de dater les nouvelles insertions de rétrotransposons à LTR, par comparaison du pourcentage d'identité entre les LTR ("paléo-génomique"). Sans l'aide des longues lectures Nanopore, il nous aurait été impossible de caractériser la région complexe du chromosome 4 de la variété de riz du TE-GWAS (128 087), analysée au cours du chapitre 7.

Grâce à la baisse des coûts des technologies longues lectures, on peut désormais envisager l'analyse de populations. De telles analyses populationnelles pourront être réalisées grâce à l'assemblage *de novo* des individus qui composent ces populations. Comparer des génomes assemblés de haute-qualité permet des analyses beaucoup plus pertinentes en génomique comparative, en particulier en sécurisant les relations d'orthologie/paralogie au sein des génomes.

Une autre conséquence importante de la "démocratisation" des technologies "longues lectures" est qu'elle rend obsolète la notion de génome de référence, puisque l'on peut réaliser des assemblages de haute qualité à faible coût (< 10k€).

Chez le riz par exemple, on dispose désormais de génomes assemblés de haute qualité ("standard platinum") pour 12 variétés qui représentent la diversité de l'espèce cultivée (Y. Zhou *et al.* 2020). La disponibilité de ces assemblages permettent de mener des études comparatives multiples bien plus pertinentes que celles basées sur le seul génome de référence de la variété *Nipponbare*.

Je mets actuellement en place de nouveaux pipelines d'analyse basés sur ces génomes "Platinum". Après assemblage des lectures Nanopore des variétés issues de l'analyse TE-GWAS, l'un de mes objectifs est de retracer l'histoire récente des variations structurales au sein de l'espèce cultivée *O.sativa* ("généalogie génomique" récente).

En 6 ans, l'évolution des technologies de séquençage et l'arrivée des longues lectures permettent une nouvelle approche des analyses au niveau micro-évolutif (Figure 7.13).

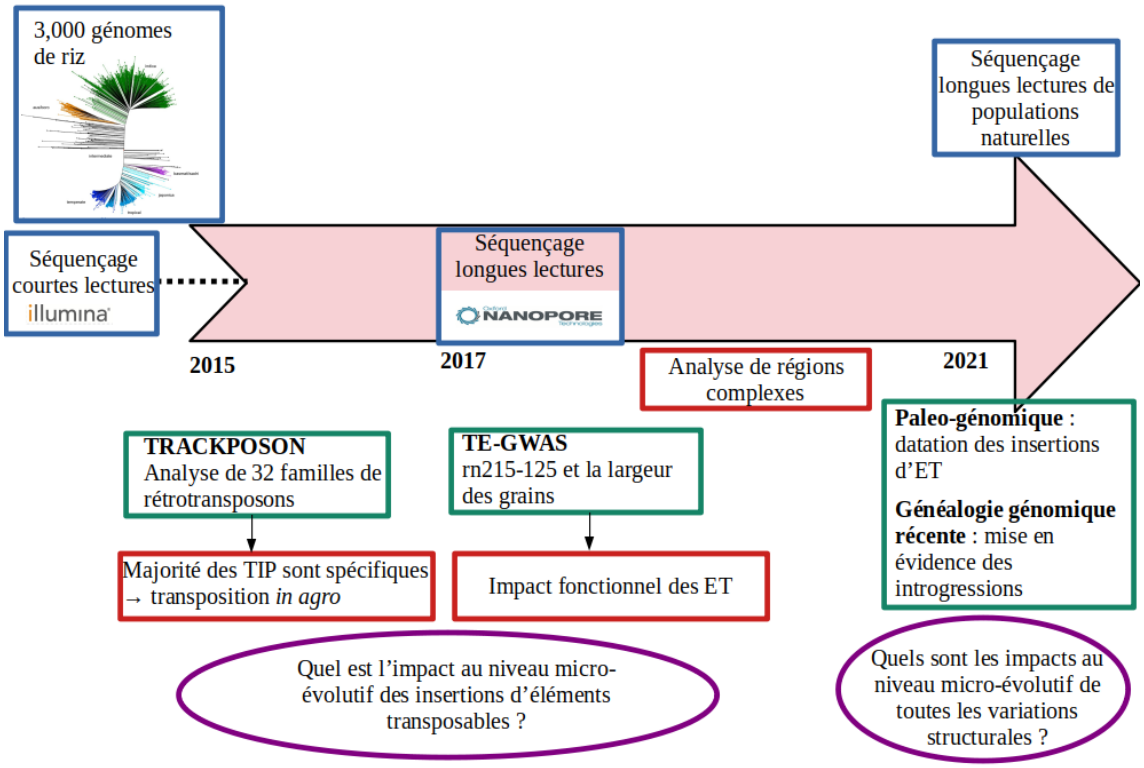


Figure 7.13 – **De l'impact des ET à l'impact de tous les SV (Variations Structurales) au sein des populations, au niveau micro-évolutif**. Les technologies de séquençage sont encadrées en bleues, en vert : les outils développés ou en perspective de développement, en rouge : les résultats obtenus, en violet : les questions biologiques.

Au début de ma thèse, nous nous sommes focalisés sur les impact de la dynamique des ET. Avec les lectures Illumina, la dynamique des ET est la principale variation structurale qui est possible de détecter avec une bonne fidélité. Maintenant avec les lectures Nanopore, en plus des ET, il est possible d'analyser toutes les variations structurales commes les introgressions, les CNV (*Copy Number Variation*), les délétions. Ainsi, cela

nous permettra de mettre en évidence la dynamique de toutes les variations structurales au sein d'une population, comme par exemple le riz cultivé.

L'étude de l'impact fonctionnel, adaptatif des ET est l'un des axes forts des activités de mon équipe de recherche. Le chapitre 7 de ma thèse s'inscrit dans cette problématique.

En parallèle de celui-ci, je participe également à un autre projet de l'équipe sur le riz sauvage *O.australiensis*. Ce riz sauvage a une taille de génome deux fois plus importante que le riz cultivé de référence *Nipponbare*. Ce doublement de taille de génome est le résultat d'une forte augmentation du nombre de copies de 4 familles principales de rétrotransposons (Piegu *et al.* 2006). Ainsi, on peut se demander comment ces insertions d'ET peuvent influencer le paysage transcriptionnelle d'une espèce sauvage ? Et est-ce que cette espèce a un meilleur potentiel adaptatif face au stress (comparé à l'espèce cultivé *Nipponbare*) ou cela n'influence en rien l'expression des gènes ?

Je peux aujourd'hui mettre à profit les compétences bioinformatiques d'analyse des données longues lectures que j'ai acquises pour l'étude du transcriptome de cette espèce. J'envisage ainsi d'effectuer un séquençage par Nanopore (MinIon) de cDNA issu de ce riz sauvage, nous permettant de mettre en évidence les co-transcrits ET-gènes potentiels, qui pourraient jouer un rôle majeur dans l'adaptation du riz sauvage aux différents stress. Je considère que l'étude des variations structurales au sein du transcriptome sera l'une des applications les plus importantes de la technologie longues lectures pour la compréhension de l'impact fonctionnel des ET. J'ai donc décidé de m'impliquer sur cette thématique à court terme.

Grâce aux longues lectures, la génomique comparative est en pleine évolution, surtout dans le domaine des variations structurales des génomes. Par leur dynamique rapide et récente, les ET au sein des génomes des plantes pourraient jouer un rôle clé dans l'adaptation des espèces aux différents changements environnementaux (Baduel *et al.* 2021). Une hypothèse envisagée serait qu'en condition de stress (biotique ou abiotique), certaines copies d'ET soient réactivées. Suite à la transposition de ces dernières, la grande majorité des néo-copies seraient éliminées mais il est possible que certaines soient gardées au cours du temps par sélection positive, si celles-ci confèrent un avantage adaptatif à la plante.

Ainsi, je pense que dans les prochaines années, grâce aux technologies "longues lectures", les impacts fonctionnels des ET et surtout leur potentiel pouvoir adaptatif au sein des populations vont être de plus en plus étudiés. De plus par ces futures études, on peut espérer répondre à des questions qui restent encore en suspens dans le domaine des ET et des études micro-évolutives. Comme par exemple comprendre quels sont les stimuli environnementaux et/ou mécanismes qui permettent la réactivation des éléments au sein des génomes de plantes ? Ou quelles sont les copies paralogues qui vont être réactivées ?

Au cours de ma thèse, je me suis intéressée aux variations structurales provoquées par les insertions d'éléments transposables. Néanmoins, ce ne sont pas les seuls responsables de ces variations au sein des génomes. Par exemple, les variations de copies de gènes (*Copy Number Variation, CNV*) peuvent également avoir un rôle non négligeable dans la dynamique des génomes des invididus. De plus l'impact des CNV sur l'adaptation des individus à certaines niches écologiques est actuellement un domaine en pleine explosion, grâce aux séquençage longues lectures qui permet de mettre en évidence efficacement le nombre de copies de gènes. Très recemment, grâce aux longues lectures Pacbio, diverses accessions de riz on été séquencées, ce qui a permis la construction d'un pan-génome de riz haute qualité (Qin *et al.* 2021). Ainsi, Qin et al. ont montré l'impact de ces variations de copies de certains gènes sur l'adaptation environnementale et les caractéristiques agronomiques de ces riz.

En conclusion, à la fin du XXème siècle, les éléments transposables étaient considérés par la majorité des chercheurs de la communauté scientifique comme de l'"ADN poubelle" (Orgel *et al.* 1980). Lors des analyses génomiques de divers organismes, ces éléments transposables étaient souvent masqués, car ne présentant aucun réel intêret biologique pour les chercheurs à l'époque : les projets se concentraient sur l'espace génique principalement.

En quelques décennies, ce postulat a bien changé : les éléments transposables sont maintenant au centre de ces analyses génomiques montrant au fur et à mesure leur fort impact sur la structure des génomes. Au delà des variations purement structurales, leur impact fonctionnel et leur rôle au cours de l'adaptation ont ces dernières années de plus en plus été mis en évidence.

Ainsi, avec les nouvelles technologies de séquençage en constante évolution, cela annonce un futur radieux pour l'étude des variations structurales, notamment les éléments transposables, et leur impact fonctionnel au niveau micro-évolutif. De par leur forte dynamique, ces éléments pourraient être une réponse effective et efficace au sein d'une population pour une rapide adaptation à différents stress.

Matériel et Méthodes

Matériel et Méthodes - Chapitre 7

Utilisation de la technologie de séquençage longues lectures pour les analyses structurales des génomes.

TRACKPOSON inversé

Les lectures pairées des 3000 génomes de riz sont alignés contre la séquence du gène de glucosyl transférase (Os04t0204100) en utilisant le programme d'alignement Bowtie2 (version 2.3.14) (Langmead *et al.* 2012).

Les paires de lectures alignées/non-alignées sont filtrées par leur FLAG à partir du fichier d'alignement SAM généré (FLAG égal à 69, 133, 165, 181, 101 ou 117).

Les lectures non-mappées de ces paires sont ensuite alignées par blast (version 2.9, Camacho *et al.* 2009) contre la séquence consensus du rétrotransposon *Hopi*. Si une lecture a un hit significatif *blast*, l'insertion d'*Hopi* à proximité du gène de GST est déclarée.

Conditions de culture du riz

Les plants de riz ont été cultivés en chambre de culture (Percival, USA). Pour mimer les conditions de culture retrouvées dans la nature ces plantes sont soumises à un cycle 12h/12h de lumière/ obscurité. Pendant la période de luminosité, les plantes sont cultivées à 28℃ et à 26℃ pour la période à l'obscurité. L'humidité relative est de 80% pendant le cycle jour et de 70% pour le cycle nuit. L'intensité lumineuse varie graduellement toutes les 40 minutes de 0% à 100% au début du cycle jour et inversement au début du cycle nuit.

Les feuilles ont été prélévées pour le séquençage en Nanopore.

Séquençage longues lectures Nanopore

Préparation des banques

Les feuilles ont été broyées dans l'azote liquide et l'ADN a été extrait dans le tampon CTAB 2X. Une purification avec le kit Zymo a été appliquée pour avoir une meilleur qualité d'ADN. La qualité de ces ADN a été vérifiée au Qubit et Nanodrop parallèlement.

Séquençage MinIon

Une quantité d'environ 3,400 ng d'ADN a été utilisé au départ.

Pour la préparation des banques Nanopore, le kit SQKLSK109 a été utlisé en suivant le protocole fourni par Nanopore (*Genomic DNA by Ligation*). Le séquençage de cet ADN a été effectué avec le séquenceur MinIon MK1B MN24609 sur 2 flowcells R9 et R10. Le basecalling a été effectué par le logiciel Guppy (version 3.2.10) en mode "High Accurancy".

Analyse des données longues lectures

Assemblage *de novo*

L'assemblage des données fastq Nanopore a été fait en utilisant le logiciel Flye (version 2.4.2, Kolmogorov *et al.* 2019) en utilisant l'option spécifique pour les données Nanopore –nano-raw et l'option -g 430m pour préciser la taille du génome qui va être assemblé.

Alignement des lectures Nanopore et/ou des contigs assemblés

Les différents alignements ont été fait avec :

— minimap2 (version 2.17, H. Li, 2018) avec les paramètres par défault

— blastn (blast+ v2.9.0)

— dotter (version 4.44.1, Sonnhammer *et al.* 1995)

Filtre des résultats d'alignement

Les fichiers d'alignement PAF issus de minimap2 ou les fichiers tabulés issus de blast ont été traités en ligne de commande avec notamment le language awk (version 5.0.1)

Création des fichiers de séquences (multifasta)

A partir des fastq, le logiciel seqtk subset (version 1.3, Shen *et al.* 2016) a été utilisé.

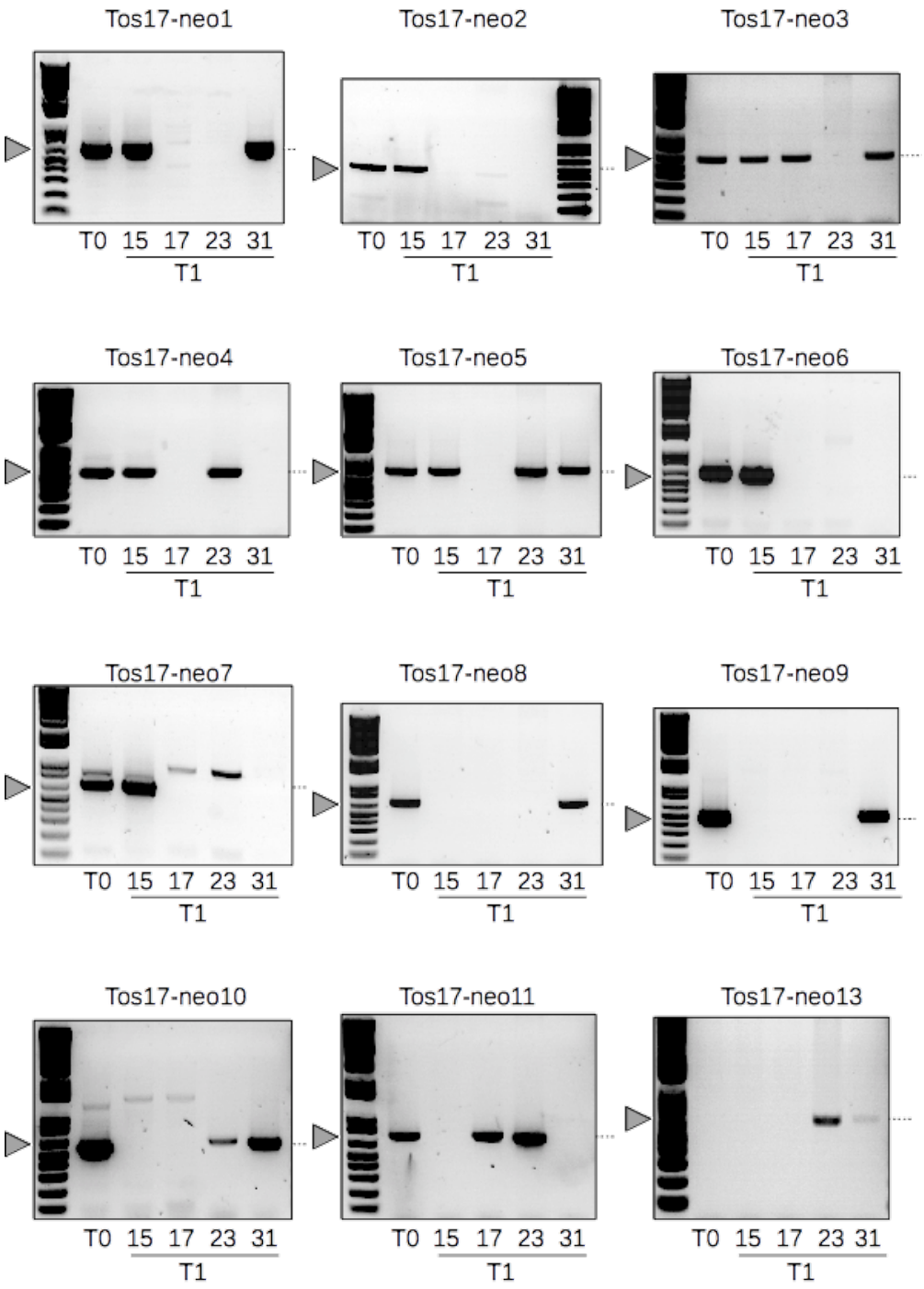
A partir de fasta, le programme bedtools getfasta (version 2.27.1, Quinlan *et al.* 2010) a été utilisé.

Séquences génomiques utilisées issues des bases de données

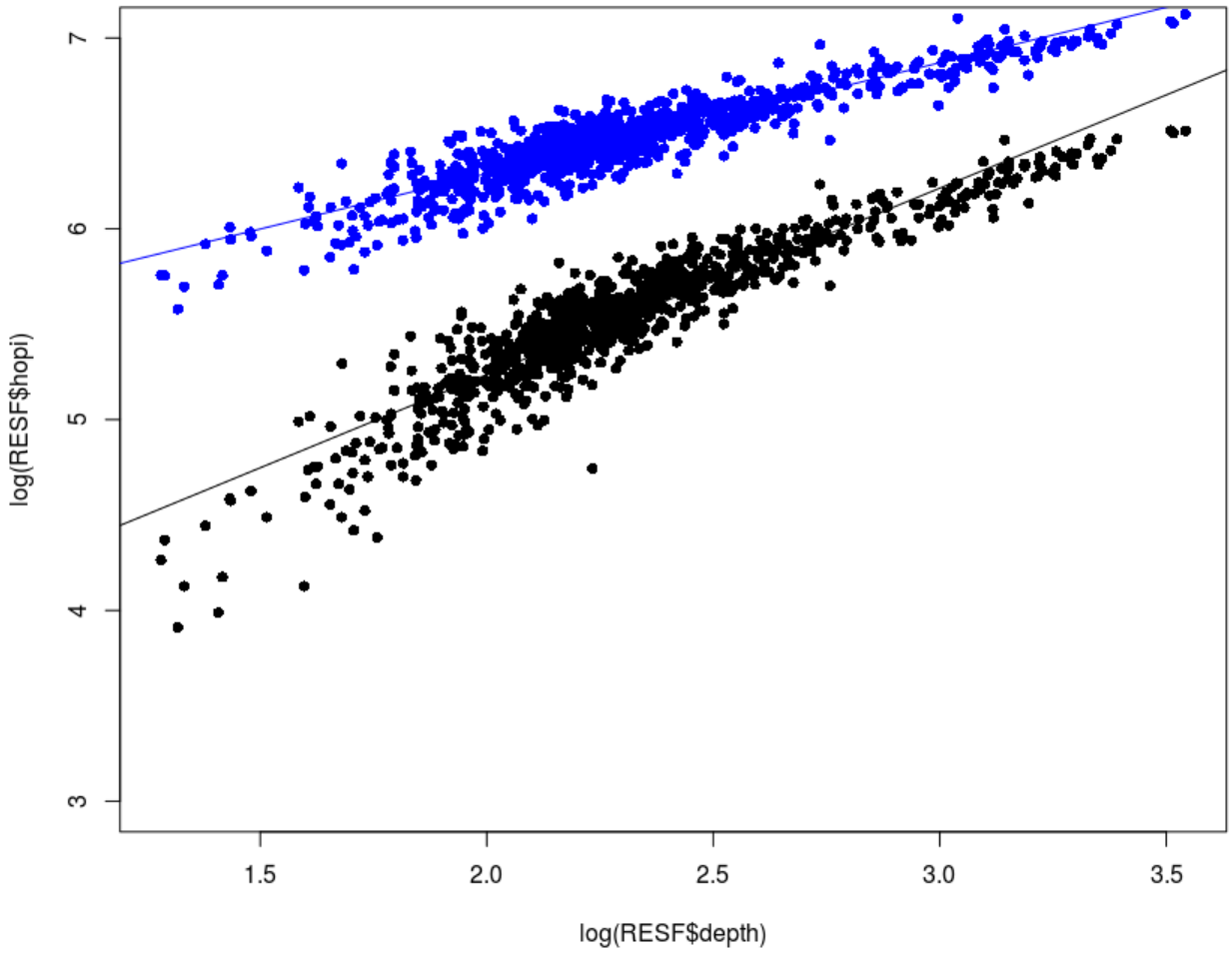
Nom	Source	Lien url
3000 génomes de riz	The 3,000 rice genomes project, 2014	3000 fastq des lectures pairées
Nipponbare, Japonica	International Rice Genome Sequencing Project *et al.* 2005	Génome assemblé IRGSP1.0
IR64, Indica	Y. Zhou *et al.* 2020	Génome assemblé IR64
Larah Mughad	Y. Zhou *et al.* 2020	Génome assemblé
Kha You Guang	Y. Zhou *et al.* 2020	Génome assemblé
Oryza rufipogon	Brozynska,M., Furtado,A. and Henry,R.J	Génome assemblé
Os04g0204100 (GST)	Rice Annotation DataBase (RAP-DB)	Séquence fasta du gène
Hopi	Carpentier *et al.* 2019	Séquence fasta de l'ET

Annexes

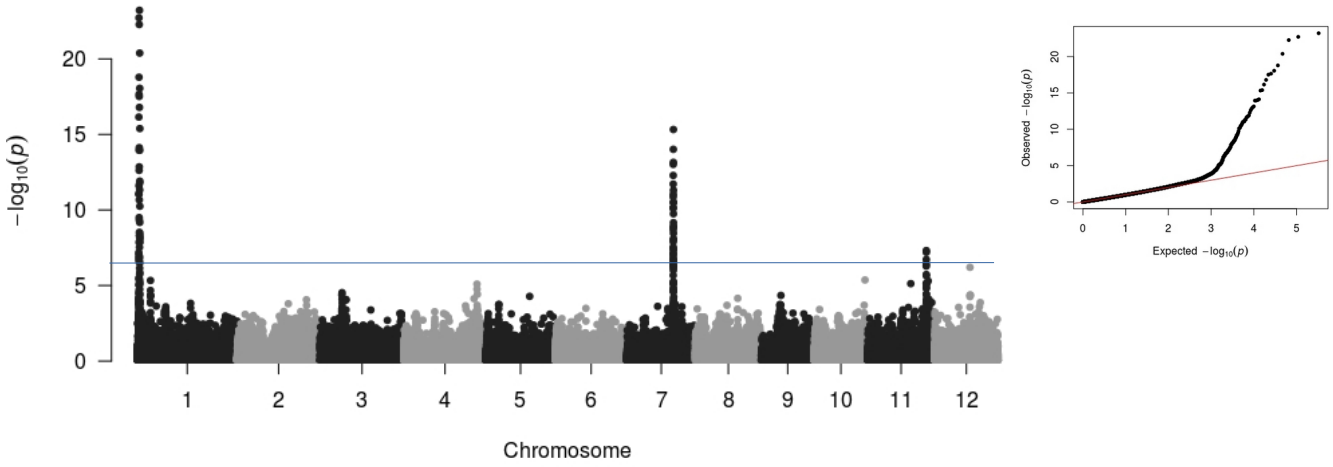
Figures supplémentaires de la publication Carpentier *et al.* 2019.



Supplementary Figure 1. Gels PCR Validation of Tos17 insertions.
A set of independently identified Tos17 insertions in a pedigree of four rice plants has been used as validation data for trackposon. T0 has been regenerated from callus after transformation by A. tumefasciens, conditions that allow Tos17 mobilization. T1-15, T1-17, T1-23 and T1-31 were obtained by selfing of the T0. After Illumina sequencing, reads have been mapped against the reference genome of Oryza sativa cv. nipponbare. Based on the position of its unique active copy, new insertions loci of Tos17 could be identified by visual analysis in a browser (IGV), of the discordantly aligned pairs, one mate of which matches the native TE copy : the discordant mates indicated the insertion position of the new copy. Agarose gels obtained after PCR. The arrow and the dashed line indicate the position of the fragments of the expected size. Other bands are the result of random non specific amplifications. Size markers are at the extreme left or right of the gels. Colors were inverted for readability.

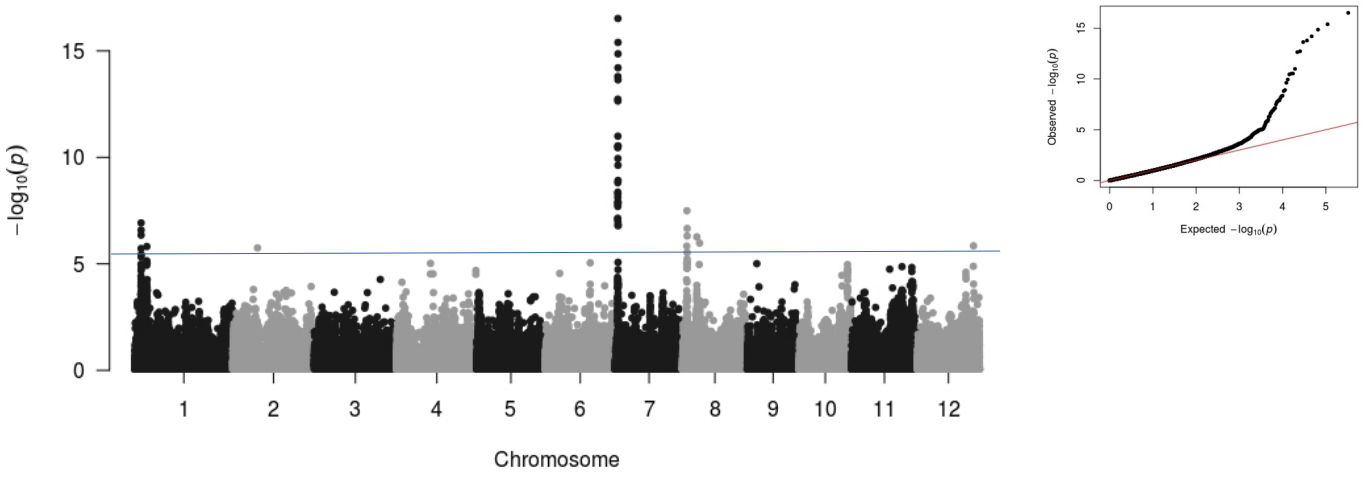


Supplementary Figure 2. The effect of the third step (threshold 2) on TIP detection. In the TRACKPOSON pipeline, the third step was added to decrease the number of false negative TE insertions.In x axis, the depth coverage for the 3000 genomes was represented in log scale, and in y axis the number of Hopi insertions also in log scale. In black, the result only the first pass at a threshold of 5 and in blue the result with the additional pass at a the threshold of 2. The lines correpond to the linear regression for the 2 representations.



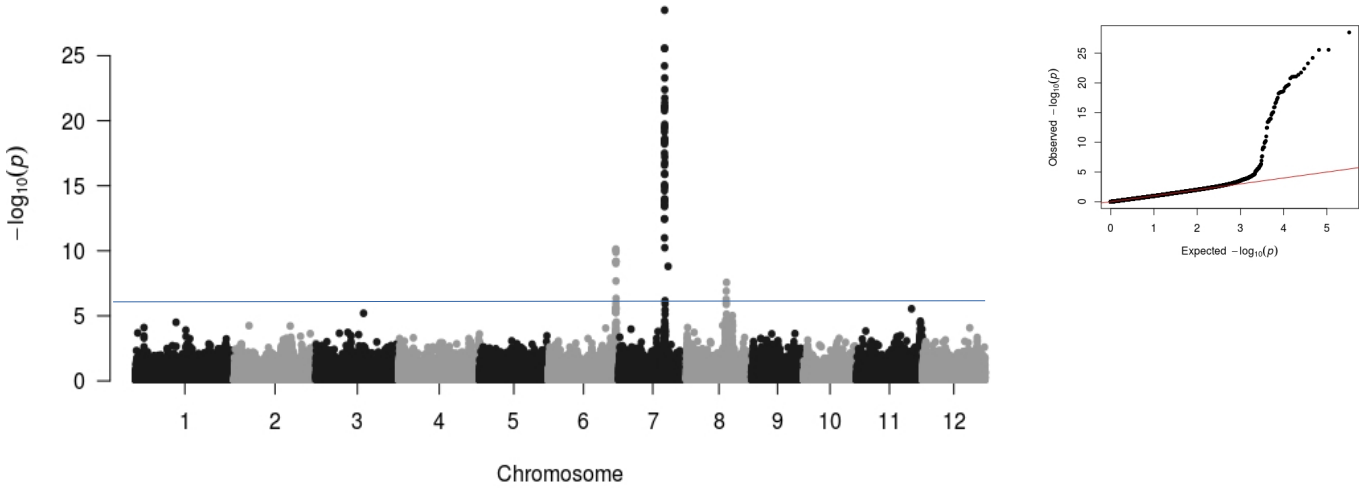
TE family	GWAS interval	TE overlapping or candidate genes	% mappability
Tos17	Chr1: 734611-1063997	Overlap with Tos17	66
Tos17	Chr7: 19995068-20201905	Overlap with Tos17	72
Tos17	Chr11:25393543-25431911	No Tos17	61

Supplementary Figure 3. GWAS results for *Tos17* retrotransposons family:
The Manhattan plot represents the association peaks. A Qqplot is also provided.
The table provides the exact localization of the peak region, together with the percentage of mappability of the region, computed using the method described in the manuscript.



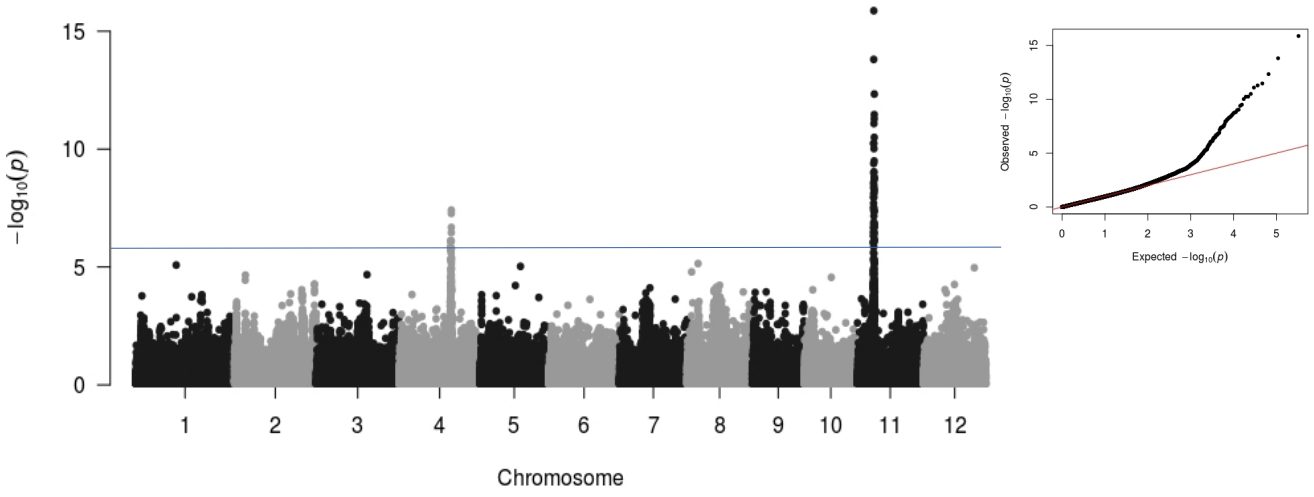
TE family	GWAS interval	Candidate TE or genes	% mappability
Karma	Chr1: 2761585-2763810	Overlap with karma copy	95
Karma	Chr7: 1043983-1159959	Overlap with karma copy	87
Karma	Chr8: 1870881-1988117	Overlap with karma copy	28.5

Supplementary Figure 4. GWAS results for *Karma* retrotransposons family:
The Manhattan plot represents the association peaks. A Qqplot is also provided.
The table provides the exact localization of the peak region, together with the percentage of mappability of the region, computed using the method described in the manuscript.



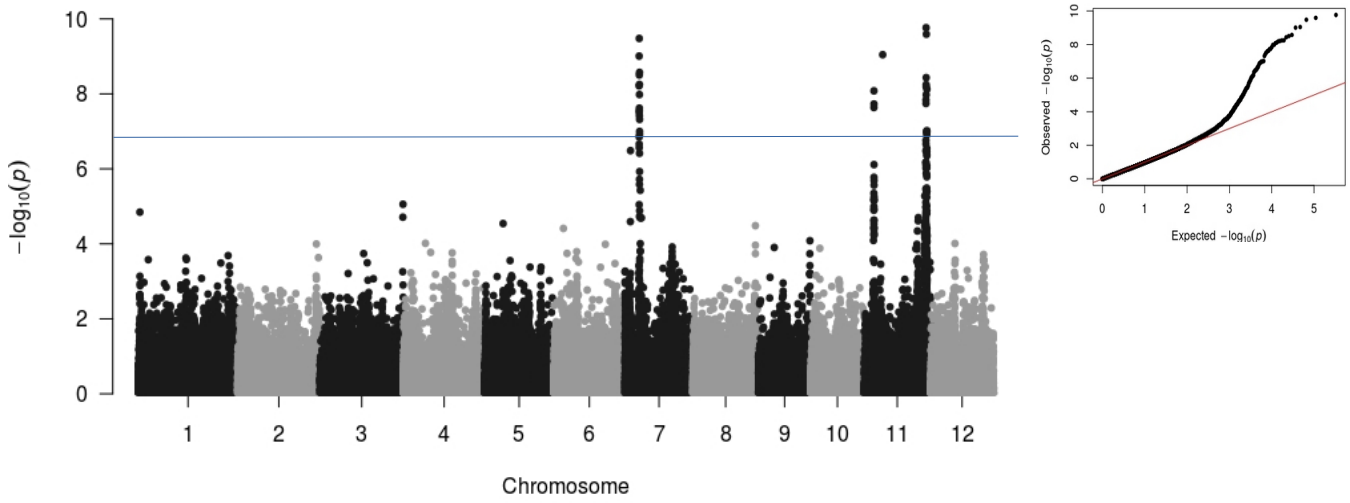
TE family	GWAS interval	Candidate TE or genes	% mappability
Fam90	Chr6: 29839249-29948887	Overlap with Fam90 copy	64
Fam90	Chr7: 19999141-20110067	Overlap with Fam90 copy	72F
Fam90	Chr8: 17428187-17558707	Overlap with Fam90 copy	58

Supplementary Figure 5. GWAS results for *Fam90* retrotransposons family:
The Manhattan plot represents the association peaks. A Qqplot is also provided.
The table provides the exact localization of the peak region, together with the percentage of mappability of the region, computed using the method described in the manuscript.



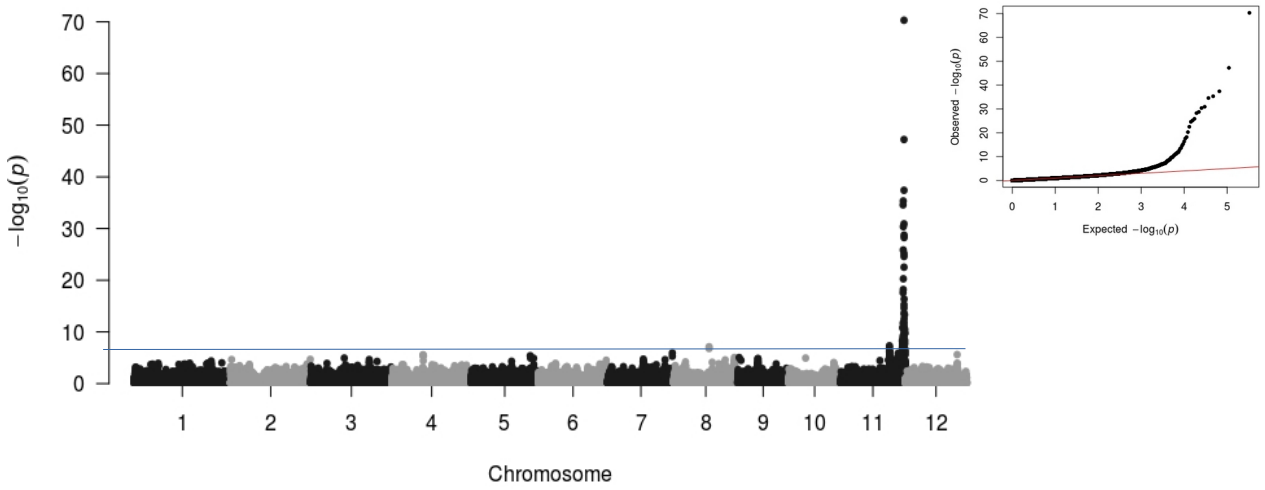
TE family	GWAS interval	Candidate TE or genes	% mappability
Fam124	Chr4: 22837141-22848622	Overlap with Fam 124	87
Fam124	Chr11: 7012603-7530068	No overlap with Fam 124	*57*

Supplementary Figure 6. GWAS results for *Fam124* retrotransposons family:
The Manhattan plot represents the association peaks. A Qqplot is also provided.
The table provides the exact localization of the peak region, together with the percentage of mappability of the region, computed using the method described in the manuscript.



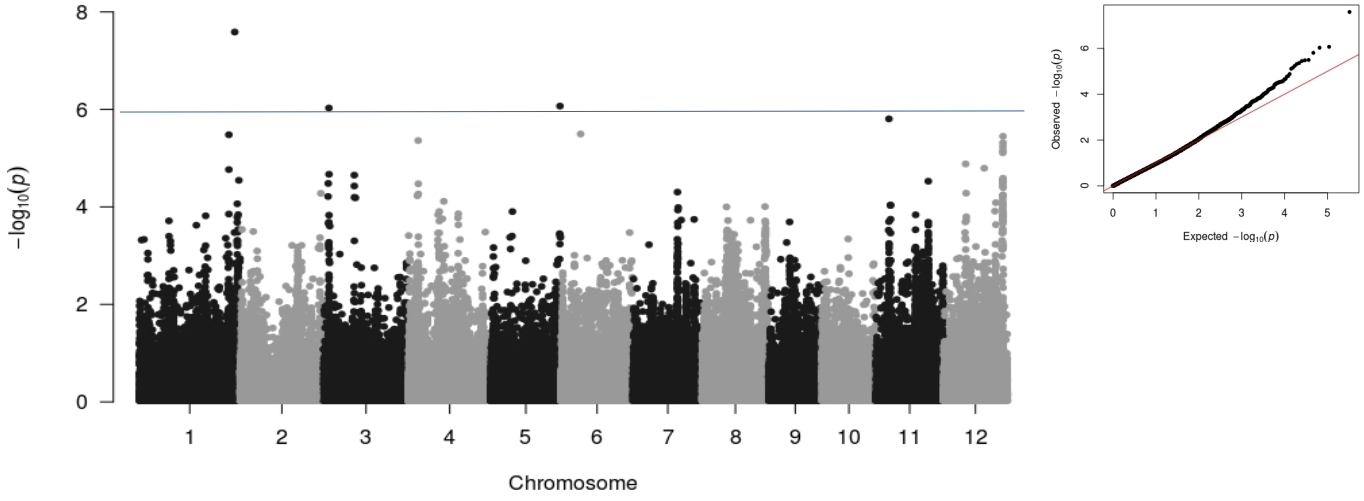
TE family	GWAS interval	Candidate TE or genes	% mappability
Fam106	Chr7: 6170573-6363341	Overlap with fam106 copy	41
Fam106	Chr11: 4316562-4375485	Overlap with fam106 copy	67
Fam106	Chr11: 27008905-27445698	Overlap with fam106 copy	62

Supplementary Figure 7. GWAS results for *Fam106* retrotransposons family:
The Manhattan plot represents the association peaks. A Qqplot is also provided.
The table provides the exact localization of the peak region, together with the percentage of mappability of the region, computed using the method described in the manuscript.



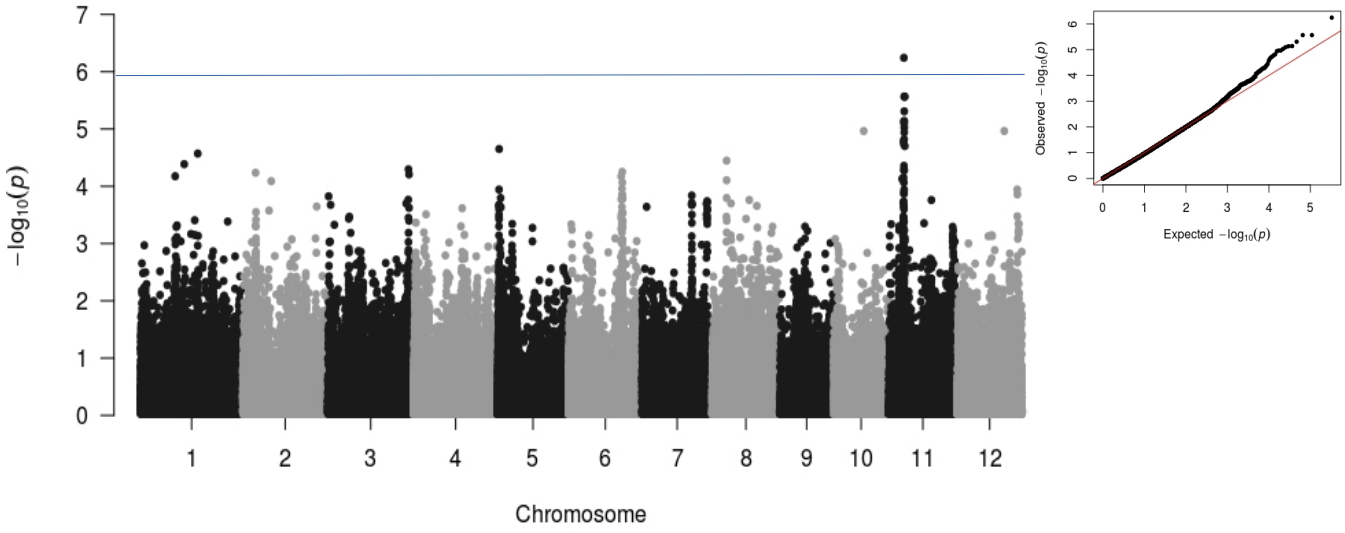
TE family	GWAS interval	Candidate TE or genes	% mappability
Lullaby	Chr11: 28186914-28477992	Overlap with lullaby copy	68

Supplementary Figure 8. GWAS results for *Lulabby* retrotransposons family:
The Manhattan plot represents the association peaks. A Qqplot is also provided.
The table provides the exact localization of the peak region, together with the percentage of mappability of the region, computed using the method described in the manuscript.



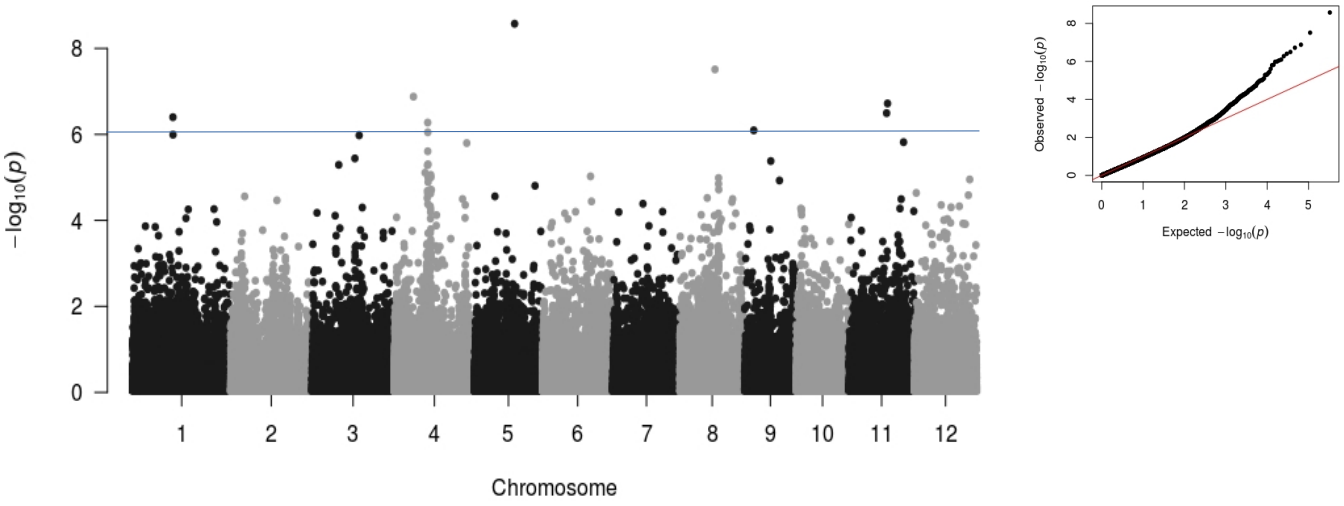
TE family	GWAS interval	Candidate TE or genes	Mappability
Dasheng	Chr1:40946066	Overlap with dasheng	76
Dasheng	Chr3:2251217-2312182	Overlap with dasheng	58

Supplementary Figure 9. GWAS results for *Dasheng* retrotransposons family:
The Manhattan plot represents the association peaks. A Qqplot is also provided.
The table provides the exact localization of the peak region, together with the percentage of mappability of the region, computed using the method described in the manuscript.



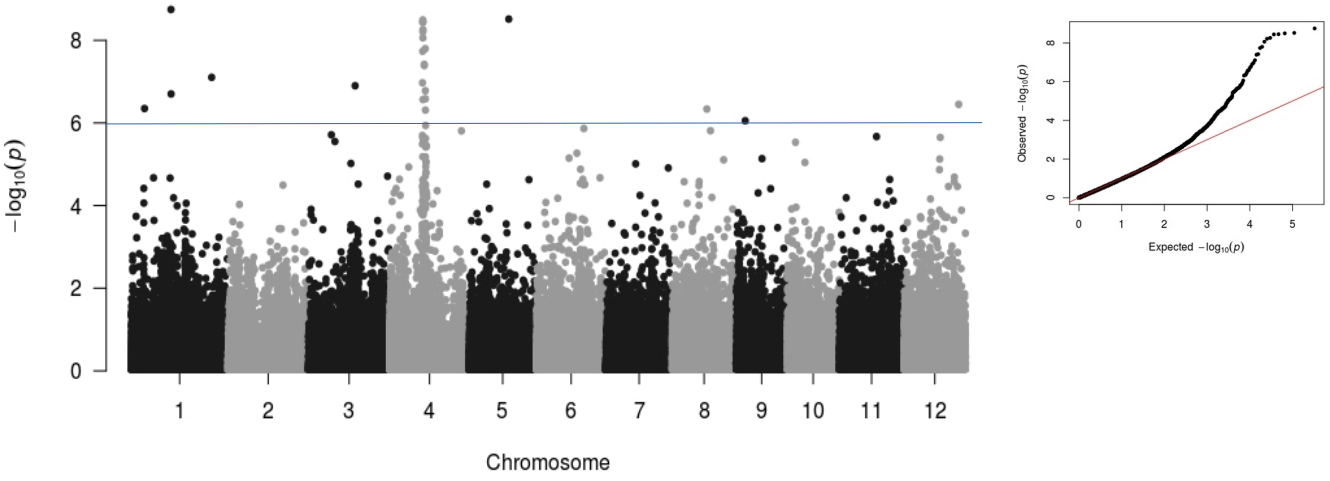
TE family	GWAS interval	Candidate TE or genes	% mappability
Fam89	Chr11: 6481747-6602990	No overlap with Fam89	56

Supplementary Figure 10. GWAS results for *Fam89* retrotransposons family:
The Manhattan plot represents the association peaks. A Qqplot is also provided.
The table provides the exact localization of the peak region, together with the percentage of mappability of the region, computed using the method described in the manuscript.



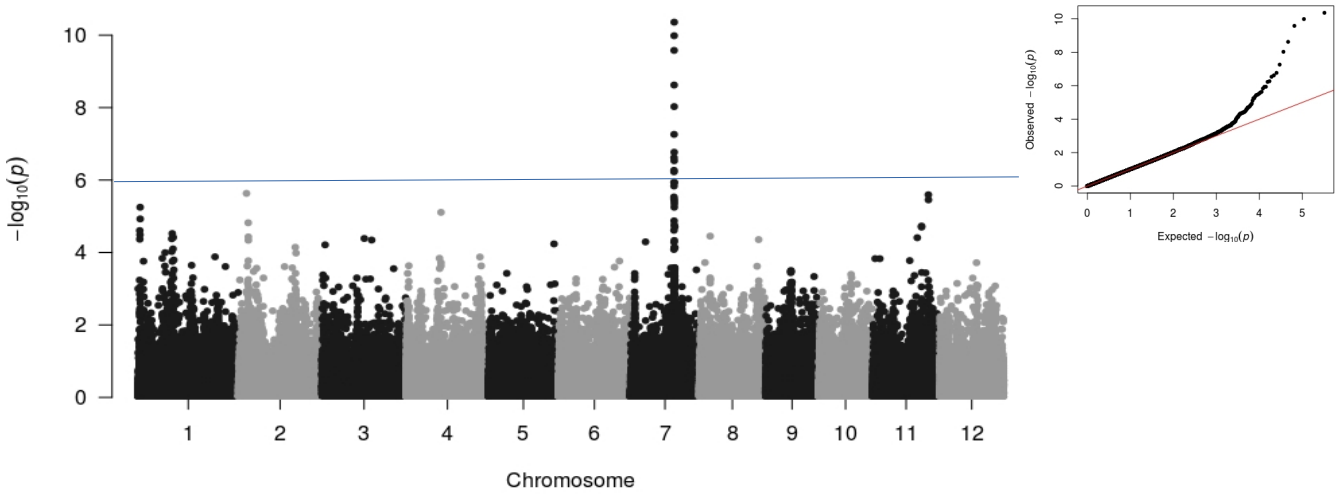
TE family	GWAS interval	Candidate TE or genes	% mappability
Houba	Chr01: 17769540-17811498	Centromere no overlap with Houba	47
Houba	Chr04: 14769176-14819944	No overlap with Houba	56

Supplementary Figure 11. GWAS results for *Houba* retrotransposons family:
The Manhattan plot represents the association peaks. A Qqplot is also provided.
The table provides the exact localization of the peak region, together with the percentage of mappability of the region, computed using the method described in the manuscript.



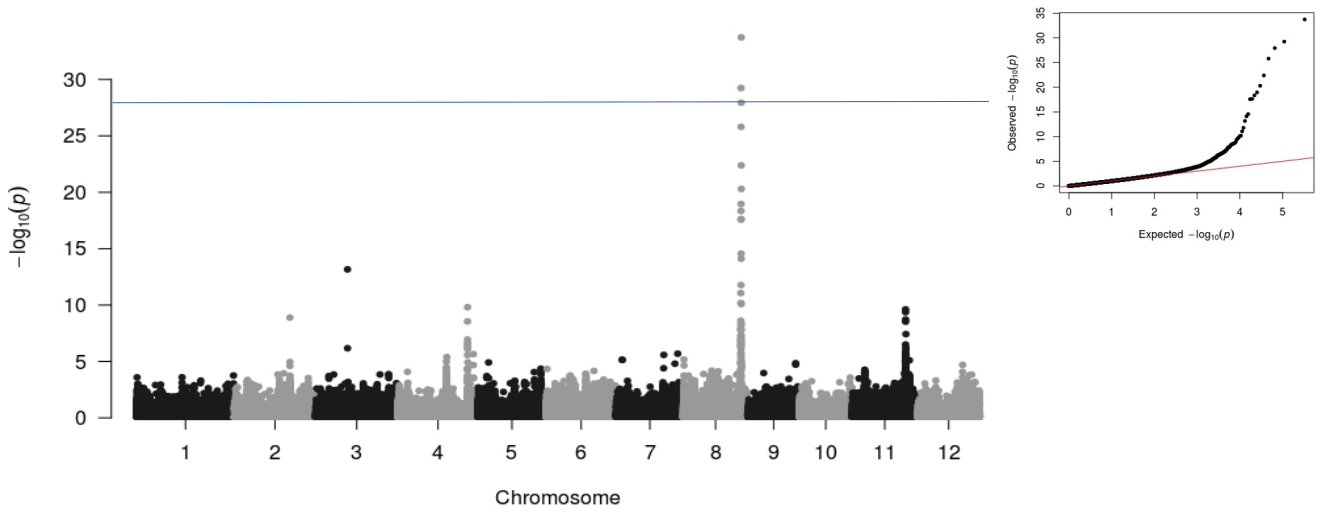
TE family	GWAS interval	Candidate TE or genes	% mappability
Fam86	Chr01: 17769540-17811498	No overlap with Fam86	47
Fam86	Chr4:14599121-14827052	No overlap with Fam86	53

Supplementary Figure 12. GWAS results for *Fam86* retrotransposons family:
The Manhattan plot represents the association peaks. A Qqplot is also provided.
The table provides the exact localization of the peak region, together with the percentage of mappability of the region, computed using the method described in the manuscript.



TE family	GWAS interval	Candidate TE or genes	% mappability
rn304	Chr07: 18726085-18820090	Overlap with rn304 copy	55

Supplementary Figure 13. GWAS results for *Rn304* retrotransposons family:
The Manhattan plot represents the association peaks. A Qqplot is also provided.
The table provides the exact localization of the peak region, together with the percentage of mappability of the region, computed using the method described in the manuscript.



TE family	GWAS interval	Candidate TE or genes	% mappability
Scaff3	Chr04: 30805389-30819182	Overlap with scaff3 copy	49
Scaff3	Chr08: 25480885-25631114	Overlap with scaff3 copy	55
Scaff3	Chr11: 23710053-23731730	Overlap with scqff3 copy	31

Supplementary Figure 14. GWAS results for *Scaff3* retrotransposons family:
The Manhattan plot represents the association peaks. A Qqplot is also provided.
The table provides the exact localization of the peak region, together with the percentage of mappability of the region, computed using the method described in the manuscript.

Supplementary Table 1. The summary of *Tos17* insertions results

			T0			
			T1-17	T1-15	T1-23	T1-31
Chr	Position	Name				
Chr01	**2729576**	**Tos17-1**		**X**	**X**		**X**
Chr02	32884818	Tos17-2		**X**	**X**		
Chr03	**5572701**	**Tos17-3**	**X**	**X**	**X**		**X**
Chr03	7989559	Tos17-4		**X**	**X**	**X**	
Chr03	**34643054**	**Tos17-5**		**X**	**X**	**X**	**X**
Chr09	21062779	Tos17-6		**X**	**X**		
Chr11	**25235933**	**Tos17-7**		**X**	**X**		
Chr01	7120092	Tos17-8			**X**		**X**
Chr01	**40103765**	**Tos17-9**			**X**		**X**
Chr07	856885	Tos17-10			**X**	**X**	**X**
Chr09	**19674711**	**Tos17-11**	**X**		**X**	**X**	
Chr01	5815622	Tos17-13				**X**	**X**

The summary of results, where a X indicates that the insertion is present.
All these insertions have been confirmed by PCR (see Supplementary Figure 1).

Supplementary Table 2. The list of primers for *Tos17* insertions validation

Name	Sequence	Position	Strand
tos17_r	GAATTGGCAGCTAGGGTTCA	tos17:157-176	-
tos17_neo1_f	TCATGCTGAATTAGATCGTGGT	chr01:2729096-2729117	+
tos17_neo2_f	TGGCCTGTGGTACTTGTGAG	chr02:32884565-32884584	+
tos17_neo3_r	AGCTGGAGCTCTGCCTAACA	chr03:5573041-5573060	-
tos17_neo4_r	GCCCACCATTGACGAATAAA	chr03:7989863-7989882	-
tos17_neo5_f	TTGTGTCCAAGGCTCTGATG	chr03:34642596-34642615	+
tos17_neo6_r	GCTAATTGAGATGCCCTTGG	chr09:21063183-21063202	-
tos17_neo7_r	ACTCGGTGGCCTCAAATCTA	chr11:25236386-25236405	-
tos17_neo8_r	CCTGCCAGAGTTCAGATTTCA	chr01:7120510-7120530	-
tos17_neo9_r	TTGCACCCCAAAAGCTAATC	chr01:40104057-40104076	-
tos17_neo10_f	CAACCCCTTCCAAAACTGAG	chr07:856513-856532	+
tos17_neo11_f	CATCCGTTTTAAATTGCTTGG	chr09:19674214-19674234	+
tos17_neo13_r	CGCTTGTTTTGCAGTCAAAG	chr01:5815999-5816018	-

The list of name, sequence, position and strand of each primer used for validation.The first in the list, tos17_r, has been used with each other primer to check for the presence of each insertion

Supplementary Table 3. Gypsy and Copia TIPs localisation

	1st quantile	Median	Mean	3th quantile
Copia	182 nt	11,296 nt	22,888 nt	32,064 nt
Gypsy	432 nt	11,665 nt	23,459 nt	32,733 nt

We determinated for all the TIPs, the distance in nucleotides with the closest gene (see Methods). We performed a Whelch t-test between these two dsitributions with p.value of 1.6e-07. The Copia and Gypsy distribution are significativly different and the Copia TIPs are closer to the genes.

Bibliographie

Alonge, M. *et al.* (2020). "Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato". en. In : *Cell* 182.1, 145-161.e23. doi : 10.1016/j.cell.2020.05.021.

Anderson, S. N. *et al.* (2019). "Transposable elements contribute to dynamic genome content in maize". en. In : *The Plant Journal* 100.5, p. 1052-1065. doi : 10.1111/tpj.14489.

Baduel, P., L. Quadrana et V. Colot (2020). *Efficient detection of transposable element insertion polymorphisms between genomes using short-read sequencing data*. en. preprint. Bioinformatics. doi : 10.1101/2020.06.09.142331.

Baduel, P. *et al.* (2021). "Genetic and environmental modulation of transposition shapes the evolutionary potential of Arabidopsis thaliana". en. In : *Genome Biology* 22.1, p. 138. doi : 10.1186/s13059-021-02348-5.

Baud, A. *et al.* (2019). "Traces of transposable element in genome dark matter co-opted by flowering gene regulation networks". en. In : *bioRxiv*, p. 547877. doi : 10.1101/547877.

Bennetzen, J. L. et H. Wang (2014). "The Contributions of Transposable Elements to the Structure, Function, and Evolution of Plant Genomes". en. In : *Annual Review of Plant Biology* 65.1, p. 505-530. doi : 10.1146/annurev-arplant-050213-035811.

Billey, E. *et al.* (2021). "LARP6C orchestrates post-transcriptional reprogramming of gene expression during hydration to promote pollen tube guidance". en. In : *The Plant Cell*, koab131. doi : 10.1093/plcell/koab131.

Bogaerts-Márquez, M. *et al.* (2019). "T-lex3 : an accurate tool to genotype and estimate population frequencies of transposable elements using the latest short-read whole genome sequencing data". en. In : *Bioinformatics*. Sous la dir. de R. Schwartz, btz727. doi : 10.1093/bioinformatics/btz727.

Bourgeois, Y. X. C. et B. H. Warren (2021). "An overview of current population genomics methods for the analysis of whole-genome resequencing data in eukaryotes". en. In : *Molecular Ecology* n/a.n/a. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1111/mec.15989. doi : https://doi.org/10.1111/mec.15989.

Bourque, G. *et al.* (2018). "Ten things you should know about transposable elements". en. In : *Genome Biology* 19.1, p. 199. doi : 10.1186/s13059-018-1577-z.

Britten, R. J. et D. E. Kohne (1968). "Repeated Sequences in DNA". en. In : *Science* 161.3841, p. 529-540. doi : 10.1126/science.161.3841.529.

Bundock, P. et P. Hooykaas (2005). "An Arabidopsis hAT-like transposase is essential for plant development". en. In : *Nature* 436.7048, p. 282-284. doi : 10.1038/nature03667.

Butelli, E. *et al.* (2012). "Retrotransposons Control Fruit-Specific, Cold-Dependent Accumulation of Anthocyanins in Blood Oranges". en. In : *The Plant Cell* 24.3, p. 1242-1255. doi : 10.1105/tpc.111.095232.

Camacho, C. *et al.* (2009). "BLAST+ : architecture and applications". en. In : *BMC Bioinformatics* 10.1, p. 421. doi : 10.1186/1471-2105-10-421.

Cameron, J. R., E. Y. Loh et R. W. Davis (1979). "Evidence for transposition of dispersed repetitive DNA families in yeast". en. In : *Cell* 16.4, p. 739-751. doi : 10.1016/0092-8674(79)90090-4.

Carlos Guimaraes, L. *et al.* (2015). "Inside the Pan-genome - Methods and Software Overview". en. In : *Current Genomics* 16.4, p. 245-252. doi : 10.2174/1389202916666150423002311.

Carpentier, M.-C., C. Bousquet-Antonelli et R. Merret (2021). “Fast and Efficient 5′P Degradome Library Preparation for Analysis of Co-Translational Decay in Arabidopsis”. en. In : *Plants* 10.3, p. 466. doi : 10.3390/plants10030466.

Carpentier, M.-C., A. Picart-Picolo et F. Pontvianne (2018). “A Method to Identify Nucleolus-Associated Chromatin Domains (NADs)”. In : *Plant Chromatin Dynamics*. Sous la dir. de M. Bemer et C. Baroux. T. 1675. Series Title : Methods in Molecular Biology. New York, NY : Springer New York, p. 99-109. doi : 10.1007/978-1-4939-7318-7_7.

Carpentier, M.-C. *et al.* (2019). “Retrotranspositional landscape of Asian rice revealed by 3000 genomes”. en. In : *Nature Communications* 10.1, p. 24. doi : 10.1038/s41467-018-07974-5.

Carpentier, M.-C. *et al.* (2020). “Monitoring of XRN4 Targets Reveals the Importance of Cotranslational Decay during Arabidopsis Development”. en. In : *Plant Physiology* 184.3, p. 1251-1262. doi : 10.1104/pp.20.00942.

Casacuberta, E. et J. González (2013). “The impact of transposable elements in environmental adaptation”. en. In : *Molecular Ecology* 22.6, p. 1503-1517. doi : 10.1111/mec.12170.

Casacuberta, J. M. et N. Santiago (2003). “Plant LTR-retrotransposons and MITEs : control of transposition and impact on the evolution of plant genes and genomes”. en. In : *Gene* 311, p. 1-11. doi : 10.1016/S0378-1119(03)00557-2.

Castanera, R. *et al.* (2020). *The replicative amplification of MITEs and their impact on rice trait variability*. en. preprint. Molecular Biology. doi : 10.1101/2020.10.01.322784.

Castanera, R. *et al.* (2021). “Amplification dynamics of miniature inverted-repeat transposable elements and their impact on rice trait variability”. en. In : *The Plant Journal*, tpj.15277. doi : 10.1111/tpj.15277.

Cavrak, V. V. *et al.* (2014). “How a Retrotransposon Exploits the Plant’s Heat Stress Response for Its Activation”. en. In : *PLoS Genetics* 10.1. Sous la dir. de T. Kakutani, e1004115. doi : 10.1371/journal.pgen.1004115.

Choi, J. Y. *et al.* (2020). “Nanopore sequencing-based genome assembly and evolutionary genomics of circum-basmati rice”. en. In : *Genome Biology* 21.1, p. 21. doi : 10.1186/s13059-020-1938-2.

Chuong, E. B., N. C. Elde et C. Feschotte (2016). “Regulatory evolution of innate immunity through co-option of endogenous retroviruses”. en. In : *Science* 351.6277, p. 1083-1087. doi : 10.1126/science.aad5497.

Chuong, E. B., N. C. Elde et C. Feschotte (2017). “Regulatory activities of transposable elements : from conflicts to benefits”. en. In : *Nature Reviews Genetics* 18.2, p. 71-86. doi : 10.1038/nrg.2016.139.

Civáň, P. *et al.* (2015). “Three geographically separate domestications of Asian rice”. en. In : *Nature Plants* 1.11, p. 15164. doi : 10.1038/nplants.2015.164.

Cleal, K. et D. M. Baird (2021). *Dysgu : efficient structural variant calling using short or long reads*. en. preprint. Bioinformatics. doi : 10.1101/2021.05.28.446147.

Copetti, D. *et al.* (2015). “RiTE database : a resource database for genus-wide rice genomics and evolutionary biology”. en. In : *BMC Genomics* 16.1, p. 538. doi : 10.1186/s12864-015-1762-3.

Cosby, R. L. *et al.* (2021). “Recurrent evolution of vertebrate transcription factors by transposase capture”. en. In : *Science* 371.6531, eabc6405. doi : 10.1126/science.abc6405.

Cretu Stancu, M. *et al.* (2017). “Mapping and phasing of structural variation in patient genomes using nanopore sequencing”. en. In : *Nature Communications* 8.1. Number : 1 Publisher : Nature Publishing Group, p. 1326. doi : 10.1038/s41467-017-01343-4.

Debladis, E. *et al.* (2017). “Detection of active transposable elements in Arabidopsis thaliana using Oxford Nanopore Sequencing technology”. en. In : *BMC Genomics* 18.1, p. 537. doi : 10.1186/s12864-017-3753-z.

Domínguez, M. *et al.* (2020). “The impact of transposable elements on tomato diversity”. en. In : *Nature Communications* 11.1, p. 4058. doi : 10.1038/s41467-020-17874-2.

Doolittle, W. F. et C. Sapienza (1980). “Selfish genes, the phenotype paradigm and genome evolution”. en. In : *Nature* 284.5757, p. 601-603. doi : 10.1038/284601a0.

El Baidouri, M. *et al.* (2014). “Widespread and frequent horizontal transfers of transposable elements in plants”. en. In : *Genome Research* 24.5, p. 831-838. doi : 10.1101/gr.164400.113.

El Baidouri, M. et O. Panaud (2013). “Comparative Genomic Paleontology across Plant Kingdom Reveals the Dynamics of TE-Driven Genome Evolution”. en. In : *Genome Biology and Evolution* 5.5, p. 954-965. doi : 10.1093/gbe/evt025.

Elbaidouri, M., C. Chaparro et O. Panaud (2013). “Use of Next Generation Sequencing (NGS) Technologies for the Genome-Wide Detection of Transposition”. In : *Plant Transposable Elements*. Sous la dir. de T. Peterson. T. 1057. Series Title : Methods in Molecular Biology. Totowa, NJ : Humana Press, p. 265-274. doi : 10.1007/978-1-62703-568-2_19.

Elliott, T. A. et T. R. Gregory (2015). “Do larger genomes contain more diverse transposable elements ?” en. In : *BMC Evolutionary Biology* 15.1, p. 69. doi : 10.1186/s12862-015-0339-8.

Ferrero-Serrano, Á., C. Cantos et S. M. Assmann (2019). “The Role of Dwarfing Traits in Historical and Modern Agriculture with a Focus on Rice”. en. In : *Cold Spring Harbor Perspectives in Biology* 11.11, a034645. doi : 10.1101/cshperspect.a034645.

Feschotte, C. (2008). “Transposable elements and the evolution of regulatory networks”. en. In : *Nature Reviews Genetics* 9.5, p. 397-405. doi : 10.1038/nrg2337.

Feschotte, C., L. Swamy et S. R. Wessler (2003). “Genome-wide analysis of mariner-like transposable elements in rice reveals complex relationships with stowaway miniature inverted repeat transposable elements (MITEs)”. eng. In : *Genetics* 163.2, p. 747-758.

Finnegan, D. J. (1989). “Eukaryotic transposable elements and genome evolution”. en. In : *Trends in Genetics* 5, p. 103-107. doi : 10.1016/0168-9525(89)90039-5.

Fuentes, R. R. *et al.* (2019). “Structural variants in 3000 rice genomes”. en. In : *Genome Research* 29.5, p. 870-880. doi : 10.1101/gr.241240.118.

Galindo-González, L. *et al.* (2017). “LTR-retrotransposons in plants : Engines of evolution”. en. In : *Gene* 626, p. 14-25. doi : 10.1016/j.gene.2017.04.051.

Gardner, E. J. *et al.* (2017). “The Mobile Element Locator Tool (MELT) : population-scale mobile element discovery and biology”. en. In : *Genome Research* 27.11, p. 1916-1929. doi : 10.1101/gr.218032.116.

Gilbert, C. et C. Feschotte (2018). “Horizontal acquisition of transposable elements and viral sequences : patterns and consequences”. en. In : *Current Opinion in Genetics & Development* 49, p. 15-24. doi : 10.1016/j.gde.2018.02.007.

Goerner-Potvin, P. et G. Bourque (2018). “Computational tools to unmask transposable elements”. en. In : *Nature Reviews Genetics* 19.11, p. 688-704. doi : 10.1038/s41576-018-0050-x.

Goldfeder, R. L. *et al.* (2017). "Human Genome Sequencing at the Population Scale : A Primer on High-Throughput DNA Sequencing and Analysis". en. In : *American Journal of Epidemiology* 186.8, p. 1000-1009. doi : 10.1093/aje/kww224.

Goubert, C. *et al.* (2020). "TypeTE : a tool to genotype mobile element insertions from whole genome resequencing data". en. In : *Nucleic Acids Research*, gkaa074. doi : 10.1093/nar/gkaa074.

Grandbastien, M.-A. *et al.* (2005). "Stress activation and genomic impact of Tnt1 retrotransposons in Solanaceae". en. In : *Cytogenetic and Genome Research* 110.1-4, p. 229-241. doi : 10.1159/000084957.

Grandbastien, M.-A. (2015). "LTR retrotransposons, handy hitchhikers of plant regulation and stress response". en. In : *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* 1849.4, p. 403-416. doi : 10.1016/j.bbagrm.2014.07.017.

Grandbastien, M.-A., A. Spielmann et M. Caboche (1989). "Tnt1, a mobile retroviral-like transposable element of tobacco isolated by plant cell genetics". en. In : *Nature* 337.6205, p. 376-380. doi : 10.1038/337376a0.

Gregory, T. R. (2005). "Synergy between sequence and size in Large-scale genomics". en. In : *Nature Reviews Genetics* 6.9, p. 699-708. doi : 10.1038/nrg1674.

Gross, B. L. et Z. Zhao (2014). "Archaeological and genetic insights into the origins of domesticated rice". en. In : *Proceedings of the National Academy of Sciences* 111.17, p. 6190-6197. doi : 10.1073/pnas.1308942110.

Gutaker, R. M. *et al.* (2020). "Genomic history and ecology of the geographic spread of rice". en. In : *Nature Plants* 6.5, p. 492-502. doi : 10.1038/s41477-020-0659-6.

Hénaff, E. *et al.* (2015). "Jitterbug : somatic and germline transposon insertion detection at single-nucleotide resolution". en. In : *BMC Genomics* 16.1, p. 768. doi : 10.1186/s12864-015-1975-5.

Hirochika, H. *et al.* (1996). "Retrotransposons of rice involved in mutations induced by tissue culture." en. In : *Proceedings of the National Academy of Sciences* 93.15, p. 7783-7788. doi : 10.1073/pnas.93.15.7783.

Hirochika, H. et H. Otsuki (1995). "Extrachromosomal circular forms of the tobacco retrotransposon Ttol". en. In : *Gene* 165.2, p. 229-232. doi : 10.1016/0378-1119(95)00581-P.

Hirochika, H. *et al.* (2004). "Rice Mutant Resources for Gene Discovery". en. In : *Plant Molecular Biology* 54.3, p. 325-334. doi : 10.1023/B:PLAN.0000036368.74758.66.

Hof, A. E. v. *et al.* (2016). "The industrial melanism mutation in British peppered moths is a transposable element". en. In : *Nature* 534.7605, p. 102-105. doi : 10.1038/nature17951.

Hollister, J. D. *et al.* (2011). "Transposable elements and small RNAs contribute to gene expression divergence between Arabidopsis thaliana and Arabidopsis lyrata". en. In : *Proceedings of the National Academy of Sciences* 108.6, p. 2322-2327. doi : 10.1073/pnas.1018222108.

Horváth, V., M. Merenciano et J. González (2017). "Revisiting the Relationship between Transposable Elements and the Eukaryotic Stress Response". en. In : *Trends in Genetics* 33.11, p. 832-841. doi : 10.1016/j.tig.2017.08.007.

Hsu, C.-C. *et al.* (2020). "Identification of high-copy number long terminal repeat retrotransposons and their expansion in Phalaenopsis orchids". en. In : *BMC Genomics* 21.1, p. 807. doi : 10.1186/s12864-020-07221-6.

Hua-Van, A. *et al.* (2011). "The struggle for life of the genome's selfish architects". en. In : *Biology Direct* 6.1, p. 19. doi : 10.1186/1745-6150-6-19.

Huang, X. *et al.* (2012). “A map of rice genome variation reveals the origin of cultivated rice”. en. In : *Nature* 490.7421, p. 497-501. doi : 10.1038/nature11532.

International Human Genome Sequencing Consortium (2001). “Initial sequencing and analysis of the human genome”. en. In : *Nature* 409.6822, p. 860-921. doi : 10.1038/35057062.

International Rice Genome Sequencing Project et T. Sasaki (2005). “The map-based sequence of the rice genome”. en. In : *Nature* 436.7052, p. 793-800. doi : 10.1038/nature03895.

Ishikawa, R., C. C. Castillo et D. Q. Fuller (2020). “Genetic evaluation of domestication-related traits in rice : implications for the archaeobotany of rice origins”. en. In : *Archaeological and Anthropological Sciences* 12.8, p. 197. doi : 10.1007/s12520-020-01112-3.

Ito, H. *et al.* (2011). “An siRNA pathway prevents transgenerational retrotransposition in plants subjected to stress”. en. In : *Nature* 472.7341, p. 115-119. doi : 10.1038/nature09861.

Jiang, N. et O. Panaud (2013). “Transposable Element Dynamics in Rice and Its Wild Relatives”. en. In : *Genetics and Genomics of Rice*. Sous la dir. de Q. Zhang et R. A. Wing. New York, NY : Springer New York, p. 55-69. doi : 10.1007/978-1-4614-7903-1_5.

Jiang, N. *et al.* (2002). “Dasheng : A Recently Amplified Nonautonomous Long Terminal Repeat Element That Is a Major Component of Pericentromeric Regions in Rice”. en. In : p. 13.

Jordan, I. *et al.* (2003). “Origin of a substantial fraction of human regulatory sequences from transposable elements”. en. In : *Trends in Genetics* 19.2, p. 68-72. doi : 10.1016/S0168-9525(02)00006-9.

Kawase, M., K. Fukunaga et K. Kato (2005). “Diverse origins of waxy foxtail millet crops in East and Southeast Asia mediated by multiple transposable element insertions”. en. In : *Molecular Genetics and Genomics* 274.2, p. 131-140. doi : 10.1007/s00438-005-0013-8.

Knip, M., S. de Pater et P. J. Hooykaas (2012). “The SLEEPER genes : a transposase-derived angiosperm-specific gene family”. en. In : *BMC Plant Biology* 12.1, p. 192. doi : 10.1186/1471-2229-12-192.

Kobayashi, S. (2004). “Retrotransposon-Induced Mutations in Grape Skin Color”. en. In : *Science* 304.5673, p. 982-982. doi : 10.1126/science.1095011.

Kofler, R. (2016). “PoPoolationTE2 : Comparative Population Genomics of Transposable Elements Using Pool-Seq”. en. In : p. 6.

Kolmogorov, M. *et al.* (2019). “Assembly of long, error-prone reads using repeat graphs”. en. In : *Nature Biotechnology* 37.5, p. 540-546. doi : 10.1038/s41587-019-0072-8.

Koren, S. *et al.* (2017). “Canu : scalable and accurate long-read assembly via adaptive *k* -mer weighting and repeat separation”. en. In : *Genome Research* 27.5, p. 722-736. doi : 10.1101/gr.215087.116.

Lanciano, S. et M. Mirouze (2018). “Transposable elements : all mobile, all different, some stress responsive, some adaptive ?” en. In : *Current Opinion in Genetics & Development* 49, p. 106-114. doi : 10.1016/j.gde.2018.04.002.

Lanciano, S. *et al.* (2017). “Sequencing the extrachromosomal circular mobilome reveals retrotransposon activity in plants”. en. In : *PLOS Genetics* 13.2. Sous la dir. de C. Feschotte, e1006630. doi : 10.1371/journal.pgen.1006630.

Landis, J. B. *et al.* (2018). "Impact of whole-genome duplication events on diversification rates in angiosperms". en. In : *American Journal of Botany* 105.3, p. 348-363. doi : 10.1002/ajb2.1060.

Langmead, B. et S. L. Salzberg (2012). "Fast gapped-read alignment with Bowtie 2". en. In : *Nature Methods* 9.4, p. 357-359. doi : 10.1038/nmeth.1923.

Li, C. (2006). "Rice Domestication by Reducing Shattering". en. In : *Science* 311.5769, p. 1936-1939. doi : 10.1126/science.1123604.

Li, H. (2018). "Minimap2 : pairwise alignment for nucleotide sequences". en. In : *Bioinformatics* 34.18, p. 3094-3100. doi : 10.1093/bioinformatics/bty191.

Lisch, D. (2002). "Mutator transposons". en. In : *Trends in Plant Science* 7.11, p. 498-504. doi : 10.1016/S1360-1385(02)02347-6.

— (2013). "How important are transposons for plant evolution ?" en. In : *Nature Reviews Genetics* 14.1, p. 49-61. doi : 10.1038/nrg3374.

Liu, N. *et al.* (2018). "Selective silencing of euchromatic L1s revealed by genome-wide screens for L1 regulators". en. In : *Nature* 553.7687, p. 228-232. doi : 10.1038/nature25179.

Lu, C. *et al.* (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". en. In : *Molecular Biology and Evolution* 29.3, p. 1005-1017. doi : 10.1093/molbev/msr282.

Lu, L. *et al.* (2017). "Tracking the genome-wide outcomes of a transposable element burst over decades of amplification". en. In : *Proceedings of the National Academy of Sciences* 114.49, E10550-E10559. doi : 10.1073/pnas.1716459114.

Lynch, V. J. *et al.* (2011). "Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals". en. In : *Nature Genetics* 43.11, p. 1154-1159. doi : 10.1038/ng.917.

Ma, J. (2004). "Analyses of LTR-Retrotransposon Structures Reveal Recent and Rapid Genomic DNA Loss in Rice". en. In : *Genome Research* 14.5, p. 860-869. doi : 10.1101/gr.1466204.

Marrano, A. *et al.* (2020). "High-quality chromosome-scale assembly of the walnut (Juglans regia L.) reference genome". en. In : *GigaScience* 9.5, giaa050. doi : 10.1093/gigascience/giaa050.

Maumus, F. et H. Quesneville (2016). "Impact and insights from ancient repetitive elements in plant genomes". en. In : *Current Opinion in Plant Biology* 30, p. 41-46. doi : 10.1016/j.pbi.2016.01.003.

McClintock, B. (1953). "Induction of Instability at Selected Loci in Maize". eng. In : *Genetics* 38.6, p. 579-599.

Merret, R. *et al.* (2015). "Heat-induced ribosome pausing triggers mRNA co-translational decay in Arabidopsis thaliana". en. In : *Nucleic Acids Research* 43.8, p. 4121-4132. doi : 10.1093/nar/gkv234.

Merret, R. *et al.* (2017). "Heat Shock Protein HSP101 Affects the Release of Ribosomal Protein mRNAs for Recovery after Heat Shock". en. In : *Plant Physiology* 174.2, p. 1216-1225. doi : 10.1104/pp.17.00269.

Michael, T. P. *et al.* (2018). "High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell". en. In : *Nature Communications* 9.1, p. 541. doi : 10.1038/s41467-018-03016-2.

Montacié, C. *et al.* (2017). "Nucleolar Proteome Analysis and Proteasomal Activity Assays Reveal a Link between Nucleolus and 26S Proteasome in A. thaliana". In : *Frontiers in Plant Science* 8, p. 1815. doi : 10.3389/fpls.2017.01815.

Murat, F. *et al.* (2015). "Understanding Brassicaceae evolution through ancestral genome reconstruction". en. In : *Genome Biology* 16.1, p. 262. doi : 10.1186/s13059-015-0814-y.

Nelson, M. G., R. S. Linheiro et C. M. Bergman (2017). "McClintock : An Integrated Pipeline for Detecting Transposable Element Insertions in Whole-Genome Shotgun Sequencing Data". en. In : *G3: Genes|Genomes|Genetics* 7.8, p. 2763-2778. doi : 10.1534/g3.117.043893.

Neumann, P. *et al.* (2006). "Significant Expansion of *Vicia pannonica* Genome Size Mediated by Amplification of a Single Type of Giant Retroelement". en. In : *Genetics* 173.2, p. 1047-1056. doi : 10.1534/genetics.106.056259.

Nishihara, H. (2019). "Transposable elements as genetic accelerators of evolution : contribution to genome size, gene regulatory network rewiring and morphological innovation". en. In : *Genes & Genetic Systems* 94.6, p. 269-281. doi : 10.1266/ggs.19-00029.

Novák, P. *et al.* (2020). "Repeat-sequence turnover shifts fundamentally in species with large genomes". en. In : *Nature Plants* 6.11, p. 1325-1329. doi : 10.1038/s41477-020-00785-x.

Ong-Abdullah, M. *et al.* (2015). "Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm". en. In : *Nature* 525.7570, p. 533-537. doi : 10.1038/nature15365.

Orgel, L., F. Crick et C. Sapienza (1980). "Selfish DNA". en. In : *Nature* 288.5792, p. 645-646. doi : 10.1038/288645a0.

Perumal, S. *et al.* (2020). *High contiguity long read assembly of Brassica nigra allows localization of active centromeres and provides insights into the ancestral Brassica genome.* en. preprint. Genomics. doi : 10.1101/2020.02.03.932665.

Picault, N. *et al.* (2009). "Identification of an active LTR retrotransposon in rice". en. In : *The Plant Journal* 58.5. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-313X.2009.03813.x, p. 754-765. doi : 10.1111/j.1365-313X.2009.03813.x.

Piegu, B. *et al.* (2006). "Doubling genome size without polyploidization : Dynamics of retrotransposition-driven genomic expansions in Oryza australiensis, a wild relative of rice". en. In : *Genome Research* 16.10, p. 1262-1269. doi : 10.1101/gr.5290206.

Plasterk, R. H., Z. Izsvák et Z. Ivics (1999). "Resident aliens : the Tc1/ mariner superfamily of transposable elements". en. In : *Trends in Genetics* 15.8, p. 326-332. doi : 10.1016/S0168-9525(99)01777-1.

Pont, C. *et al.* (2019). "Paleogenomics : reconstruction of plant evolutionary trajectories from modern and ancient DNA". en. In : *Genome Biology* 20.1, p. 29. doi : 10.1186/s13059-019-1627-1.

Pontier, D. *et al.* (2019). "The m ^{6} A pathway protects the transcriptome integrity by restricting RNA chimera formation in plants". en. In : *Life Science Alliance* 2.3, e201900393. doi : 10.26508/lsa.201900393.

Pontvianne, F. *et al.* (2016). "Identification of Nucleolus-Associated Chromatin Domains Reveals a Role for the Nucleolus in 3D Organization of the A. thaliana Genome". en. In : *Cell Reports* 16.6, p. 1574-1587. doi : 10.1016/j.celrep.2016.07.016.

Purugganan, M. D. (2019). "Evolutionary Insights into the Nature of Plant Domestication". en. In : *Current Biology* 29.14, R705-R714. doi : 10.1016/j.cub.2019.05.053.

Qin, P. *et al.* (2021). "Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations". en. In : *Cell*, S009286742100581X. doi : 10.1016/j.cell.2021.04.046.

Quadrana, L. *et al.* (2016). "The Arabidopsis thaliana mobilome and its impact at the species level". en. In : *eLife* 5, e15716. doi : 10.7554/eLife.15716.

Quadrana, L. *et al.* (2019). "Transposition favors the generation of large effect mutations that may facilitate rapid adaption". en. In : *Nature Communications* 10.1, p. 3421. doi : 10.1038/s41467-019-11385-5.

Quesneville, H. (2020). "Twenty years of transposable element analysis in the Arabidopsis thaliana genome". en. In : *Mobile DNA* 11.1, p. 28. doi : 10.1186/s13100-020-00223-x.

Quinlan, A. R. et I. M. Hall (2010). "BEDTools : a flexible suite of utilities for comparing genomic features". en. In : *Bioinformatics* 26.6, p. 841-842. doi : 10.1093/bioinformatics/btq033.

Rey, O. *et al.* (2016). "Adaptation to Global Change : A Transposable Element–Epigenetics Perspective". en. In : *Trends in Ecology & Evolution* 31.7, p. 514-526. doi : 10.1016/j.tree.2016.03.013.

Rigal, M. et O. Mathieu (2011). "A "mille-feuille" of silencing : Epigenetic control of transposable elements". en. In : *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* 1809.8, p. 452-458. doi : 10.1016/j.bbagrm.2011.04.001.

Rishishwar, L., L. Mariño-Ramírez et I. K. Jordan (2017). "Benchmarking computational tools for polymorphic transposable element detection". In : *Briefings in Bioinformatics* 18.6, p. 908-918. doi : 10.1093/bib/bbw072.

Rodić, N. et K. H. Burns (2013). "Long Interspersed Element–1 (LINE-1) : Passenger or Driver in Human Neoplasms ?" en. In : *PLoS Genetics* 9.3. Sous la dir. de S. M. Rosenberg, e1003402. doi : 10.1371/journal.pgen.1003402.

Ruan, J. et H. Li (2020). "Fast and accurate long-read assembly with wtdbg2". en. In : *Nature Methods* 17.2, p. 155-158. doi : 10.1038/s41592-019-0669-3.

Ruprecht, K. *et al.* (2008). "Endogenous retroviruses : Endogenous retroviruses and cancer". en. In : *Cellular and Molecular Life Sciences* 65.21, p. 3366-3382. doi : 10.1007/s00018-008-8496-1.

Sabot, F. et A. H. Schulman (2006). "Parasitism and the retrotransposon life cycle in plants : a hitchhiker's guide to the genome". en. In : *Heredity* 97.6, p. 381-388. doi : 10.1038/sj.hdy.6800903.

Sabot, F. *et al.* (2011). "Transpositional landscape of the rice genome revealed by paired-end mapping of high-throughput re-sequencing data : Transposition in rice". en. In : *The Plant Journal* 66.2, p. 241-246. doi : 10.1111/j.1365-313X.2011.04492.x.

Sanchez, D. H. *et al.* (2017). "High-frequency recombination between members of an LTR retrotransposon family during transposition bursts". en. In : *Nature Communications* 8.1, p. 1283. doi : 10.1038/s41467-017-01374-x.

SanMiguel, P. *et al.* (1998). "The paleontology of intergene retrotransposons of maize". en. In : *Nature Genetics* 20.1, p. 43-45. doi : 10.1038/1695.

Schulman, A. H. (2013). "Retrotransposon replication in plants". en. In : *Current Opinion in Virology* 3.6, p. 604-614. doi : 10.1016/j.coviro.2013.08.009.

Scott, E. C. *et al.* (2016). "A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer". en. In : *Genome Research* 26.6, p. 745-755. doi : 10.1101/gr.201814.115.

Secco, D. *et al.* (2015). "Stress induced gene expression drives transient DNA methylation changes at adjacent repetitive elements". en. In : *eLife* 4, e09343. DOI : 10.7554/eLife.09343.

Sedlazeck, F. J. *et al.* (2018). "Accurate detection of complex structural variations using single molecule sequencing". In : *Nature methods* 15.6, p. 461-468. DOI : 10.1038/s41592-018-0001-7.

Shafin, K. *et al.* (2020). "Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes". en. In : *Nature Biotechnology* 38.9, p. 1044-1053. DOI : 10.1038/s41587-020-0503-6.

Shen, W. *et al.* (2016). "SeqKit : A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation". en. In : *PLOS ONE* 11.10. Sous la dir. de Q. Zou, e0163962. DOI : 10.1371/journal.pone.0163962.

Sonnhammer, E. L. et R. Durbin (1995). "A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis". en. In : *Gene* 167.1-2, GC1-GC10. DOI : 10.1016/0378-1119(95)00714-8.

Stein, J. C. *et al.* (2018). "Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus Oryza". en. In : *Nature Genetics* 50.2, p. 285-296. DOI : 10.1038/s41588-018-0040-0.

Stuart, T. *et al.* (2016). "Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation". en. In : *eLife* 5, e20777. DOI : 10.7554/eLife.20777.

Studer, A. *et al.* (2011). "Identification of a functional transposon insertion in the maize domestication gene tb1". en. In : *Nature Genetics* 43.11, p. 1160-1163. DOI : 10.1038/ng.942.

The 3,000 rice genomes project (2014). "The 3,000 rice genomes project". en. In : *GigaScience* 3.1, p. 7. DOI : 10.1186/2047-217X-3-7.

Vitte, C., O. Panaud et H. Quesneville (2007). "LTR retrotransposons in rice (Oryza sativa, L.) : recent burst amplifications followed by rapid DNA loss". In : *BMC Genomics* 8.1, p. 218. DOI : 10.1186/1471-2164-8-218.

Wang, W. *et al.* (2018). "Genomic variation in 3,010 diverse accessions of Asian cultivated rice". en. In : *Nature* 557.7703, p. 43-49. DOI : 10.1038/s41586-018-0063-9.

Wicker, T., J. P. Buchmann et B. Keller (2010). "Patching gaps in plant genomes results in gene movement and erosion of colinearity". en. In : *Genome Research* 20.9, p. 1229-1237. DOI : 10.1101/gr.107284.110.

Wicker, T. *et al.* (2007). "A unified classification system for eukaryotic transposable elements". en. In : *Nature Reviews Genetics* 8.12, p. 973-982. DOI : 10.1038/nrg2165.

Xia, L. *et al.* (2017). "Rice Expression Database (RED) : An integrated RNA-Seq-derived gene expression database for rice". en. In : *Journal of Genetics and Genomics* 44.5, p. 235-241. DOI : 10.1016/j.jgg.2017.05.003.

Xiao, H. *et al.* (2008). "A Retrotransposon-Mediated Gene Duplication Underlies Morphological Variation of Tomato Fruit". en. In : *Science* 319.5869, p. 1527-1530. DOI : 10.1126/science.1153040.

Xin, Y. *et al.* (2019). "Amplification of miniature inverted-repeat transposable elements and the associated impact on gene regulation and alternative splicing in mulberry (Morus notabilis)". en. In : *Mobile DNA* 10.1, p. 27. DOI : 10.1186/s13100-019-0169-0.

Xu, Z.-S. *et al.* (2019). "Changing Carrot Color : Insertions in *DcMYB7* Alter the Regulation of Anthocyanin Biosynthesis and Modification". en. In : *Plant Physiology* 181.1, p. 195-207. DOI : 10.1104/pp.19.00523.

Zhang, T. *et al.* (2019). “Genome of *Crucihimalaya himalaica* , a close relative of *Arabidopsis* , shows ecological adaptation to high altitude”. en. In : *Proceedings of the National Academy of Sciences* 116.14, p. 7137-7146. doi : 10.1073/pnas.1817580116.

Zhao, K. *et al.* (2011). “Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa”. en. In : *Nature Communications* 2.1, p. 467. doi : 10.1038/ncomms1467.

Zhao, Q. *et al.* (2018). “Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice”. en. In : *Nature Genetics* 50.2, p. 278-284. doi : 10.1038/s41588-018-0041-z.

Zhou, A., T. Lin et J. Xing (2019). “Evaluating nanopore sequencing data processing pipelines for structural variation identification”. en. In : *Genome Biology* 20.1, p. 237. doi : 10.1186/s13059-019-1858-1.

Zhou, Y. *et al.* (2020). “A platinum standard pan-genome resource that represents the population structure of Asian rice”. en. In : *Scientific Data* 7.1, p. 113. doi : 10.1038/s41597-020-0438-2.