Contribution des éléments transposables à l'adaptabilité de ravageurs de cultures en absence de reproduction sexuée

Djampa Kozlowski

UNIVERSITÉ CÔTE D'AZUR

ÉCOLE DOCTORALE
SCIENCES DE LA VIE ET DE LA SANTÉ

THÈSE DE DOCTORAT

Contribution des éléments transposables à l'adaptabilité de ravageurs de cultures en absence de reproduction sexuée.

Djampa KOZLOWSKI

Institut Sophia-Agrobiotech,
UMR INRAE 1355, CNRS 7254
Equipe « **Interaction Plante-Nématode** »

Présentée en vue de l'obtention du grade de docteur en Biologie des interactions et écologie d'Université Côte d'Azur
Dirigée par : Etienne DANCHIN
Co-encadrée par : Marc BAILLY-BECHET
Soutenue le : 13/11/2020

Devant le jury, composé de :
Dr. Gianni LITI, DR, IRCAN, Président du jury
Dr. Emmanuelle LERAT, CR, LBBE, Rapporteure
Dr. Clément GILBERT, CR, LEGC, Rapporteur
Dr. Pierre PONTAROTTI, DR, IHU, Examinateur
Dr. Hadi QUESNEVILLE, DR, URGI, Examinateur
Dr. Didier FORCIOLI, MCU, UCA, Examinateur
Dr. Marc BAILLY-BECHET, MCU, UCA, Co-encadrant de thèse
Dr. Etienne DANCHIN, DR, ISA, Directeur de thèse

INRAE
la science pour la vie, l'humain, la terre

Contribution des éléments transposables à l'adaptabilité de ravageurs de cultures en absence de reproduction sexuée

Jury :

Président du jury
Gianni LITI, DR, IRCAN (Nice, France)

Rapporteurs
Emmanuelle LERAT, CR, LBBE (Lyon, France)
Clément GILBERT, CR, LEGC (Gif sur Yvette, France)

Examinateurs
Pierre PONTAROTTI, DR, IHU (Marseille, France)
Hadi QUESNEVILLE, DR, URGI (Versailles, France)
Didier FORCIOLI, MCU, UCA (Nice, France)

Co-encadrant de thèse
Marc BAILLY-BECHET, MCU, UCA (Nice, France)

Directeur de thèse
Etienne DANCHIN, DR, ISA (Sophia Antipolis, France)

Résumé

Les nématodes à galles (genre *Meloidogyne*) sont parmi les parasites de plantes les plus nuisibles. Ces organismes se distinguent par la diversité de leurs modes de reproduction. Étonnement, il a été observé que les espèces les plus néfastes se reproduisent de manière strictement asexuée et certaines sont capables de contourner la résistance de la plante hôte en un nombre de générations restreint. Ainsi, bien qu'incapables de combiner des mutations bénéfiques provenant de différents individus, ces espèces peuvent s'adapter à des changements du milieu. L'adaptabilité et le succès parasitaire de ces espèces malgré l'absence de reproduction sexuée semblent paradoxaux et doivent reposer sur d'autres mécanismes capables de générer de la plasticité génétique.

Les éléments transposables (ETs) sont des fragments d'ADN capables de se déplacer et de se multiplier dans les génomes. De ce fait, les ETs peuvent avoir des répercussions fonctionnelles et structurales sur les génomes. Les ETs pourraient constituer un des mécanismes permettant de générer la diversité génétique nécessaire à l'adaptabilité chez *Meloidogyne*.

En réalisant une analyse de génomique comparative entre 7 espèces de *Meloidogyne*, j'ai montré que le contenu en ETs actuellement observé chez ces espèces semble suivre leur histoire évolutive et la dérive entre espèces, plutôt que des traits d'histoires de vie tels que le mode de reproduction. Par ailleurs, cette analyse soutient une activité récente des ETs au sein de la plupart des espèces. Ces résultats suggèrent que bien que les ETs aient récemment été actifs au sein du genre *Meloidogyne*, leur dynamique dans les génomes semble spécifique à chaque espèce et nécessite donc une étude ciblée.

Dans cette optique, j'ai concentré mes efforts sur *M. incognita*, l'espèce à reproduction asexuée la plus préjudiciable pour l'agriculture. Dans un premier temps, j'ai annoté en détail le contenu en ETs du génome de *M. incognita*. L'analyse du contenu en ETs a confirmé que ces éléments ont probablement été récemment actifs dans le génome. Afin de mieux caractériser cette activité et ses potentiels effets, j'ai ensuite estimé la mobilité de ces ETs via une analyse de génomique comparative portant sur 12 isolats géographiques. J'ai pu identifier plusieurs milliers de loci dans le génome où les fréquences de présence d'ETs varient entre les différents isolats. Par une approche phylogénétique, j'ai montré que ces variations de fréquence d'ETs suivent l'histoire évolutive des isolats étudiés. Par rapport au génome de référence, j'ai prédit des néo-insertions d'ETs, certaines ayant un potentiel impact fonctionnel. Les validations expérimentales réalisées pour plusieurs de ces insertions confirment le rôle probable des ETs dans la plasticité du génome de cette espèce.

Lors de cette analyse, j'ai également identifié des ETs présents à des fréquences intermédiaires (différentes de 0 ou 1) au sein de chaque isolat, signe d'une variabilité entre individus. Or *M. incognita* est un organisme supposé clonal et chaque isolat étudié est issu d'une seule femelle. En nous concentrant sur l'analyse d'un de ces isolats, nous avons validé expérimentalement plusieurs polymorphismes de présence, ce qui confirme qu'il existe une hétérogénéité génétique non négligeable au sein d'un même isolat. Par ailleurs, en comparant des données de séquençage issues du même isolat à deux points de cinétique différents, nous avons pu prédire que quelques ETs varient en fréquences au sein de l'isolat en un faible nombre de générations, ce qui sous-entend que ces ETs participent à la dynamique de la diversité génétique de cet organisme.

Ces résultats posent les bases pour de futures analyses visant à déterminer si l'activité des ETs joue un rôle actif dans la capacité d'espèces à s'adapter à leur environnement en absence de reproduction sexuée.

Mots clefs : Eléments transposables, bio-informatique, reproduction asexuée, plasticité génomique, *Meloidogyne*.

Abstract

Root-knot nematodes (genus *Meloidogyne*) are among the most devastating plant parasites. These organisms present an important diversity of reproductive modes. Surprisingly, it has been observed that the most damaging species reproduce in a strictly asexual manner and some can bypass the host plant's resistance in a limited number of generations. Thus, although being unable to combine beneficial mutations from different individuals, these species can adapt to environmental changes. The adaptability and parasitic success of these species despite the absence of sexual reproduction seem paradoxical and must rely on other mechanisms capable of generating genetic plasticity.

Transposable Elements (TEs) are DNA fragments capable of moving and multiplying in genomes. As a result, TEs can have functional and structural repercussions on genomes. Hence, TEs could be one of the mechanisms involved in generating the genetic diversity necessary for adaptability in Meloidogyne. By performing a comparative genomics analysis between 7 *Meloidogyne* species, I have shown that the TE landscape currently observed in these species seems to follow their evolutionary history and interspecies drift rather than life-history traits such as the reproduction mode. Furthermore, this analysis supports recent TE activity within all these species. The results also suggest that although TEs have recently been active within the genus *Meloidogyne*, their dynamics in the genomes appear to be species-specific and thus require targeted study.

With this in mind, I have focused my efforts on *M. incognita*, arguably the most detrimental asexually reproductive species to agriculture. As a first step, I have annotated in detail the TE content in the genome of *M. incognita*. The TE content analysis confirmed these elements have probably been recently active in the genome. To better characterize this activity and its potential effects, I then estimated the mobility of these TEs through a comparative genomics analysis of 12 geographic isolates. I was able to identify several thousand loci in the genome where the frequencies of TE presence varied substantially between different isolates. Using a phylogenetic approach, I showed that these TE frequency variations followed the evolutionary history of the studied isolates. Compared to the reference genome, I have predicted TE neo-insertions, some with potential functional impact. Experimental validations carried out for several of these insertions confirmed the potential role of TEs in the genome plasticity in this species.

During this analysis, I also identified TEs present at intermediate frequencies (different from 0 or 1) within each isolate, indicating variability between individuals despite the fact *M. incognita* is a supposedly clonal organism and that each isolate studied was derived from a single female. Focusing on the analysis of one of these isolates, we have experimentally validated several TE polymorphisms, confirming that there is significant genetic heterogeneity within the same isolate. Furthermore, by comparing sequencing data from the same isolate at two different time points, we predicted that a few TEs varied in frequency within the isolate within a small number of generations, implying these TEs participate in the dynamics of genetic diversity in this organism.

These results lay the foundation for future analyses to determine whether TEs play an active role in the ability of species to adapt to their environment in the absence of sexual reproduction.

Key words : Transposable elements, computational biology, asexual reproduction, genomic plasticity, *Meloidogyne*.

Remerciements

Les remerciements, c'est le point de non-retour, l'étape qui marque la fin de l'aventure. Et quelle aventure ! Ces trois années resteront sans doute l'une des périodes les plus riches de ma vie, tant sur le plan personnel que professionnel. J'en ressors transformé et je ne peux qu'en être reconnaissant à toutes les personnes ayant croisé mon chemin.
Je tiens par avance à préciser aux lecteurs que témoigner de son affection de manière publique est un exercice de style particulièrement compliqué pour moi. Partez donc du principe que chaque mot que je pourrais écrire ici doit voir son intensité multipliée par 10 (minimum) pour refléter ma pensée.

Je remercie de manière générale l'ensemble des membres de l'équipe IPN et tout particulièrement Pierre ADAD pour m'avoir accueilli en leur sein.

J'aimerais ensuite remercier l'ensemble des membres de mon jury pour avoir accepté d'évaluer mes travaux ainsi que pour leurs retours constructifs.

Je remercie tout particulièrement mes directeurs de thèses Etienne DANCHIN et Marc BAILLY-BECHET qui m'ont permis de profiter du meilleur encadrement que j'aurai pu espérer. Vous êtes des modèles de rigueur et de sérieux scientifique. Au cours de ces 3 années, vous m'avez au quotidien transmis votre passion pour la recherche et l'enseignement et avez réussi me faire considérer qu'il s'agit des meilleurs métiers du monde. Vous m'avez d'abord accompagné au quotidien puis, petit à petit, m'avez laissé de plus en plus d'autonomie. Je ne peux que vous remercier pour ce témoignage de confiance qui m'a permis de m'épanouir, de prendre confiance en mes capacités, et maintenant de prendre mon envol. Merci. Merci enfin d'avoir été disponibles dès que besoin, d'avoir su gérer mes moments de stress, et de m'avoir épaulé dans les moments un peu moins joyeux. Je vous estime et vous respecte au plus haut point, vous être les enseignants et les chercheurs que j'aimerais être si je continue dans ces voies

Corinne RANCUREL et Martine DA ROCHA, vous faites partie des personnes les plus douces et les plus gentilles que je connaisse (en plus d'être les meilleures ingé' bio-info de la terre !). Travailler avec vous a été un plaisir au quotidien. Restez telles que vous êtes ! Dominique COLINET, ça a été un plaisir de découvrir l'enseignement avec toi. Les facs mériteraient bien plus d'enseignants aussi passionnés et dédiés à leur métier que toi.

Merci à l'ensemble des personnes avec lesquelles j'ai eu la chance d'avoir des échanges scientifiques passionnants. Je pense tout particulièrement à Aurélie SEASSAU, Marc MAGLIANO, Laura PERROT et Laetitia ZURLETTO.

Je remercie également les membres de l'URGI et tout particulièrement Véronique JAMILLOUX, Joelle ANSELHEM, & Hadi QUESNVILLE pour leur temps et leurs conseils.

Arrive maintenant les remerciements aux amis et la liste est longue. Difficile de mettre ces gens dans des cases, d'autant plus que les personnes à venir sont toutes aussi exceptionnelles les unes que les autres.

Merci à tous les copains de l'INRA pour votre bonne humeur au quotidien et pour toutes les discussions enflammées que nous avons pu avoir. Tout d'abord merci aux deux amis Joffrey et Georgios pour m'avoir rendu accro au café, aux rimes douteuses (Joffrey) et aux débats incessants (Georgios). Merci à Rahim, mon binôme expérimentaliste de choc, tant pour les moments à se gratter la tête sur des expériences infructueuses que pour les moments passés en dehors. Merci Carole pour ces moments passés à discuter de

tout et de rien et pour toutes tes petites attentions aux moments les plus propices. Merci Danila, Lucie et Camille pour votre bonne humeur et pour tous les moments de rigolade passés ensemble.

Merci à mes deux « coupines » de galère Marie et Marion. Vous représentez tant de choses pour moi et si j'en suis là aujourd'hui c'est en grande partie grâce à vous deux. Merci aussi à Léo. Merci pour ce voyage, merci pour ton écoute dans les moments difficiles. C'est en partie grâce à toi que je me suis lancé dans cette aventure. Merci aux amis de la « team ZEAZY» pour tous les moments IRLs passés ensemble et leur mauvaise foi on-line : Alexis, Pierre, Steve, Simon. Mention spéciale à Alexis pour avoir été présent pendant toutes ces années et de m'avoir accueilli pendant ce qui a dû être trois longues semaines pour lui au début de l'écriture de ce manuscrit. (« Trop la honte ce mec ! »). Merci à tous les amis un peu moins connectés mais tout aussi importants pour moi pour ce qu'ils sont et ce qu'ils représentent : un grand bol d'air frais à chaque inspiration et une amitié indéfectible. Par ordre alphabétique (donc pas de jaloux !) : Amaury, Cas, Célim, César, Djé, Emeric, Eric, Flo & Samson. Merci à Franck et Nathalie pour les weekends « cochon » (en tout bien tout honneur). Vous êtes des personnes magnifiques.

Merci à Tatiana MROZINSKY de m'avoir transmis sa passion pour la biologie et d'avoir été une si bonne professeure. Je souhaite a tout le monde de pouvoir avoir des cours aussi passionnants que ce que les tiens ont été pour moi.

Arrive enfin les remerciements à la famille. Ici encore, (et encore plus), il m'est difficile d'exprimer ce que je ressens pour vous et ce que vous représentez pour moi. J'aurais donc un style encore plus « nordique » que dans ce que vous avez lu précédemment. Merci d'abord à ma mère pour tout ce qu'elle s'est évertuée à me transmettre tant sur le plan personnel qu'intellectuel. Ce sont les valeurs de dépassement personnel que tu m'as donné qui m'ont permis d'avancer les jours un peu moins fructueux. Merci du fond du cœur. Merci à Chris', la plus belle des belles mères qui a su me dompter et m'accompagner. Ta douceur et ta bonne humeur font du bien au cœur ! Merci à Elo, la meilleure des belles sœurs dont on puisse rêver ! Tu es une personne extraordinaire, tout simplement.

Merci à mes frères (par ordre d'apparition) Iko, Neel, & Wado pour leur soutien indéfectible et tout leur amour. Vous êtes des rocs, des montagnes pour moi. Rien ne peut nous arrêter.

Enfin, il m'est impossible d'exprimer ici ce que je ressens pour ma moitié, mon brin complémentaire Marjorie. Je te dirais juste que tu as été le meilleur compagnon d'aventure possible pour celle-ci et que j'ai hâte d'en vivre de nouvelles avec toi. Je sais que tu me comprendras.

Cette thèse est dédiée à mon père et à ma grand-mère qui ont dû quitter l'aventure en cours de route. Rien de tout cela n'aurait été possible sans vous.

Table des matières

I – Contexte théorique et modèle d'étude

A- Reproduction asexuée : définition et implications dans l'évolution des organismes

En biologie, la reproduction est un processus qui permet la production de nouveaux individus à partir d'individu(s) existant(s). Le lecteur n'aura pas manqué de remarquer les « s » entre parenthèses dans la phrase précédente. La reproduction peut en effet faire intervenir un ou plusieurs individus suivant que l'on parle de reproduction asexuée ou sexuée. On notera que plutôt que de consister en un système binaire, la reproduction recouvre en réalité un continuum de modes intermédiaires pouvant même parfois être alternés au sein d'un organisme donné. Face à cette diversité et aux débats qu'elle engendre, il est difficile d'énoncer des généralités sans se perdre dans un vocabulaire et des concepts techniques. Les notions abordées dans cette introduction (majoritairement centrée sur la reproduction asexuée) se résumeront donc à celles nécessaires à la compréhension de ce manuscrit. Dans cette optique, nous nous concentrerons majoritairement sur les formes rencontrées chez les organismes eucaryotes pluricellulaires (voir Figure 1.A.1).

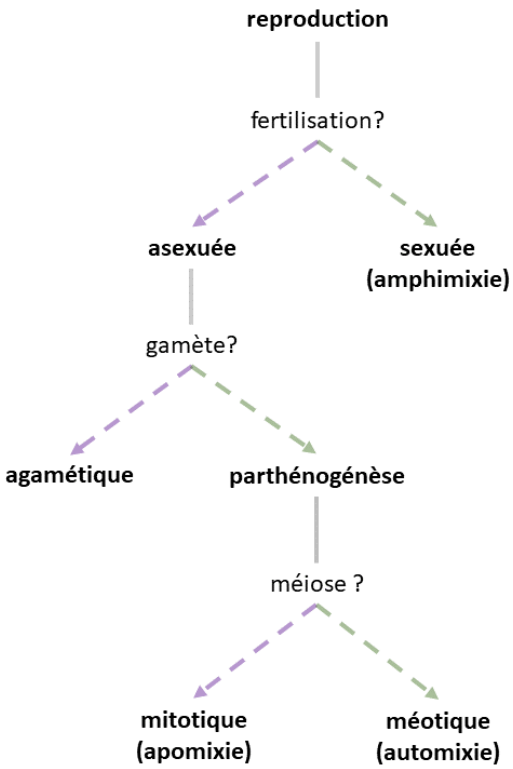


Figure 1.A.1 : diagramme schématique des modes de reproduction rencontrés chez les organismes eucaryotes.

Les distinctions entre les différents modes de reproduction sont représentées sous forme d'un arbre de décision, chaque nœud (question) permettant d'évoluer dans l'arbre. Les flèches en pointillé vertes correspondent à une réponse positive à la question, les flèches violettes à une réponse négative.

1 - Qu'est-ce que la reproduction asexuée ?

La reproduction sexuée, aussi appelée amphimixie, correspond à la fusion des matériels génétiques issus des gamètes mâles et femelles dans l'ovule fécondé, constituant le point de départ d'un nouvel individu (voir Figure 1.A.1). Du fait de la recombinaison méiotique lors de la gamétogenèse, suivi du brassage génétique engendré par la fécondation, l'individu produit sera génétiquement différent de ses parents. Ce mode de reproduction assure donc un flux d'allèles et le maintien d'une diversité génétique au sein des populations.

En opposition à la reproduction sexuée, la reproduction asexuée n'implique pas de fertilisation et donc de conjonction de matériel génétique de la part de plusieurs individus. La reproduction asexuée se décline en deux grandes voies : la reproduction agamétique et la reproduction parthénogénétique (voir Figure 1.A.1).

Brièvement, la reproduction agamétique (ou reproduction végétative en sciences végétales) décrit la formation de clones à partir de structures ou de cellules somatiques d'un individu et n'est donc pas liée à l'utilisation d'organes reproducteurs ou de gamètes (de Meeûs et al., 2007). Ce mode de reproduction recouvre des mécanismes variés.

Chez les plantes, le bouturage (création d'un nouvel individu à partir d'un fragment d'organe isolé comme un morceau de rameau ou de tige), le marcottage (branches qui s'enracinent et poussent comme de nouveaux individus), ou encore la création de bulbes ou de tubercules (organes de stockage qui peuvent produire de nouveaux individus), sont autant d'exemples de mécanismes possibles (de Meeûs et al., 2007). On notera tout de même que dans le cas où l'individu produit reste attaché à la plante mère, ce nouvel individu peut aussi être considéré comme une ramification d'un même individu ou encore comme une colonie clonale. C'est par exemple le cas de « pando », une colonie clonale de peupliers faux-trembles (*Populus tremuloides*) considéré comme l'organisme vivant le plus lourd et le plus âgé de la planète (Rogers and Gale, 2017).

Chez les métazoaires, la reproduction agamétique se manifeste aussi de nombreuses manières. On distinguera les formes actives et passives. Dans les formes passives, on retrouvera la fragmentation (un individu se forme à partir d'un fragment d'un autre individu), qui est commune chez les coraux et les éponges. En ce qui concerne les formes actives, on recense par exemple la lacération podale (un individu se détache de son pied et migre puis un nouvel individu se forme à partir du pied) chez les anémones de mer, la fission (un individu se sépare en n parties puis chaque nouvel individu reconstruit les parties manquantes) qui peut être observée chez les étoiles de mer ou encore le bourgeonnement qui est commun chez les cnidaires (de Meeûs et al., 2007).

A l'inverse de la reproduction agamétique, la parthénogénèse décrit quant à elle la production d'une descendance à partir d'une cellule germinale femelle mais sans contribution d'un gamète masculin. Au sens strict du terme, la parthénogénèse correspond donc au développement d'un nouvel organisme à partir d'un gamète femelle non fécondé (de Meeûs et al., 2007). Selon cette définition, la parthénogenèse peut être divisée en 3 classes : arrhénotoque, deutérotoque, et thélytoque (de Meeûs et

al., 2007). Afin de simplifier la compréhension de ces concepts, nous ne considérerons ici que des organismes diploïdes.

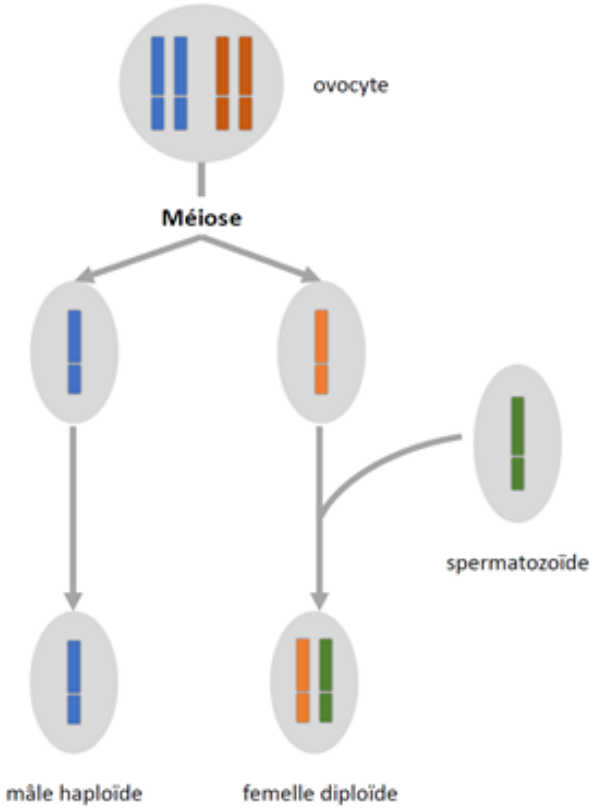


Figure 1.A.2 : parthénogénèse arrhénotoque.

Adapté et traduit de (Rabeling and Kronauer, 2013).

Lors d'une parthénogénèse arrhénotoque, l'œuf non fécondé conduira à la production de mâles, ces mâles étant nécessairement haploïdes et donc génétiquement différents de leur mère (voir Figure 1.A.2). L'œuf fécondé conduira quant à lui à la production de femelles, ce qui devient donc une forme de reproduction sexuée. La parthénogénèse arrhénotoque est donc un mode de reproduction asexuée facultatif. Ce mode de reproduction peut par exemple être observé chez les fourmis et les abeilles.

La parthénogénèse deutérotoque produit des mâles et des femelles mais l'apparition de mâles est cyclique et intervient à certaines périodes de l'année. Cette forme de reproduction fait alterner les phases parthénogénétiques et les phases sexuées en fonction des besoins ou des changements environnementaux tels que les saisons et ne peut donc, là encore, pas être considérée comme un mode de reproduction asexuée obligatoire. Ce mode de reproduction peut par exemple être observé chez les daphnies et les pucerons.

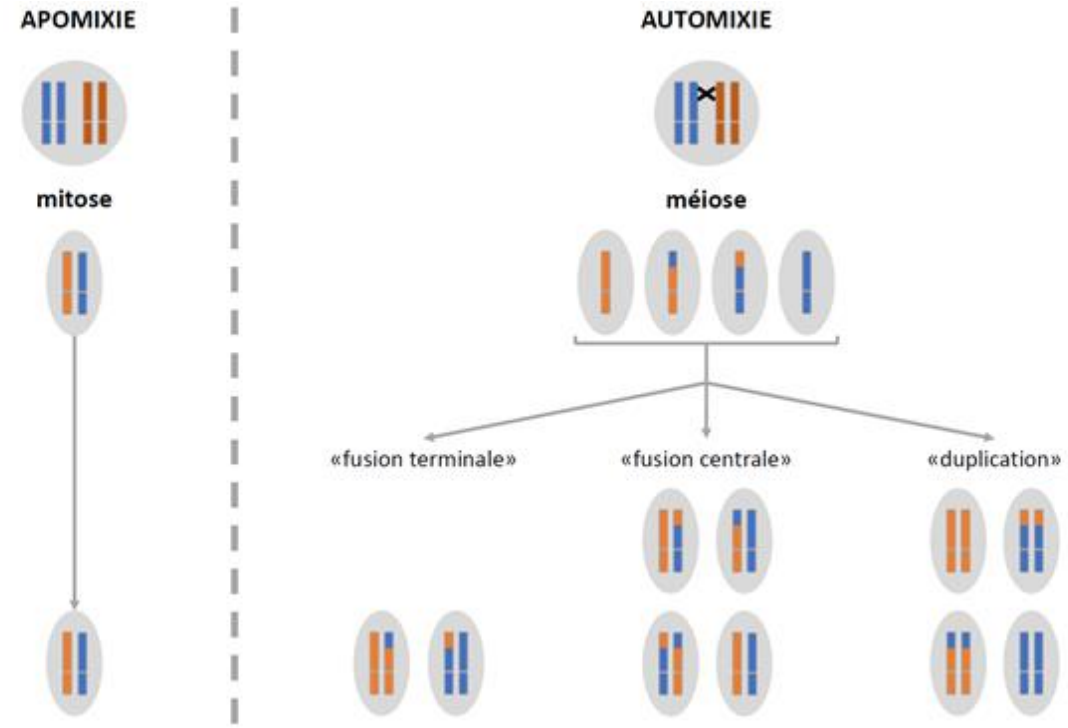


Figure 1.A.3 : parthénogenèse thélytoque.

Adapté et traduit de (Rabeling and Kronauer, 2013). L'ensemble des individus produits sont des femelles diploïdes. Les deux formes de parthénogénèse thélytoques sont séparées par le trait vertical gris. Pour l'automixie, la croix noire dans la cellule en haut symbolise un événement de recombinaison génétique.

La forme thélytoque de la parthénogénèse mène à la production de femelles à partir d'œufs non fécondés. Deux formes de parthénogenèse thélytoque existent et peuvent être distinguées par l'absence ou la présence d'un processus de méiose lors de la formation du zygote (voir Figures 1.A.1 et 1.A.3).

La parthénogénèse méiotique (aussi appelée automixie) fait intervenir le processus de méiose, la diploïdie étant rétablie sans l'apport d'un gamète (sans fécondation). La diploïdie de l'œuf peut être restaurée par i) fusion centrale (fusion des produits de la première division méiotique), ii) fusion terminale (fusion des produits de la seconde division méiotique) ou iii) duplication du génome avant la méiose, créant une cellule avec initialement 4n chromosomes (4 jeux de chromosomes). Il est important de noter que lors de la méiose, l'appariement des chromosomes homologues peut mener à des évènements de recombinaison génétique créant à un transfert d'information entre chromosomes. Ce transfert de matériel peut se faire par la rupture et la réunion de portions d'ADN entre chromosomes (principe de l'enjambement génétique ou « crossing-over ») ou bien sans échange physique, une partie du matériel génétique étant copiée d'un chromosome à l'autre sans que le chromosome donneur ne soit modifié. Ainsi la méiose permet l'échange de matériel génétique entre paires de chromosomes

homologues, ce qui peut conduire à un brassage intrachromosomique des allèles. La parthénogénèse méiotique peut donc potentiellement conduire à la production d'une progéniture génétiquement différente de la génitrice (voir Figure 1.A.3).

La parthénogénèse mitotique (aussi appelée apomixie) ne fait pas intervenir le processus de méiose lors de la production du gamète femelle. Les œufs (non fécondés) sont diploïdes et proviennent simplement d'ovocytes non réduits. L'absence de méiose évite quasiment tout réarrangement chromosomique. La descendance est donc génétiquement identique à la génitrice en tout point du génome à l'exception de sites ayant subi des mutations. Les individus produits par parthénogénèse mitotique peuvent donc être considérées comme des clones de la génitrice (voir Figure 1.A.3).

Dans notre définition, la reproduction asexuée peut être résumée comme la production d'un organisme à partir du matériel génétique d'un seul individu. Or, comme nous avons pu le voir, cette définition recouvre en réalité une diversité importante de modes, chacun ayant ses particularités propres. Le terme de reproduction asexuée nécessite donc d'être adjoint de qualificatifs supplémentaires.

Nous avons par exemple vu que les formes de parthénogénèses arrhénotoque et deutérotoque ont chacune conservé une composante sexuée à leur manière. L'introduction de phases sexuées, même limitée à l'un des deux genres, permet d'introduire de la diversité génétique dans la population via l'apport (même réduit) de matériel génétique externe (Bengtsson, 2009). Il est donc nécessaire de différencier les organismes se reproduisant obligatoirement de manière asexuée au cours du temps de ceux pour qui ce mode de reproduction est facultatif puisque comme nous le verrons dans le chapitre suivant, ceci a d'importantes conséquences sur l'évolution des organismes.

De plus, même en cas de reproduction asexuée obligatoire, l'absence de conjonction de matériel génétique de la part de plusieurs individus n'implique pas nécessairement que la descendance soit génétiquement identique à la génitrice. En effet, comme illustré en Figure 1.A.3, seule la parthénogénèse mitotique et certaines formes de parthénogénèse méiotique conduisent à la production d'une descendance clonale (de Meeûs et al., 2007). Seules ces formes de reproduction peuvent donc être considérées comme reproduction asexuée obligatoire stricte. Néanmoins, comme nous le verrons plus loin, même si de potentiels évènements de recombinaison peuvent induire de la plasticité génétique chez les organismes parthénogénétiques méiotiques obligatoires, leur impact à long terme sur la diversité génétique au sein de la population reste limité, ce brassage génétique se faisant toujours à partir du même matériel de base au fil des générations.

2 - Quelles sont les causes/origines de la reproduction asexuée ?

Il existe de nombreuses causes directes et indirectes à l'apparition de la reproduction asexuée dans une lignée.

On notera par exemple une association importante entre ce mode de reproduction et la polyploïdie (fait de posséder plus de deux ensembles de chromosomes homologues) et/ou l'hybridation (processus consistant à combiner différentes variétés d'organismes pour créer un hybride). Cela s'explique principalement par le fait que la méiose échoue souvent dans les organismes polyploïdes, en particulier dans ceux possédant un nombre impair de lots de chromosomes en raison des difficultés à apparier plus de deux chromosomes de chaque type (Bengtsson, 2009). Des difficultés à apparier les chromosomes homologues peuvent aussi intervenir lorsque ces chromosomes sont trop divergents. Ce type de phénomène est souvent observé en cas d'hybridation d'espèces suffisamment distantes, et peut là aussi mener à un mode de reproduction asexuée. L'hybridation peut aussi avoir comme conséquence d'altérer la régulation des voies de signalisation de la reproduction sexuée (Loxdale, 2010; Ozias-Akins and Conner, 2020). De plus, l'hybridation a souvent pour conséquence d'induire la polyploïdie, et donc de combiner leurs effets respectifs.

D'autres causes, plus directes, existent néanmoins. On peut par exemple citer l'existence de « supergènes » (groupes de gènes étroitement liés) qui peuvent agir chez certains végétaux angiospermes comme des déterminants chromosomiques de l'asexualité (Schwander et al., 2014). Certains partenaires symbiotiques peuvent aussi être des déterminants de l'asexualité. L'exemple le plus courant est la bactérie *Wolbachia* qui excelle dans les stratégies de manipulation de la reproduction de ses hôtes afin d'assurer sa transmission via la féminisation des populations hôtes. Chez certains hôtes hyménoptères, *Wolbachia* peut par exemple directement induire une parthénogenèse thélytoque en intervenant lors du premier cycle mitotique (Stouthamer et al., 1999).

3 - Quelles sont les conséquences génétiques et évolutives de la reproduction strictement asexuée ?

D'un point de vue génétique, les conséquences d'une reproduction asexuée sont multiples et peuvent varier suivant que l'organisme se reproduit obligatoirement de manière asexuée au cours du temps ou non et si cette parthénogenèse fait intervenir ou non, une étape de méiose.

Par exemple, lorsque l'asexualité est complète dans une lignée, il est prédit que les copies de gènes homologues divergent lentement au cours du temps mais sans limite jusqu'à ce que les copies au sein des individus deviennent plus divergentes que ne le sont normalement les allèles dans les populations sexuelles (Bengtsson, 2009; Ellegren and Galtier, 2016). Cette théorie, communément appelé « effet Meselson » avait été proposée pour expliquer la longévité exceptionnelle des rotifères bdelloïdes (Welch, 2000), des organismes ayant réussi à persister pendant des millions d'années malgré l'absence de reproduction sexuée; ce qui leur a valu d'être caractérisés de « scandales évolutifs » par Maynard Smith (1978). Néanmoins, la production et l'analyse du génome de cet organisme a permis de réfuter cette hypothèse ; la forte divergence observée ne se situant pas entre allèles d'un même gène mais entre gènes ohnologues (gènes paralogues issus d'un évènement de duplication complète du génome) (Flot et al., 2013).

Par ailleurs, chez les organismes à reproduction asexuée, les allèles ne peuvent diffuser dans les populations au-delà d'une seule et même lignée puisqu'il n'y a pas de fusion de gamètes provenant d'individus distincts (syngamie) (Glémin and Galtier, 2012). Ainsi, chez les organismes à reproduction asexuée, chaque lignée clonale évolue indépendamment et diverge donc progressivement des autres.

Néanmoins, la conséquence principale de la reproduction mitotique (i.e. strictement asexuée) est l'absence de recombinaison entre chromosomes homologues, et ce en tout point du génome (Rice, 2002).

La recombinaison permet l'échange de matériel génétique entre paires de chromosomes homologues, ce qui conduit à un brassage intrachromosomique des allèles et donc à la production d'une progéniture avec des combinaisons de caractéristiques différentes de celles que l'on trouve chez le ou les parent(s). Chez les eucaryotes, la recombinaison génétique intervient au cours de la méiose et implique l'appariement des chromosomes homologues. On notera que de la recombinaison peut également se produire pendant la mitose où elle implique les chromosomes frères formés après l'étape de réplication chromosomique. Néanmoins, dans ce cas, de nouvelles combinaisons d'allèles ne sont normalement pas produites puisque les chromosomes frères sont généralement identiques. La limitation ou l'absence de recombinaison génétique, comme c'est le cas chez les organismes à reproduction asexuée obligatoire, a d'importantes répercussions génétiques et évolutives.

L'un des principaux avantages théoriques de la recombinaison génétique concerne sa capacité à réduire la charge de mutation (aussi appelé "fardeau génétique" i.e. ensemble des mutations génétiques défavorables dans une population) (Rice, 2002). Dans les populations sexuées, les processus de recombinaison et de brassage génétique permettent aux génomes de la progéniture d'être différents de

ceux des parents, augmentant ainsi la diversité génétique. A l'inverse, en l'absence de recombinaison comme c'est le cas chez les espèces à reproduction strictement asexuée, la totalité du génome de la mère est transmis à l'identique à la progéniture (hors mutations inter-générations). Ainsi, en supposant que les mutations inverses soient rares, la progéniture supporte donc au moins la même charge de mutation que sa mère. Il est donc communément admis que l'absence de recombinaison rencontré chez les organismes à reproduction asexués entraîne une accumulation de mutations délétères (mutations nocives) de manière irréversible au cours du temps. Ce procédé est appelé « cliquet de Muller » (Muller, 1964). Des validations expérimentales de ce modèle ont été conduites chez des lignées isogéniques (tous les individus partagent le même patrimoine génétique) de la levure *Saccharomyces cerevisiae* et ont bien conclu à une accumulation de mutations ayant mené à l'extinction de certaines populations (Zeyl et al., 2001).

Le deuxième avantage théorique de la recombinaison est qu'elle permet d'unir dans le même génome des mutations favorables qui surviennent dans différentes lignées. En revanche, chez les espèces se reproduisant par clonage, les différentes mutations bénéfiques doivent se produire successivement, de manière cumulative au sein d'une même lignée pour s'unir dans un même génome, ce qui ralentit le taux d'accumulation des mutations bénéfiques (évolution progressive) (Rice, 2002). La combinaison d'allèles favorables est donc théoriquement plus lente chez les organismes asexués. Par ailleurs, lorsque différentes mutations bénéfiques sont présentes simultanément dans une population à reproduction asexuée, il est prédit que leur vitesse de fixation est réduite par l'interférence clonale, qui impose une "limite de vitesse" à leur fixation progressive dans la population. L'interférence clonale se produit parce que différentes mutations bénéfiques sont en concurrence les unes avec les autres, ce qui dilue leur avantage par rapport aux génomes qui ne portent pas de mutations bénéfiques (Rice, 2002).

Enfin, moins il y a de recombinaison, plus le déséquilibre de liaison (i.e. l'association préférentielle entre des allèles) sera important et maintenue dans les génomes via des contraintes mécaniques. Ainsi, dans les populations se reproduisant de manière strictement asexuée, l'ensemble des loci du génome sont liés et devraient donc partager la même destinée. La sélection à un locus entraînera avec lui l'ensemble du génome, un phénomène connu sous le nom « d'auto-stop ». Dans le cas où plusieurs loci liés seraient soumis à la sélection, l'effet de la sélection sur un locus interférera avec la sélection sur l'autre, ce qui peut donc réduire l'efficacité de la sélection (de Meeûs et al., 2007).

Pour résumer, la théorie prédit que l'absence de recombinaison génétique due à la reproduction strictement asexuée i) induit l'accumulation de mutations délétères au cours du temps, ii) réduit le potentiel adaptatif des organismes et iii) induit une diminution de l'efficacité de la sélection. De ce fait, il est communément considéré que sous ces effets délétères, le patrimoine génétique des organismes asexués obligatoire va progressivement se dégrader jusqu'à un point de non-retour ("genome decay").

Ainsi, dans le cas où les conditions du milieu viendraient à changer, l'organisme, dont la fitness (valeur sélective) a probablement déjà été affaiblie, ne sera pas en mesure de s'adapter à ses nouvelles conditions de vie. Aussi, il a régulièrement été fait l'hypothèse que la reproduction asexuée devrait majoritairement être rencontrée chez des organismes vivant dans des niches écologiques plus ou moins restreintes dont les conditions du milieu sont stables ("geographic parthenogenesis hypothesis").

4 - Occurrence, persistance et succès évolutif des animaux à reproduction asexuée

Chez les métazoaires, l'observation de la répartition des espèces dans l'arbre du vivant en fonction de leurs modes de reproduction semble venir appuyer cette hypothèse.

En effet, on peut constater que la reproduction asexuée obligatoire est un phénomène rare chez les animaux car on estime que seul 0.1% à 1% de ces organismes se reproduisent ainsi (de Meeûs et al., 2007; Galis and Alphen, 2020; Rice, 2002) bien que la plupart des groupes comportent des lignées asexuées à l'exception des mammifères et des « oiseaux » (Rice, 2002). De plus, il est communément admis que les lignées asexuées dérivent toutes de lignées sexuées (Bengtsson, 2009; Rice, 2002). Enfin, ces lignées asexuées sont le plus souvent retrouvées dans les branches les plus jeunes de l'arbre de la vie, signe du fait qu'elles ne persistent pas sur des temps géologiques. En effet, aucun genre de taille importante, ni aucun groupe taxonomique supérieur, n'est entièrement composé de lignées asexuées (Rice, 2002).

Selon Bengtsson, le sort le plus probable d'une lignée asexuée nouvellement formée est une extinction rapide, même si elle est initialement favorisée de manière sélective. Cela s'applique à tous les nouveaux traits dans les processus d'évolution stochastique. Ce n'est que si la lignée se développe rapidement en nombre et peut se créer une niche appropriée qu'il est probable qu'elle continuera à subsister pendant un temps d'évolution conséquent. Une lignée asexuée peut même en venir à dominer un vaste espace géographique mais de telles opportunités peuvent cependant tout aussi bien se refermer à cause, par exemple, de changements environnementaux. De même, un organisme asexué bien établi peut après un certain temps, être frappé par un parasite particulièrement adapté à son génotype spécifique (Bengtsson, 2009). En effet, la théorie du « cliquet pathogène » prédit un avantage au sexe lorsque la recombinaison réduit la similarité des facteurs de résistance codés génétiquement entre les parents et la progéniture qui sont regroupés dans l'espace, et réduit ainsi la transmission des pathogènes entre les parents et la progéniture (Rice, 2002). La reproduction sexuée et la création de nouvelles combinaisons de gènes qui l'accompagne sont donc les moteurs d'une interminable course à l'armement (hypothèse de la Reine Rouge) rendue nécessaire dans un monde en constante évolution.

L’ensemble de ces observations ont contribué à créer le dogme selon lequel on considère que se reproduire de manière strictement asexuée constitue une impasse évolutive à long terme, même si quelques contre-exemples existent comme les rotifères bdelloïdes évoqués plus haut qui ont réussi à persister pendant des millions d'années. Sur les mêmes bases d’observation, Bengtsson conclut que la recombinaison génétique est un trait favorisé par l'évolution et que son absence causée par la reproduction strictement asexuée est donc un handicap (Bengtsson, 2009).

Pour pondérer cette conclusion, on peut tout de même constater que la reproduction sexuée est incapable de maintenir les meilleures combinaisons d’allèles ensemble au cours du temps car elles sont brassées à la génération suivante (de Meeûs et al., 2007). Par ailleurs, la reproduction asexuée permet de réduire i) le double "coût de production des mâles", qui fait référence à la réduction du taux de croissance intrinsèque d'une population sexuelle lorsque les mâles ne fournissent pas de ressources qui augmentent la fécondité de leurs compagnes; ii) le double "coût de la méiose", qui réduit le rapport parent-production de 1, chez une femelle qui se reproduit parthogénétiquement, à 0,5 chez une femelle qui se reproduit sexuellement ; et iii) la rupture des combinaisons de gènes co-adaptés, conséquence du brassage provoqué par la reproduction sexuée à chaque génération (Rice, 2002). Enfin, si lors de l’apparition de l’asexualité chez l’organisme son génotype de départ était particulièrement adapté et diversifié, il faudra parfois des périodes très longues et de nombreuses générations avant que les aspects délétères de l’absence de recombinaison ne se fassent ressentir. Dans ces cas-là, l’absence de recombinaison peut donc au contraire être un facteur clef de propagation des organismes clonaux (Bengtsson, 2009). C’est par exemple le cas chez certains végétaux issus d’une hybridation chez lesquels on observe une « vigueur hybride » (performances supérieures aux « parents » sexués), qui peut être conservée au cours du temps grâce à l’asexualité (Ozias-Akins and Conner, 2020). Néanmoins, ce phénomène n’est pas automatique puisqu’à l’inverse il a été montré au sein de poissons du genre *Phoxinus* que des hybrides arboraient des performances de nage inférieures à celle des donneurs sexués (Mee et al., 2011).

Nous allons par la suite nous intéresser particulièrement à un groupe d’organismes présentant une gamme de modes de reproduction diversifiée et pour lesquels des espèces à reproduction strictement asexuée présentent un succès évolutif apparent: les nématodes du genre *Meloidogyne*.

B- Les nématodes du genre *Meloidogyne* : organismes d'intérêts agronomique et évolutif

1 - Intérêt agronomique et mode d'action parasitaire

Avec 11 % de la population souffrant de sous-nutrition en 2018 et une population estimée à 9,7 milliards d'individus à l'horizon 2050, la sécurité alimentaire mondiale est chaque jour plus menacée (un.org, fao.org). Pour subvenir aux besoins de cette population, la production agricole devra croître de 20 à 70% et ceci sans possibilité d'étendre la surface cultivée dans les mêmes proportions (Hunter et al., 2017). Dans ce contexte, limiter les pertes agricoles quelle que soit leur cause est donc primordial.

À eux seuls, les nématodes parasites de plantes causent plus de 80 milliards de dollars (US) de dommages par an à la production agricole mondiale et il est communément admis que les espèces les plus néfastes sont celles appartenant au genre *Meloidodgyne* (Jones et al., 2013). Les *Meloidogyne* sont des endoparasites obligatoires de la racine dans laquelle ils vont établir leur site nourricier et passer la majeure partie de leur cycle de vie. L'infection d'une plante par l'un de ces nématodes se traduit par l'apparition de boursouflures symptomatiques appelées « galles » sur la racine (voir Figure 1.B.1-b).

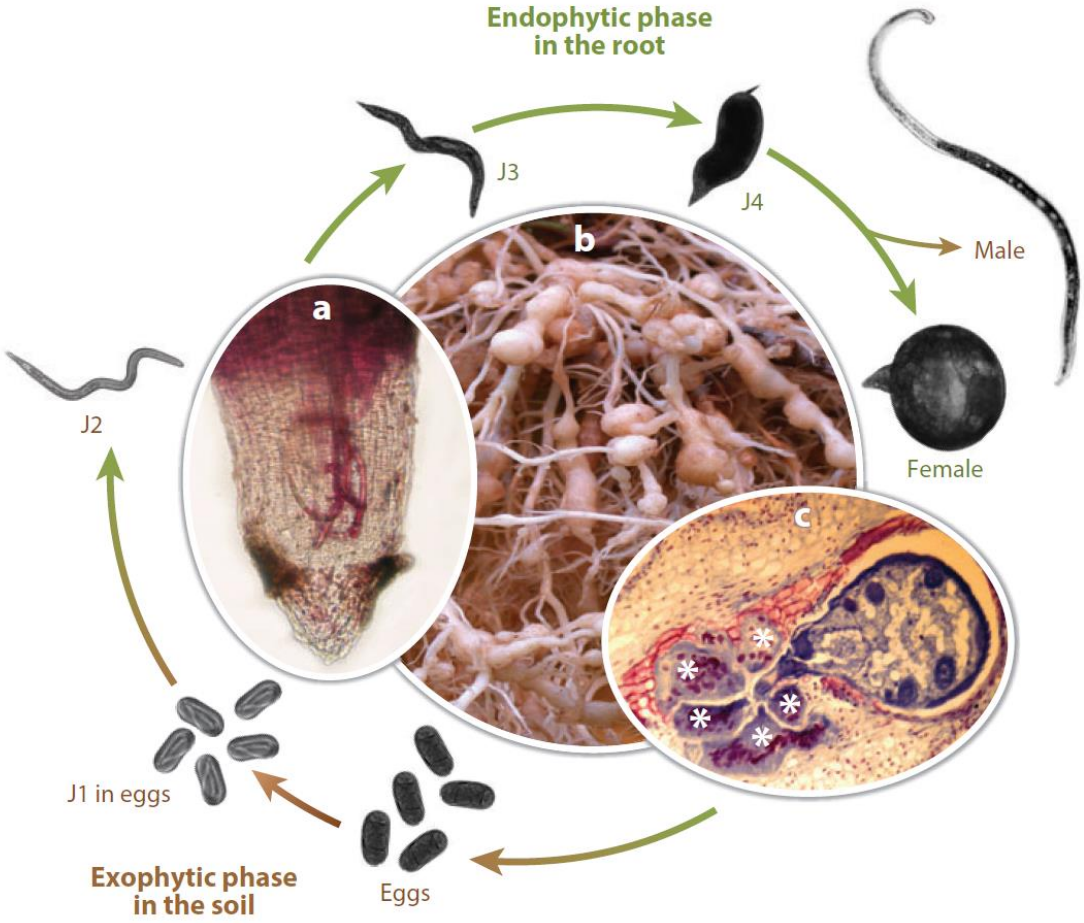


Figure 1.B.1: cycle de vie du nématode *M. incognita* .

Figure issue de (Castagnone-Sereno et al., 2013) (Figure 1)

A - Coupe longitudinale de l'extrémité d'une racine montrant les juvéniles de deuxième stade (J2) (colorés à la fuchsine acide) se retournant au niveau du méristème racinaire pour migrer dans le cylindre vasculaire.

B - Symptômes typiques (i.e. galles) sur les racines de tomates.

C - Coupe longitudinale d'une racine infestée montrant une femelle mature et cinq cellules géantes (∗) constituant le site nourricier du nématode.

Le cycle de vie des *Meloidogyne* dure 3 à 6 semaines en fonction des espèces et des conditions environnementales (Castagnone-Sereno et al., 2013). Ce cycle de vie, commun à l'ensemble des espèces de *Meloidogyne*, se déroule en 4 phases principales (voir Figure 1.B.1). La femelle mature pond ses œufs dans une matrice protectrice formant une masse d'œufs à la surface de la racine ou bien incrustée dans la galle ; chaque masse d'œuf contenant un nombre variable d'œufs de l'ordre du millier d'unités. Après embryogenèse, les juvéniles au premier stade larvaire (J1) muent à l'intérieur de l'œuf puis éclosent sous forme de larves juvéniles de second stade (J2) ayant acquis leurs caractères infectieux. La

larve J2, mobile, pénètre dans la racine puis migre jusqu'à la zone de différenciation où elle induit la formation de 5 à 7 cellules géantes hyperactives métaboliquement qui constitueront son site nourricier. Une fois nourrie, la larve J2 gonfle et subit encore 3 mues successives avant d'atteindre le stade d'adulte reproducteur (Jones et al., 2013). On notera que si des mâles sont bien produits, leur proportion dans la population est anecdotique. Par ailleurs, ils ne contribuent pas au patrimoine génétique de la descendance (Triantaphyllou, 1981).

Bien que généralement non létale, ce détournement des ressources de la plantes a des effets néfastes notables sur sa croissance, sa résistance aux aléas climatiques, et par conséquent sur son rendement.

2 - Intérêt en génomique de l'évolution : un groupe d'espèces diversifié

Le genre *Meloidogyne* constitue un large groupe d'espèces avec 98 espèces décrites en 2013 (Jones et al., 2013) (voir Figure 1.B.2). Les *Meloidogyne* sont répartis sur l'ensemble du globe. Néanmoins, comme illustré en Figure 1.B.2, leur aire de répartition géographique varie grandement d'une espèce à l'autre. Certaines espèces comme *M. incognita, M. javanica, M. arenaria* ou encore *M.hapla* sont distribuées mondialement alors que d'autres comme *M. dunensis, M.indica* ou encore *M. nataliei* n'ont été décrites que dans certaines aires géographiques restreintes. On trouve donc des *Meloidogyne* dans des conditions de vies très différentes et parfois extrêmes. Par exemple, *Meloidogyne incognita*, l'espèce de *Meloidogyne* la plus répandue, est retrouvée des régions tempérées aux régions tropicales, dans tout endroit où la température annuelle la plus basse est supérieure à 3°C (Sasser et al, 1983).

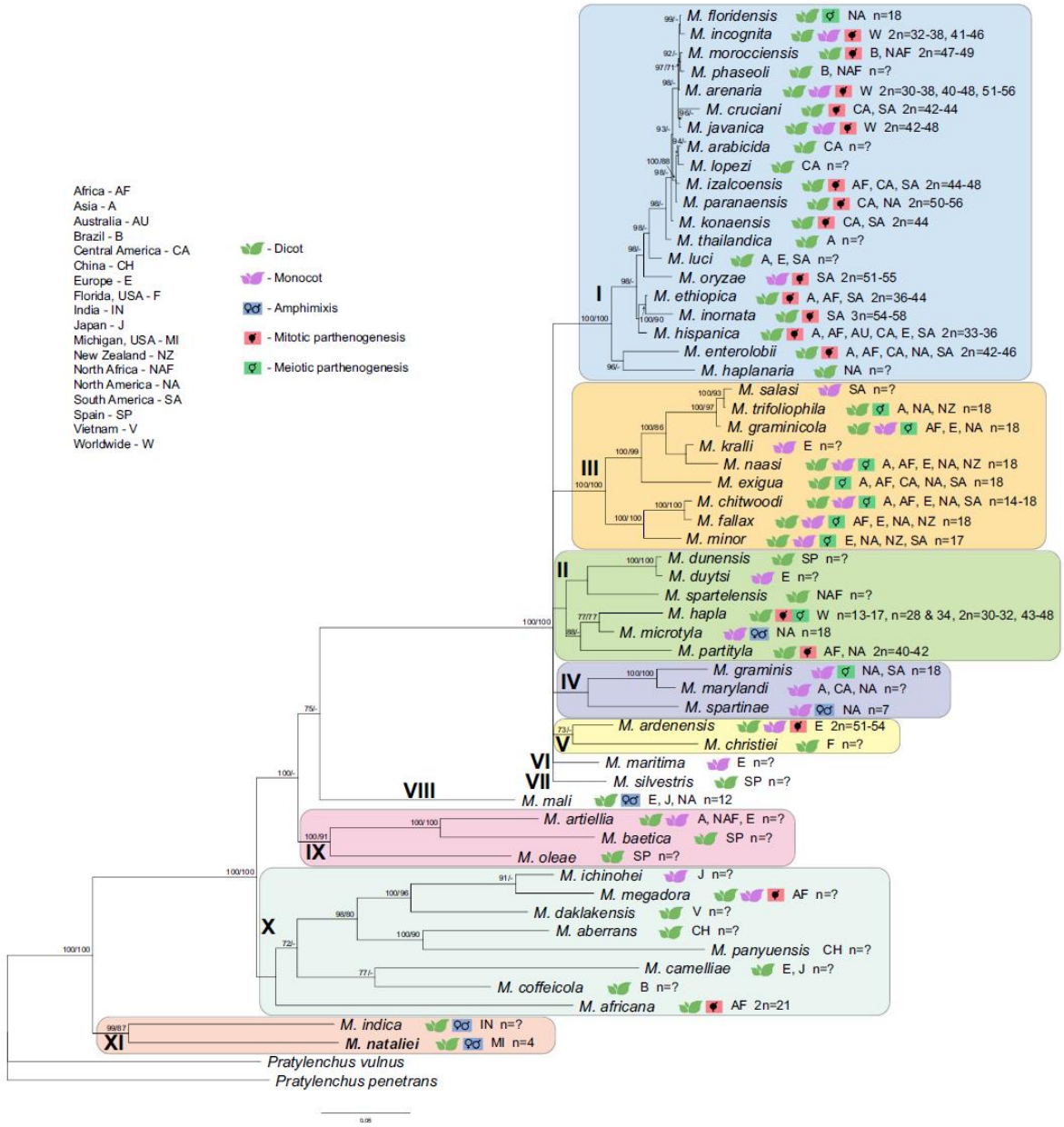


Figure 1.B.2: arbre phylogénétique des espèces du genre Meloidogyne.

Figure issue de (Álvarez-Ortega et al., 2019) (Figure 4)

Arbre consensus majoritaire (50%) obtenu par inférence bayésienne (modèle GTR+I+G) à partir des alignements multiples des séquences de l'ARNr 18S, de l'ARNr ITS1, des segments d'expansion D2-D3 de l'ARNr 28S, du gène COI et l'ARNr COII-16S. La topologie n'est pas résolue pour les noeuds dont la valeur de support est inférieure à 70%. Les valeurs de support sont indiquées comme suit : valeur des probabilités postérieures dans l'analyse d'inférence bayésienne/valeur bootstrap de l'analyse de maximum de vraisemblance. n représente le nombre de chromosomes de l'espèce (n = ? -information sur le nombre de chromosomes inconnue).

Les *Meloidogyne* ont une gamme d'hôtes (nombre d'espèces de plantes qu'un parasite est capable d'infecter) extrêmement large qui englobe la plupart des plantes à fleurs cultivées que ce soit pour l'alimentation humaine, du bétail ou l'ornement (Trudgill, 1997). On constatera néanmoins que là aussi, cette gamme d'hôtes varie grandement d'une espèce à l'autre, voire même au sein d'une espèce (« race d'hôte »). *M. incognita, M. javanica* et *M. arenaria* sont par exemple capables d'infecter l'ensemble des plantes d'intérêt agricole et plus encore. Selon ces deux critères (aire de répartition géographique et gamme d'hôtes), il est communément admis que les espèces les plus nuisibles sont *M. incognita, M. javanica, M. arenaria* (espèces tropicales), et *M. hapla* (espèce tempérée) auxquelles on peut rajouter 2 espèces émergentes: *M. enterolobii* et *M. chitwoodi* (Castagnone-Sereno et al., 2013; Jones et al., 2013).

Les *Meloidogyne* présentent aussi une grande diversité cytogénétique tant en termes de niveau de ploïdie, que du nombre de chromosomes. Cette observation est bien sûr vraie entre espèces différentes mais aussi au sein d'une même espèce. Par exemple l'espèce *M. hapla* présente différentes populations pour lesquelles le nombre de chromosomes (26 à 68) et la ploïdie (2 à 4-5 n) varient. Les espèces apomictiques sont majoritairement tri ou tétra-ploïdes, et la plupart présentent de 40 à 50 chromosomes en moyenne (Trudgill and Blok, 2001).

Enfin, l'aspect le plus intéressant de leur diversité concerne sans doute leurs modes de reproduction. Le genre *Meloidogyne* présente un large éventail de modes de reproduction (Castagnone-Sereno et al., 2013) qui s'étend de l'amphimixie stricte (reproduction sexuée obligatoire) à la parthénogenèse mitotique obligatoire (ou reproduction asexuée stricte) en passant par la parthénogenèse méiotique facultative.

Le genre *Meloidogyne* constitue donc un groupe d'espèce avec des traits d'histoire de vie diversifiés (répartition géographique, mode de reproduction ou gamme d'hôte) représentant donc un modèle d'étude de choix pour réaliser des analyses de génomique comparative en relation avec des variations phénotypiques.

3 - Une énigme évolutive.

De manière surprenante, il a pu être constaté qu'au sein des *Meloidogyne*, les espèces présentant la plus grande gamme d'hôtes et causant le plus de dégâts sont des espèces à reproduction asexuée stricte (apomictiques). Sur les 54 espèces examinées par Jepson (Jepson, 1987), il a été constaté que l'ensemble des espèces amphimictiques n'infectent en général qu'une plante ou un groupe restreint de plantes alors que les espèces parthénogénétiques sont polyphages. Par exemple, l'espèce sexuée *M. subartica* est restreinte aux *Commelinidae* (sous-groupe de plantes monocotylédones) alors que les espèces

apomictiques (parthénogénétiques mitotiques) *M. incognita, M. javanica,* et *M. arenaria* présentent une gamme d'hôte virtuellement illimitée puisque capables d'infecter presque l'ensemble des familles de plantes (Trudgill and Blok, 2001). A cet égard, leur distribution réelle va bien au-delà des niches écologiques restreintes et hautement spécialisées supposées être habitées par des organismes asexués dans le cadre de la parthénogenèse géographique (Castagnone-Sereno, 2006).

Par ailleurs, il a pu être constaté que des espèces strictement asexuées de *Meloidogyne* sont capables de contourner les défenses de la plante hôte et ce en un nombre de générations restreint. Via des expériences contrôlées en laboratoire, il a par exemple été démontré des populations avirulentes de *M. incognita* (i.e contrôlées par un gène de résistance chez la plante et donc majoritairement incapables d'infecter ladite plante) étaient capables de surmonter la résistance de la plante en quelques générations, conduisant à des sous-populations virulentes (Castagnone-Sereno et al., 1994; Castagnone-Sereno et al., 2019; Tzortzakakis et al., 2014). Le même constat d'émergence de populations virulentes qui ne sont plus contrôlées par des gènes de résistance, a également été signalé sur le terrain (Barbary et al., 2015).

La question des mécanismes évolutifs permettant à ces organismes asexués de s'adapter rapidement à un environnement changeant relatif au mode de vie parasitaire se pose donc, et va être abordée dans la section suivante.

C- Facteurs de plasticité génomique en absence de recombinaison

1 - Hybridation, polyploïdie et structure du génome

En l'absence de recombinaison sexuelle et méiotique (*e.g.* en cas de reproduction asexuée), la polyploïdie (fait de posséder plus de deux ensembles de chromosomes homologues) et l'hybridation (processus consistant à combiner différentes variétés d'organismes pour créer un hybride) peuvent être d'importantes sources de diversité génomique et donc potentiellement d'adaptabilité.

Tout d'abord, la polyploïdie peut fournir la matière première pour la néo- et la sous-fonctionnalisation de copies de gènes dupliquées, ce qui entraîne une nouvelle variation génétique (Cuypers and Hogeweg, 2014; Soltis et al., 2014). Il a par exemple été démontré chez la levure que le niveau de ploïdie est corrélé à une adaptation plus rapide (Selmecki et al., 2015). De plus, il a été suggéré que la polyploïdie pourrait masquer les allèles récessifs délétères (Madlung, 2013) et limiter leur accumulation via la conversion de gènes entre régions homologues, la copie fraîchement mutée étant corrigée par la présence des nombreuses autres copies « sauvages » (e.g. non-mutées, « Wild type »). Ainsi selon Maciver, la polyploïdie permettrait à des espèces asexuées, en l'occurrence des amibes, d'échapper au « cliquet de Muller » (Maciver, 2016)

Un génome peut devenir polyploïde suite à i) la duplication complète du génome à la suite de la non-disjonction des chromosomes dans la lignée germinale au cours de la méiose (autopolyploidie), ou ii) l'hybridation de deux espèces étroitement apparentées (allopolyploïdie). L'allopolyploïdie, en combinant plusieurs génomes dans une espèce, peut conduire à des phénotypes transgressifs (caractéristiques phénotypiques différentes des donneurs) qui surpassent ceux de l'espèce parente par le biais d'une nouvelle combinaison génétique (Bar-Zvi et al., 2017; Dittrich-Reed and Fitzpatrick, 2013; Madlung, 2013). Ce phénomène, connu sous le nom d'hétérosis, décrit l'augmentation des capacités et/ou de la « vigueur » d'un hybride par rapport à ses parents (basé sur des critères de tailles, de poids, de voracité, etc). Certaines salamandres du genre *Ambystoma* constituent un cas clair de phénotype transgressif chez les animaux. En effet, il a pu être montré que les hybrides entre une espèce indigène et une espèce introduite sont écologiquement plus adaptés et plus performants que l'espèce indigène parentale ainsi que d'autres espèces apparentées dans le milieu naturel (Fitzpatrick and Shaffer, 2007). Ces hybrides menacent d'ailleurs la survie de l'espèce indigène.

M. incognita, M. javanica et M. arenaria sont trois espèces parthénogénétiques mitotiques pour lesquelles la perte de reproduction sexuée est probablement due à un nombre indéterminé d'évènements d'hybridation interspécifique ayant mené à l'apparition de leur génome polyploïde (Blanc-Mathieu et

al., 2017). Au sein de chacun de ces génomes, il a été constaté un très fort taux de divergence (6 à 8%) entre régions homéologues (régions homologues ayant divergées) impliquant la présence de *n* génomes en 1 au sein de ces espèces (avec *n* = 3 à 4-5 suivant les espèces). A titre d'exemple, ce taux de divergence entre régions homéologues est supérieur à celui observé entre l'Humain et le macaque (Rogers and Gibbs, 2014) dont on date la séparation entre 25 et 28 millions d'années (même si ces chiffres ne sont pas tout à fait comparables, en raison de la taille différente des génomes et de traits d'histoires de vies distincts). Par ailleurs, le séquençage du génome complet de *M. hapla*, une espèce à reproduction parthénogénétique facultative, n'a révélé aucune structure génomique comparable à celle observée chez les 3 espèces parthénogénétiques strictes précédemment citées (Opperman et al., 2008). Ceci sous-entend que cette structure génomique avec des régions homéologues extrêmement divergentes est propre aux espèces asexuées de *Meloidogyne*. Chez ces espèces, les copies de gènes résultant de l'allopolyploïdie divergent non seulement au niveau de leur séquence mais aussi dans leurs patrons d'expression, ce qui suggère que cette structure particulière du génome pourrait promouvoir une diversité de fonctions, qui pourrait à son tour être impliquée dans leur succès parasitaire malgré l'absence de reproduction sexuée (Blanc-Mathieu et al. 2017). Cette hypothèse semble cohérente avec le concept de "génotype généraliste" ("general purpose genotype"), qui propose que les organismes asexués ayant réussi à se maintenir aient un génotype généraliste qui leur confère une capacité d'adaptation importante à une variété d'environnements différents (Vrijenhoek and Parker, 2009). Une autre hypothèse, non mutuellement exclusive, est le concept de "variations de la niche gelée" ("frozen niche variation"), qui propose que les organismes asexués réussissent à se maintenir dans des environnements stables parce qu'ils ont un génotype gelé adapté à cet environnement spécifique (Vrijenhoek et Parker 2009). Cependant, bien qu'un "génotype à usage général" apporté par l'hybridation puisse contribuer à la large gamme d'hôtes et à la répartition géographique des espèces parthénogénétiques de *Meloidogyne* précédemment citées, cela ne peut expliquer à lui seul comment ces espèces évoluent et s'adaptent à de nouveaux hôtes ou environnements en absence de recombinaison.

2 - Variations structurales

La variation de la structure des chromosomes, ou variation structurale, englobe un large éventail de mutations, y compris les insertions, les inversions, les translocations, ou encore des variations du nombre de copies de régions entières via des mécanismes de duplication ou de délétion. En raison de la variété des types d'altérations possibles, de leur ubiquité et de leur longueur variable, les variations structurales ont un fort potentiel pour induire la réorganisation des génomes et ainsi produire de nouveaux phénotypes adaptatifs (Radke and Lee, 2015). Dennis et Eichler montrent par exemple que les humains et les grands singes présentent plus de différences génétiques en termes de contenu et de

structure au sein de duplications segmentaires récentes que dans toute autre région codante. Ils décrivent en outre la découverte de nouveaux gènes spécifiques à l'homme (ARHGAP11B et SRGAP2C) issus d'un événement de duplication qui auraient eu un rôle potentiel dans l'expansion néocorticale et l'augmentation de la densité neuronale de la colonne vertébrale (Dennis and Eichler, 2016).

Parmi les variations structurelles, les variations du nombre de copies (VNC) de régions codantes ou non codantes constituent des sources importantes de polymorphismes génétiques contribuant à la diversité phénotypique des populations. Les VNC sont définies comme des séquences de taille variable (de 50 pb jusqu'au chromosome entier) qui, en raison de duplications et/ou de délétions, varient dans leur nombre de copies entre les individus d'une population (Steenwyk and Rokas, 2018) et sont dispersées dans l'ensemble du génome. Les VNC peuvent influencer de manière significative la diversité phénotypique au sein d'une population. Les VNC couvrant des gènes peuvent constituer une plate-forme majeure pour la divergence fonctionnelle des duplications de gènes (par exemple, par la sous-fonctionnalisation ou le partitionnement d'un ensemble de fonctions ancestrales entre les duplications), y compris l'évolution de nouvelles fonctions (néo fonctionnalisation) (Steenwyk and Rokas, 2018). Par exemple chez les serpents, il a été montré que des évènements de duplication de gènes codant pour des phospholipases suivis de néo-fonctionnalisations seraient responsable de la diversité des venins rencontrés chez ces espèces et donc potentiellement de la diversité de ces espèces (Lynch, 2007). Les VNC peuvent être héritées de la génération précédente ou apparaître de novo par des événements de duplication/délétion, et leur fixation par dérive ou sélection peut contribuer à la création d'une nouveauté génétique entraînant l'adaptation des espèces à des environnements stressants ou nouveaux (Castagnone-Sereno et al., 2019). Par exemple, il a été démontré que les VNC peuvent conduire à des phénotypes adaptatifs tels que la résistance au cuivre chez la levure (Hull et al., 2017) ou encore la résistance à des médicaments antifongiques chez *Candida albicans,* un champignon pathogène de l'humain (Selmecki et al., 2010).

Chez *M. incognita,* Castagnone-Sereno et collaborateurs ont détecté des variations du nombre de copies de gènes (majoritairement des délétions) qu'ils ont pu associer à la capacité des nématodes à surmonter la résistance de la plante hôte. Par ailleurs, ils ont pu observer que certaines de ces variations du nombre de copies de gènes étaient convergentes entre deux lignées initialement avirulentes (i.e. incapable de parasiter l'hôte) de *M. incognita* d'origines géographiques distinctes. Leurs résultats soutiennent donc l'idée que les variations du nombre de copies de gènes (en particulier la perte de copie) pourraient constituer des mécanismes génétiques adaptatifs communs en réponse à des variations des conditions du milieu chez les animaux clonaux (Castagnone-Sereno et al. 2019).

Néanmoins, les mécanismes responsables des variations du nombre de copies de gènes et d'autres variations structurales éventuellement impliquées dans l'évolution adaptative de *M. incognita* restent à décrire.

3 - Mutations ponctuelles

Les mutations nucléotidiques ponctuelles sont des altérations de la séquence d'ADN intervenant de manière aléatoire dans les génomes via des mécanismes d'insertion, de délétion ou de substitution d'un seul nucléotide à la fois. Les mutations ponctuelles sont ubiquitaires dans le vivant, bien que tous les génomes ne mutent pas à la même « vitesse ».

Si une mutation intervient dans une région non codante et non régulatrice, elle n'aura la plupart du temps aucune conséquence fonctionnelle ou régulatoire immédiate et passera donc le plus souvent inaperçue (mutation silencieuse). En revanche, si cette mutation intervient dans la région codante ou régulatrice d'un gène, elle pourra avoir des conséquences bien plus importantes. Par exemple, certaines substitutions peuvent être non-synonymes et donc modifier la séquence et la fonction de la protéine codée. Les insertions/délétions de nucléotides peuvent engendrer un décalage de phase et causer des codons stop prématurés donnant des protéines tronquées. Comme le résument Deng et collaborateurs, un grand nombre de gènes associés à divers types de cancer contiennent des mutations ponctuelles (Deng et al., 2017). Dans les régions régulatrices, les mutations peuvent modifier l'affinité de reconnaissance par des éléments régulateurs et donc modifier le patron d'expression des gènes. Chez la bactérie *Pseudomonas fluorescens*, il a par exemple été montré qu'une seule mutation ponctuelle (substitution) à eu pour effet de « recâbler » un réseau génique entier (Knight et al., 2006).

De simples modifications mono-nucléotidiques peuvent donc parfois avoir des répercussions importantes sur les organismes dont des modifications phénotypiques pouvant potentiellement être transmises à la descendance, créant ainsi de la variabilité dans la population. Il a par exemple été montré qu'un certain nombre de phénotypes chez les mammifères, notamment l'épaisseur des poils, la masse musculaire ou encore la locomotion, peuvent directement être reliées à des polymorphismes mono-nucléotidiques (PMN) (Radke and Lee, 2015).

Chez *M. incognita*, de récentes analyses de génomique des populations comparant différents isolats brésiliens présentant des gammes distinctes de compatibilité avec l'hôte ont montré que seulement ~0.19% des positions sur l'ensemble du génome de référence présentaient un polymorphisme (Koutsovoulos et al., 2020). L'ajout d'autres isolats provenant de différentes régions géographiques du monde n'a pas eu d'effet significatif sur l'augmentation du nombre de positions variables dans le génome. Le peu de PMN identifiés n'ont montré aucune corrélation significative avec l'emplacement géographique, la gamme d'hôtes ou les espèces de cultures actuellement infectées. Néanmoins, l'utilisation de ces PMN a permis de confirmer l'absence de recombinaison méiotique sexuelle chez *M. incognita* (Koutsovoulos et al. 2020). Ainsi, la faible variabilité nucléotidique qui a été observée entre

les isolats n'est probablement pas le principal moteur de la plasticité génomique qui sous-tend l'adaptabilité de *M. incognita*.

Ce résultat n'est pas surprenant car i) l'apparition d'une mutation ponctuelle est majoritairement aléatoire, et ii) bien que la mutation soit instantanée, sa propagation dans une population est quant à elle un processus lent, en particulier chez les espèces asexuées (cf chapitre I). Pour avoir un impact significatif dans des délais d'évolution courts (par exemple, la capacité à surmonter la résistance des plantes en quelques générations), le taux de mutation doit nécessairement être élevé. Cependant, comme l'accumulation de mutations délétères n'est pas purgée par recombinaison chez les organismes asexués, un taux de mutation élevé pourrait dangereusement conduire à une saturation de mutations délétères dans le génome (Muller, 1964). Étant donné l'impact négatif de ce mécanisme sur les espèces asexuées, on peut considérer que l'accumulation de mutations ponctuelles dans les génomes de ces espèces n'est probablement pas le principal facteur de plasticité génétique. En outre, il n'existe actuellement aucune donnée indiquant une différence dans les taux de mutation entre des espèces sexuées et asexuées de *Meloidogyne*.

4 - Transferts horizontaux

Les transfert horizontaux (TH) sont des mouvements de matériel génétique entre des organismes autrement que par la transmission ("verticale") de l'ADN du parent à la progéniture (*e.g.* reproduction) (Keeling and Palmer, 2008).

La quantité de TH entre organismes peut représenter un apport non négligeable de matériel génétique. Par exemple, Il a été montré que d'importantes quantités d'ADN ont été transférées depuis le génome de la bactérie endosymbiotique *Wolbachia* vers le génome nucléaire de plusieurs insectes et nématodes hôtes (Hotopp et al., 2007; Nikoh et al., 2008). Le cas le plus extrême à ce jour concerne la mouche *Drosophila ananassae* chez laquelle il a été montré que plus de 2% de son génome (~5 Mbp) est issu de l'insertion de plusieurs copies du génome de *Wolbachia* (Klasson et al., 2014). Cependant, à ce jour, ces insertions de matériel génétique de *Wolbachia* dans les génomes d'arthropodes et nématodes ne semblent pas associées à une fonction ou à un caractère phénotypique. A ce jour, la seule exception est l'insertion d'un élément féminisant à partir de *Wolbachia* dans le génome de l'isopode, un crustacé terrestre (Cordaux and Gilbert, 2017).

Les Transferts Horizontaux de Genes (THG) ont un rôle reconnu dans la plasticité des génomes et l'évolution des organismes (Keeling and Palmer, 2008). Chez le rotifère bdelloïde, un organisme évoluant sans sexe depuis plus de soixante millions d'années, environ 8 % des gènes codant pour les protéines sont issus de THG. Il a été fait l'hypothèse que ce taux élevé de THG constituerait un

mécanisme favorisant la diversité génétique en l'absence de sexe et pourrait donc être impliqué dans la longévité exceptionnelle de cette espèce (Eyres et al., 2015; Gladyshev et al., 2008).

Chez les nématodes et les insectes, les THG sont un mécanisme particulièrement répandu ; probablement à cause de l'association fréquente de ces organismes avec des endosymbiontes bactériens. Certains cas de THG pourraient être impliqués dans l'adaptation des arthropodes ou des nématodes aux plantes, soit parce que ces gènes codent des enzymes spécifiques permettant la dégradation et le métabolisme des produits végétaux, soit parce qu'ils permettent la détoxification de composants végétaux potentiellement nocifs (Drezen et al., 2017; Wybouw et al., 2016). Par exemple, il a été montré que le génome du scolyte du caféier (*Hypothemus hampei*, un coléoptère) renferme un gène codant pour une hydrolase qui a vraisemblablement été acquis via un TH. Cette hydrolase lui permet de dégrader certains polysaccharides de la baie de café dans laquelle il peut ensuite rentrer et établir son site nourricier (Acuna et al., 2012). En ce qui concerne les nématodes, il a été constaté qu'une part non négligeable des génomes des espèces du genre *Meloidogyne* (>3% des gènes codant pour les protéines) provient de transferts horizontaux (Paganini et al., 2012). Certains de ces THG codent pour des protéines de dégradation de la paroi cellulaire des plantes et il a donc été fait l'hypothèse que ces THG ont favorisé la capacité de ces nématodes à parasiter des plantes (Danchin et al., 2010). Cependant, le nombre de THG observé ne diffère pas de manière significative entre les nématodes à reproduction sexuée et asexuée. L'hypothèse actuelle est donc que plutôt que de constituer un mécanisme d'évolution en l'absence de sexe, ces THG auraient joué un rôle important dans l'évolution du phyto-parasitisme chez les nématodes dans leur ensemble (*i.e* aussi chez d'autres genres de nématodes phytoparasites) (Haegeman et al., 2011).

Pris séparément, aucun des mécanismes précédemment présentés ne peut complètement générer la plasticité génétique nécessaire à l'adaptabilité a priori paradoxale des espèces à reproduction asexuée de *Meloidogyne*, surtout en un laps de temps restreint.

Un autre facteur de plasticité et de diversité génétique qui n'a pas encore été étudié chez ces espèces pourrait à la fois jouer un rôle propre mais aussi permettre de combiner l'action de plusieurs des facteurs précédemment cités. Il s'agit des Eléments Transposables (ETs).

II – Les Eléments transposables

A- Définition, classification des ETs

Les Eléments Transposables (ETs) sont des séquences d'ADN capables de se déplacer et de se multiplier dans les génomes. Longtemps considérés comme des « gènes égoïstes » ou encore comme de l'ADN « poubelle », les ETs sont dorénavant vus comme des acteurs de premier plan de la dynamique des génomes (comme nous le verrons plus loin dans ce chapitre). Barbara McClintock fut la première à mettre en évidence le concept de transposition (i.e mouvement de séquence d'ADN) et ses impacts dans le génome du maïs (McClintock, 1953). D'abord largement ignorés, ses travaux et leurs implications furent finalement reconnus à partir des années 60-70 et lui valurent d'être la première femme à recevoir le prix Nobel de Physiologie et Médecine en 1983. Mieux vaut tard que jamais ?

En 1989, Finnegan propose de diviser les ETs en deux grands groupes en fonction de leur mécanisme de transposition et de la nature de l'intermédiaire utilisé pour leur mobilisation (voir Figure 2.A.1): les rétrotransposons, dont la mobilisation nécessite un intermédiaire d'ARN, et les transposons à ADN (Finnegan, 1989). A l'heure actuelle ce critère constitue toujours la pierre angulaire de la caractérisation des ETs, néanmoins d'autres caractères discriminants ont dû être trouvés afin de rendre compte de la diversité toujours grandissante des éléments nouvellement découverts. Ainsi depuis 1989, de nombreux sous-groupes ont été créés en fonction de critères variés tels que le mécanisme d'intégration chromosomique de l'ET, la machinerie enzymatique utilisée, la présence de certains motifs, ou encore sur des critères d'homologie avec des séquences connues (Arkhipova, 2017; Piégu et al., 2015).

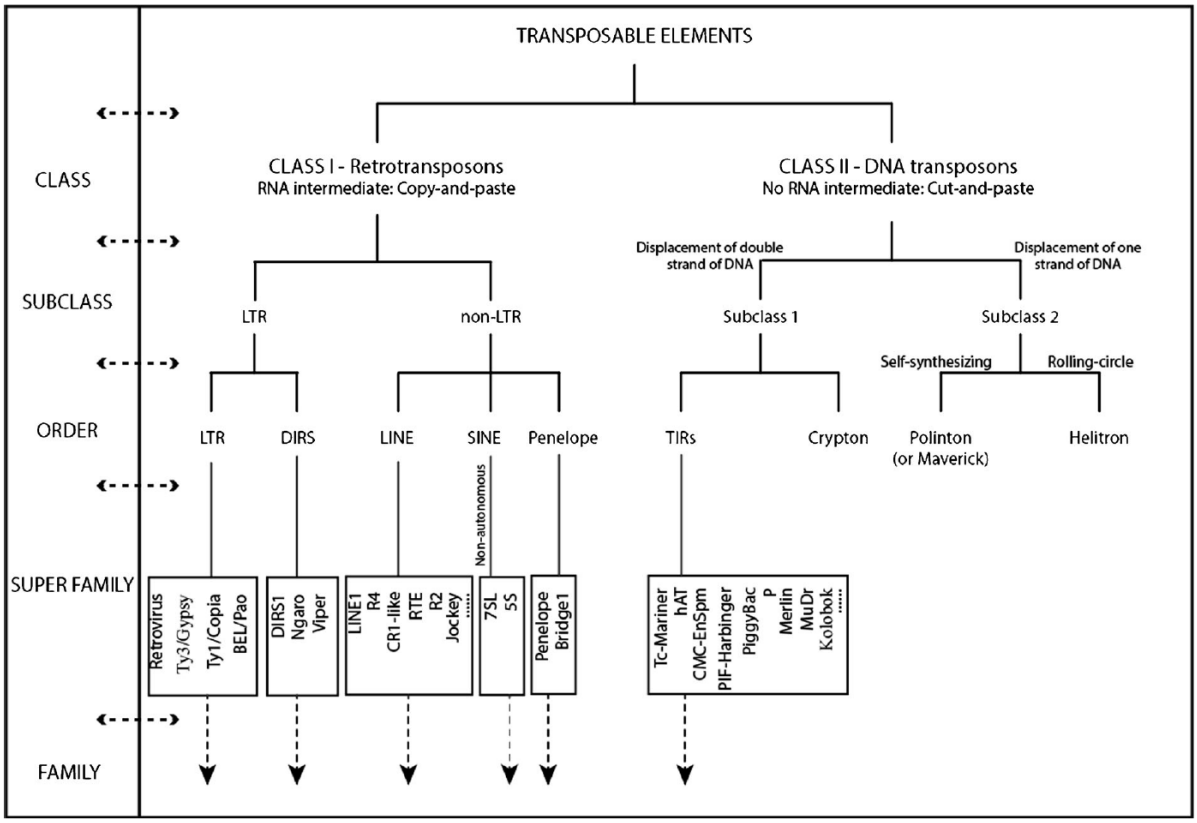


Figure 2.A.1 : Classification des ETs (classification de « Wicker »).

Figure issue de (Chalopin et al., 2015) (Figure 1)

Compte tenu de la combinatoire entraînée par l'analyse de ces caractéristiques, la classification des ETs (i.e. leur regroupement selon un ensemble de critères donné) fait débat et constitue un domaine en constante évolution (cf en fin de partie). Actuellement, deux systèmes de classification principaux cohabitent. Le premier que nous appellerons dès maintenant classification de « Wicker » est explicité en détail dans la publication (Wicker et al., 2007). Le deuxième que nous appellerons dès lors classification de « repbase » est présenté dans la publication (Kapitonov and Jurka, 2008). Bien que certains regroupements d'ETs varient entre les deux systèmes, ces différences sont principalement de l'ordre de variation de nomenclature (voir Figure 2.A.2) et sont somme toute minimes (Piégu et al., 2015).

En effet, les classifications Wicker et Repbase sont toutes deux basées sur les caractéristiques de l'ADN et de la séquence d'acides aminés des ETs. Les deux systèmes divisent les ETs en deux groupes : les rétrotransposons et les transposons à ADN. Cette division basale est appelée "type" dans la classification Repbase et "classe" dans celle de Wicker. Chacune de ces deux classes ou types est ensuite subdivisée en "classes" dans la classification Repbase ou en "ordres" dans celle de Wicker. Dans

l'ensemble, les "classes" de Repbase et les "ordres" de Wicker sont très similaires et chaque groupe contient les mêmes superfamilles d'ET (Piégu et al., 2015).

Dans les paragraphes suivants, je vais me servir de la classification de Wicker (que j'ai le plus utilisée durant ma thèse) pour rendre compte de la diversité des ETs rencontrés chez les eucaryotes et rapporter les caractéristiques principales de chaque classe (ou "groupe") et ordre (ou "sous-groupes") d'ETs. Cette synthèse sera majoritairement basée sur les revues et articles suivants (Bourque et al., 2018; Makałowski et al., 2019; Piégu et al., 2015; Wicker et al., 2007) qui traitent exhaustivement de ce sujet.

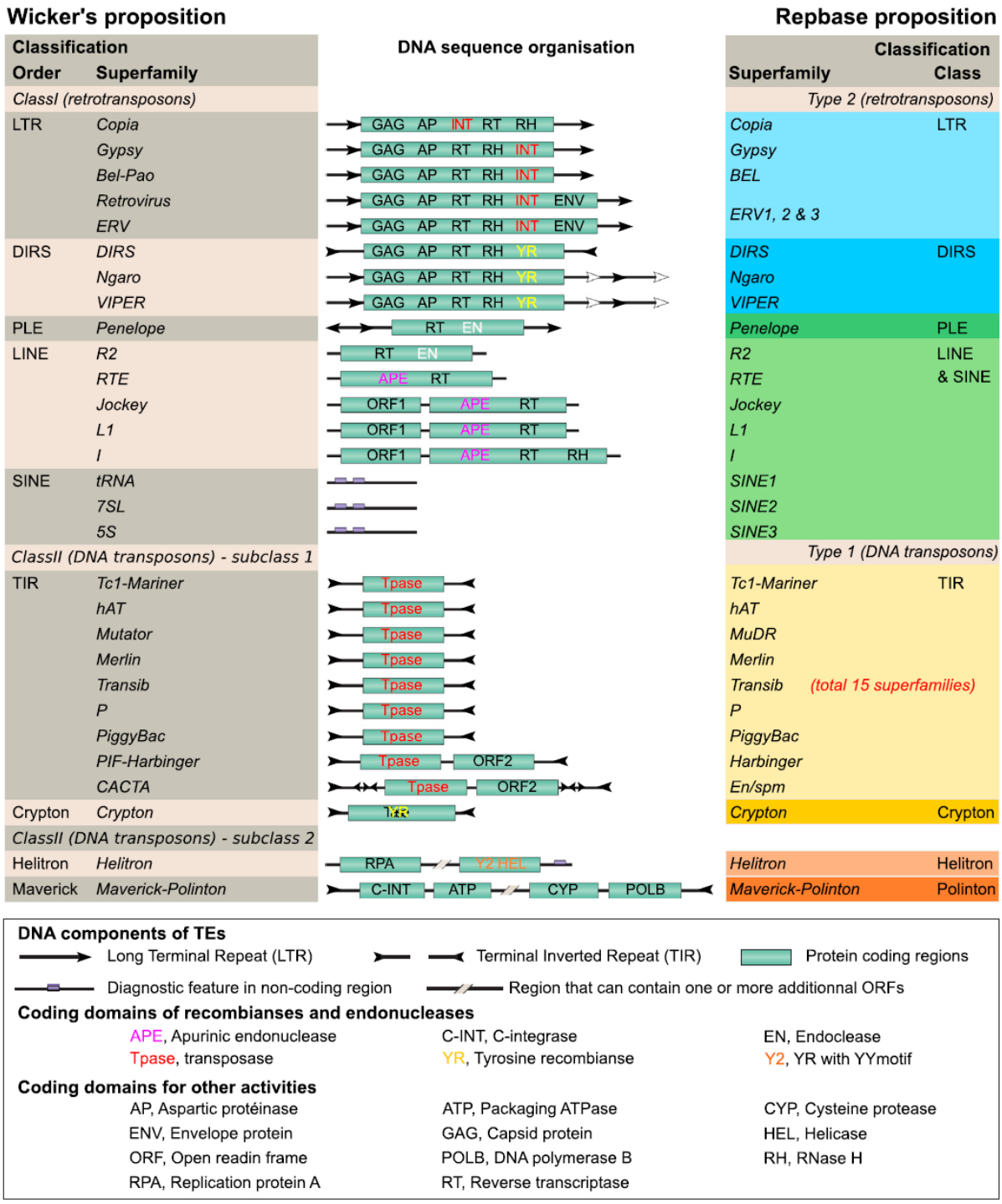


Figure 2.A.2 : Comparaison des classifications d'ET de « Wicker » et de « Repbase »

Figure issue de (Piégu et al., 2015) (Figure2).

Comme précédemment évoqué, les ETs sont généralement classifiés en deux groupes (ou Classes) en fonction de leur intermédiaire de mobilisation.

Les éléments de classe I (ou rétrotransposons) sont mobilisés via un intermédiaire à ARN. Un rétrotrotransposon présent à une position génomique donnée est i) transcrit en un intermédiaire à ARN, ii) l'intermédiaire à ARN subit une transcription inverse en ADNc, iii) l'ADNc est intégré ailleurs dans le génome. Les rétrotransposons sont donc mobilisés dans les génomes via un mécanisme de "copier-coller" qui induit la création d'une nouvelle copie lors de chaque cycle de réplication. Les rétrotransposons ont longtemps été séparés en deux grand groupes : les éléments LTR et les non-LTR qui se distinguent par la présence ou l'absence de longues régions terminales répétées (« Long Terminal Repeat » i.e. « LTR ») à leur extrémités, longues de quelques centaines de paires de bases (pb) à plusieurs kilobases (kb). Dans la classification de Wicker cette distinction n'existe pas. Les rétrotransposons (autonomes i.e. qui codent pour leur propre machinerie de transposition) sont répartis entre les LTRs (« Long Terminal Repeat transposon »), les DIRS (« DIRS-like »), les PLE (« Penelope-Like Element »), les LINE (« Long INterspersed Element »). La classification de Wicker décrit aussi 3 ordres de rétrotransposons « non-autonomes » : les SINE (« Short INterspersed Element »), les LARD (« LArge Retrotransposon Derivative ») et les TRIM (« Terminal Repeat retrotransposon In Miniature »). La discrimination entre ces 7 ordres d'ET se fait principalement sur des critères de machinerie enzymatique et donc conséquemment sur des différences dans leur mécanisme d'intégration (déroulement du cycle de rétrotransposition).

Les rétrotransposons « LTR » (au sens du regroupement de Wicker) sont longs de quelques centaines de pb jusqu'à plus de 20 kb. Ils débutent en 5‘ par un motif 'TG' et finissent par 'CA' en 3'. Les « LTR » codent généralement pour GAG (pour « Group-specific Antigen » ; une protéine structurelle pour les particules de type viral), pour le complexe POL, et enfin pour leurs origines de réplication fonctionnelles (ORF) respectives. Le complexe enzymatique POL regroupe une protease aspartique (aspartic proteinase ; AP), une transcriptase inverse (reverse transcriptase ; RT), une RNase H (RH), et une intégrase (INT). De ce fait, on notera que les rétrotransposons LTRs présentent d'importantes similitudes avec certains rétrovirus. La composition et l'ordre des différentes enzymes citées constitue l'un des critères permettant de subdiviser les LTRs en différentes superfamilles (i.e types) comme Copia ou Gypsy.

Les DIRSs se différencient des LTRs principalement par le remplacement de l'intégrase par une tyrosine recombinase, et par le fait que leurs extrémités sont similaires à des répétitions inversées. Le mécanisme d'intégration des DIRS est donc différent des LTRs.

Les éléments PLEs présentent une machinerie de transposition bien différente des ordres précédemment cités. En effet, ces éléments ne codent que pour deux enzymes : une RT (plus proche des télomérases que des RT des LTR) et une endonucléase (EN). Les Penelopes présentent en outre de longues régions terminales répétées mais dont l'orientation peut varier par rapport aux LTRs.

Les LINEs, dont la taille peut atteindre plusieurs kb, ne présentent pas de longues régions terminales répétées à leur extrémités. Les LINEs sont très diversifiés et sont donc subdivisés en une multitude de superfamilles. Les LINEs encodent au minimum une RT et une nucléase et peuvent présenter une queue poly-A, une répétition en tandem ou encore une région riche en adénine à leur extrémité 3'.

Les SINEs sont des rétrotransposons relativement courts (80-500 pb) qui ne disposent pas de la machinerie nécessaire à leur auto-réplication (i.e. « non-autonomes »). Les SINEs sont le plus souvent le produit accidentel d'une réplication de l'ADN par une polymérase III. La plupart des SINE de mammifères sont ainsi issus de la combinaison d'une tête 5′ qui est dérivée d'un pseudogène ribosomique ou ARNt et d'une queue 3′ homologue à un LINE. La région homologue à un LINE du SINE est utilisée pour parasiter la machinerie enzymatique des LINEs pour transposer (en particulier leur RT).

Les LARDs et les TRIMs sont deux groupes d'ETs non-autonomes qui décrivent respectivement des dérivés longs (> 4 kb) et courts (< 4kb) de LTRs. Les LARDs comportent une région non-codante proche des gènes GAG. Les TRIMs présentent quant à eux des régions non-codantes proches des RT ou INT. Les partenaires autonomes des LARDs et des TRIMs (i.e. les rétrotransposons autonomes dont la machinerie est utilisée) sont le plus souvent méconnus.

Les éléments de classe II (ou transposons à ADN) sont mobilisés via un intermédiaire d'ADN ; soit directement par un mécanisme de "couper-coller" (sous-classe 1 dans la classification de Wicker) ou soit par un mécanisme de réplication du type « peler-copier » (sous-classe 2) impliquant un intermédiaire circulaire d'ADN pour les Helitrons (voir ci-dessous).

Les transposons "couper-coller" (sous-classe 1) comprennent des éléments autonomes répartis entre les TIR (« Terminal Inverted Repeat elements ») et les Cryptons, mais aussi des éléments non-autonomes regroupés dans l'ordre des MITEs (« Miniature Inverted-repeat Transposable Element »). Les éléments « coupés-collés » sont excisés du génome sous forme d'ADN double brin puis réintroduits à une autre position via l'action d'une transposase, une enzyme codée par des instances autonomes de la famille d'éléments en question.

Les TIRs sont caractérisés par des répétitions terminales inversées (« TIR ») et codent pour une transposase qui se lie à proximité des répétitions inversées, coupe l'ADN double brin et sert de médiateur de la mobilité. La transposition des TIR n'est généralement pas un processus réplicatif. Néanmoins il arrive parfois que lors de la réplication des chromosomes une position déjà répliquée transpose à une autre position où la fourchette de réplication n'est pas encore passée. Il arrive aussi que le « trou » causé par l'excision de l'ET soit comblé en se servant de la chromatide sœur comme modèle. L'insertion d'un TIR à un nouveau locus provoque la duplication de courtes séquences au niveau du site cible (TSD pour « target sequence duplication »). La longueur de ces TSD ainsi que la séquence et la taille des répétitions terminales inversées constituent les critères permettant de différencier les 9 superfamilles appartenant à l'ordre TIR.

Les MITEs constituent un groupe hétérogène de courts transposons non-autonomes (< 500 pb) comportant des répétitions terminales inversées flanquées de TSD. Les MITEs dérivent d'éléments TIRs dont ils utilisent la machinerie de transposition pour se répliquer.
L'ordre des Cryptons se distingue des TIRs entre autre par le fait qu'ils encodent une tyrosine recombinase (YR) en lieu et place de la transcriptase. Les extrémités des Cryptons sont difficiles à caractériser car ces éléments ne présentent pas de répétitions terminales inversées ni de longues répétitions mais des répétitions courtes. Par analogie avec l'YR procaryote, il a été proposé que lors de leur mobilisation, les Cryptons sont excisés du génome de l'hôte sous forme d'ADN circulaire extra-chromosomique puis intégrés à un locus différent dans le génome.

La sous-classe 2 de transposons à ADN regroupe les Helitrons et les Mavericks. Les ETs de la sous-classe 2 subissent un processus de transposition très différent de celui de la sous-classe 1 puisque la réplication n'implique le clivage que d'un seul des deux brins d'ADN au site donneur et s'apparente donc à mécanisme du type « peler-copier ».
Les Helitrons, d'une longueur de l'ordre la dizaine de kb, semblent se répliquer via une transposition en « cercle roulant » qui ne génèrent pas de TSD. Les Helitrons autonomes codent pour une protéine tyrosine recombinase de type Y2 (similaire à celle trouvée chez les procaryotes) comportant un domaine hélicase et une séquence initiatrice de réplication. Les Helitrons peuvent également coder pour une protéine de liaison simple brin ou d'autres protéines.
Les Mavericks (aussi appelés Polintrons) mesurent 10 à 20 kb, et sont bordés de longues répétitions terminales inversées. Ces ET peuvent encoder jusqu'à 11 protéines, mais celles-ci varient en nombre et en ordre. Au minimum, les Mavericks codent pour une protéine ADN polymérase B et une intégrase de type C (similaire à celles trouvées chez certains ETs de classe I) mais ne contiennent pas de RT, ce qui suggère qu'ils subissent une transposition réplicative sans intermédiaire à ARN. Le mécanisme proposé

est que l'excision d'un seul brin d'ADN au site donneur est suivie d'une réplication extrachromosomique et que l'ADN double-brin nouvellement formé est ensuite réintégré au niveau du site accepteur.

La classification des ETs est aujourd'hui encore débattue et en constante évolution. Par exemple, les deux systèmes de classification précédemment présentés se concentrent sur les ETs eucaryotes. Or comme le fait remarquer I. Arkhipova, même si en termes de mécanistique les éléments mobiles bactériens et archéens sont parfois très similaires à ceux des eucaryotes, leurs systèmes de classification n'en demeurent pas moins déconnectés (Arkhipova, 2017). C'est pourquoi certains comme Piégu et collaborateurs sont en faveur d'un système de classification universel qui engloberait tous les règnes de la vie et qui permettrait aussi d'inclure certains éléments "négligés" jusqu'alors (Piégu et al., 2015). Par ailleurs, il a aussi été avancé qu'«un système de classification unifié des éléments transposables eucaryotes devrait refléter leur phylogénie » (Seberg and Petersen, 2009), ce qui n'est pas nécessairement le cas dans la classification de Wicker. Cependant, comme le note I. Arkhipova, la nature polyphylétique des ETs, c'est-à-dire la ressemblance entre des ETs ne provenant pas d'une séquence ancestrale commune, rend difficile une telle approche (Arkhipova, 2017).

B- Rôle dans la plasticité fonctionnelle et structurale des génomes

De par leur capacité à se multiplier et à se déplacer, les ETs ont de multiples implications actives et passives sur la plasticité des génomes. La nature de l'impact des ETs est dépendante de la localisation de l'insertion mais aussi des caractéristiques intrinsèques de l'ET et de la composition en ET dans cette région/dans le génome (effets passés en revue de manière extensive dans (Bonchev and Parisod, 2013; Bourgeois and Boissinot, 2019; Bourque et al., 2018; Klein and O'Neill, 2018)).

L'impact des ETs sur la plasticité des génomes peut arbitrairement être réparti entre les effets fonctionnels, régulateurs, et structuraux. Néanmoins, comme nous le verrons au fil des exemples, cette catégorisation est très souvent poreuse, l'action d'un ET pouvant jouer simultanément sur plusieurs tableaux.

1 - Impact sur la séquence et la fonction des gènes

En se déplaçant dans le génome, il arrive qu'un ET s'insère dans un gène et conduise à une perte ou à une modification de fonction observable/notable de ce dernier.

L'un des exemples les plus iconiques concerne les « pois de Mendel ». A la fin du 19 ème siècle, le botaniste autrichien Gregor Mendel réalise des travaux pionniers sur les concepts d'hérédité biologique. Mendel étudie entre autres la transmission de l'aspect lisse ou ridé du pois (*Pisum sativum*) et note que le phénotype « graine ridée » est récessif par rapport au phénotype « graine lisse ». L'étude (entre autres) de ce polymorphisme phénotypique lui permet de formuler les lois qui jetteront les bases de la génétique formelle. Mais ce n'est qu'en 1990 que Bhattacharyya et collaborateurs trouvent l'origine de ce polymorphisme. Le phénotype ridé est causé par l'insertion d'un transposon à ADN TIR de 0.8 kb dans le gène 'rugosa' du pois, qui code pour une enzyme de ramification de l'amidon (Bhattacharyya et al., 1990). Chez l'humain, des insertions d'ETs survenant dans les lignées germinales ont aussi été rapportées ; certaines étant associées à des maladies génétiques héréditaires. Ainsi, en 2016, Hancks & Kazazian comptabilisent 124 maladies humaines pouvant directement être reliées à l'activité de rétrotransposons L1 (LINE) (Hancks and Kazazian, 2016). Par exemple, il a été montré que l'insertion d'un rétrotransposon Alu (SINE ; tributaire de la machinerie de transposition des LINEs) dans le gène FGFR2 (pour « fibroblast growth factor receptor 2 ») constitue l'une des mutations responsables du syndrome d'Apert dont les symptômes sont une malformation importante du crâne (craniosynostose) et des extrémités des membres, souvent accompagnées de déficiences mentales (O'Donnell and Burns, 2010).

Même si dans ce dernier exemple, la mutation a peu de chance d'être transmise sur plusieurs générations, l'exemple des pois de Mendel illustre on ne peut mieux le fait que les mutations engendrées par les ETs,

lorsqu'elles interviennent dans les cellules germinales, sont potentiellement verticalement transmissibles. De ce fait, les ETs sont à l'origine de nombreux polymorphismes génétiques dans les populations (Bourgeois and Boissinot, 2019). Bourque et collaborateurs rapportent ainsi que plus de la moitié de tous les mutants phénotypiques connus (isolés en laboratoire) de la mouche *Drosophila melanogaster* sont causés par des insertions spontanées de divers ETs. Les auteurs de cette revue rapportent aussi que les événements de transposition sont également courants et mutagènes chez les souris de laboratoire, où l'activité continue de plusieurs familles d'éléments LTR est responsable de 10 à 15 % de tous les phénotypes mutants hérités (Bourque et al., 2018).
En plus d'être mobilisés dans les cellules germinales, il a aussi pu être montré que les ETs peuvent être actifs dans les cellules somatiques (Bourque et al., 2018; O'Donnell and Burns, 2010). Chez l'humain, il a ainsi été constaté que l'activité de rétrotransposons L1 (LINE) dans des cellules somatiques pouvait être associée avec des cas de cancer. Il a par exemple été montré que l'insertion d'un ET dans un gène tumo-suppresseur (APC) de cellules du colon était à l'origine de cas de cancers colorectaux (Scott et al., 2016).

Outre le fait de pouvoir induire une perte de fonction, l'insertion d'un ET dans un gène (et en particulier les introns), peut aussi induire un épissage alternatif du transcrit. L'insertion de l'ET peut ainsi potentiellement moduler la séquence, la stabilité, la maturation ou la localisation des ARNs et donc modifier la fonction du gène.
Dans le génome du murier (*Morus notabilis*), une première étude focalisée sur l'impact des MITEs (transposons à ADN non-autonomes) a révélé que ces ETs étaient fréquemment associés avec des évènements d'épissage alternatif et en particulier d'exonisation (acquisition de nouveaux exons issue de régions initialement non codantes ou de séquences d'ETs dans ce cas) (Xin et al., 2019). De plus, une deuxième étude réalisée sur le même génome montre que les évènements d'épissage alternatifs associés à des ETs (tous types confondus) comptent pour plus de 7.5 % de l'ensemble des épissages alternatifs recensés dans ce génome (Ma et al., 2019). Les métazoaires comptent aussi des exemples d'épissage alternatifs induits par des ETs. Les gènes TMPO et ZNF451 sont deux gènes communs à l'ensemble des vertébrés. Chez les mammifères, ces deux gènes présentent chacun plusieurs isoformes, dont un contenant un domaine LAP2alfa. Ces isoformes LAP2alfa sont issus d'un épissage alternatif qui serait dû à l'insertion indépendante d'un rétrotransposon DIRS dans chacun de ces deux gènes (Abascal et al., 2015).

Mais l'impact fonctionnel des ETs ne se limite pas à leur rôle mutagène. En effet, dans certains cas, les insertions d'ETs peuvent fournir la matière première pour l'émergence de nouveaux gènes codant pour des protéines ou des ARNs non-codants.

Par exemple, il a été montré que le gène Arc, un gène neuronal impliqué dans la plasticité synaptique et le transfert intracellulaire d'ARN chez les mammifères, est issu de la domestication du gène gag d'un rétrotransposon Ty3/gypsy (LTR) chez l'ancêtre commun des vertébrés (Pastuzyn et al., 2018). Il est intéressant de noter qu'un mécanisme similaire de transport trans-synaptique d'ARN (Arc1) impliquant une autre version du gène gag d'origine virale existe chez la drosophile (Ashley et al., 2018). La fonction moléculaire de la protéine gag issue d'éléments mobiles a donc été domestiquée indépendamment à plusieurs reprises et a favorisé l'émergence d'innovations cellulaires convergentes chez différents organismes (Bourque et al., 2018). Chez certaines espèces, l'exaptation de gènes issus d'ETs pourrait être un mécanisme fréquent de création de nouveauté génétique comme le suggèrent les travaux de Hoen & Bureau sur le génome d'*Arabidopsis thaliana* (Hoen and Bureau, 2015).

En plus de fournir du matériel génétique susceptible d'être coopté, les ETs contribuent aussi substantiellement aux fonctions cellulaires impliquant des ARNs non-codant. Chez l'humain, HERVH est un rétrotransposon (LTR) préférentiellement exprimé dans les cellules souches embryonnaires. Il a ainsi été montré qu'en plus d'avoir un rôle d'amplificateur ("enhancer"), cet élément code aussi pour un long ARN non-codant nécessaire au maintien de la pluripotence des cellules souches embryonnaires (Gemmell et al., 2019; Lu et al., 2014).

L'ensemble de ces exemples illustre le fait que les ETs, par leurs mouvements, sont des agents mutagènes à part entière dont l'impact peut être au minimum aussi important que celui des mutations nucléotidiques ponctuelles abordées en première partie de l'introduction. L'insertion d'un ET, bien que souvent délétère lorsqu'elle intervient dans un gène, est parfois génitrice de nouveautés génétiques pouvant assumer des fonctions cellulaires importantes ou même essentielles pour l'organisme. Le pouvoir adaptatif des ETs sera analysé plus en détail dans la dernière partie de ce chapitre.

2 - Impact sur la régulation des gènes

En plus de leur impact sur la séquence et la fonctionnalité des gènes, les ETs jouent un rôle prépondérant dans la régulation de leur expression ; et ce de manière active (« régulation cis ») comme passive (« régulation trans ») (voir Figure 2.B.2.1).

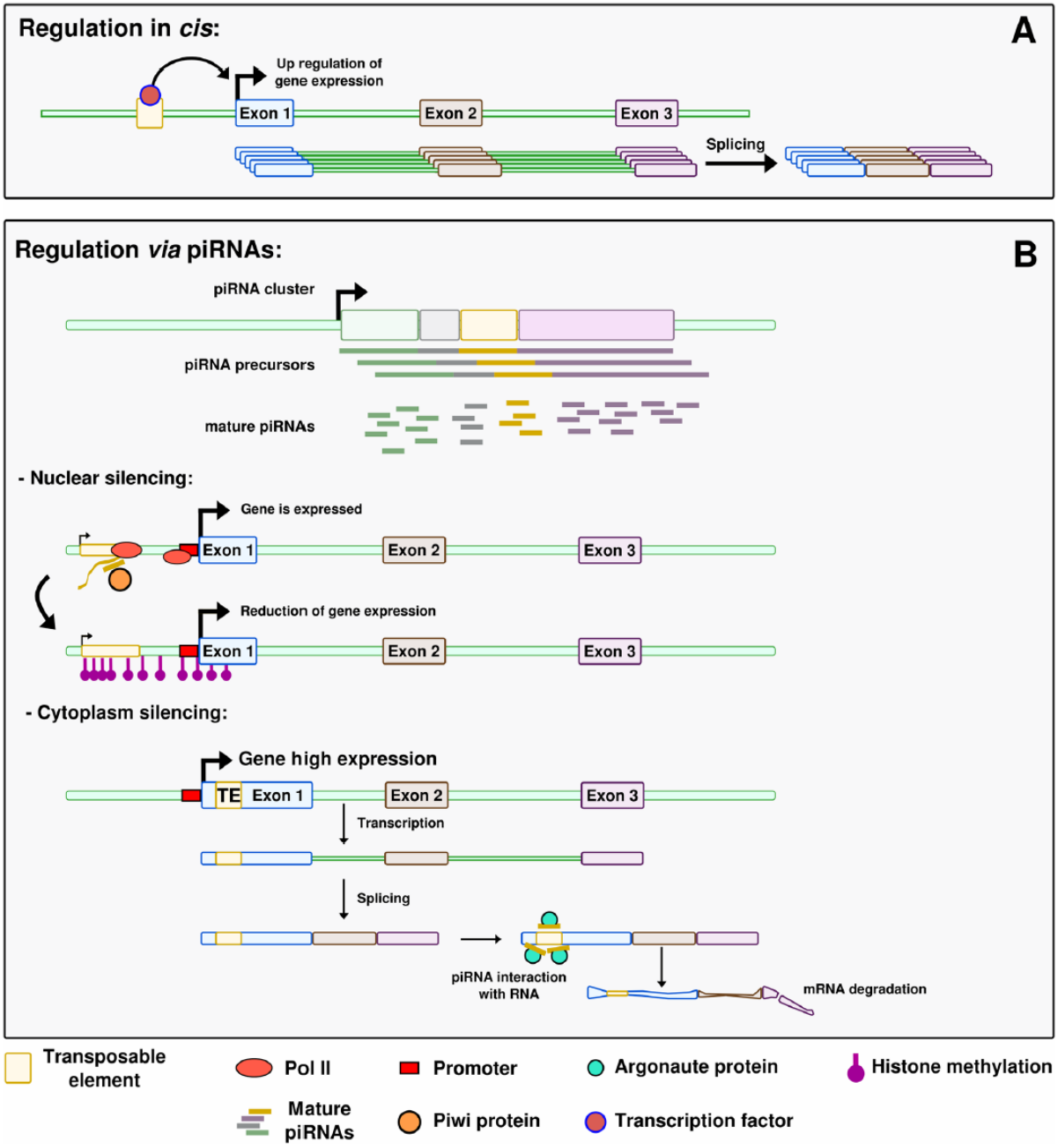


Figure 2.B.2.1 : Régulation en « cis » et « trans » de l'expression des gènes par les ETs.

Figure issue de *(Dechaud et al., 2019)* (Figure 2)

A - Régulation en « cis ». L' ET apporte une séquence de régulation prête à l'emploi qui porte un site de liaison du facteur de transcription. Le facteur de transcription peut se fixer sur ce site et influencer l'expression du gène voisin.

B - Régulation par les piRNAs (« trans »). Répression nucléaire : un ET est présent à proximité du gène d'intérêt. Le piRNA, via la protéine Piwi, déclenche des modifications d'histones qui impactent l'ET mais qui affectent également la région de liaison de l'ARN polymérase du gène voisin. En raison de la modification épigénétique de l'ET, l'expression du gène est réduite. Répression cytoplasmique : une séquence dérivée de l'ET est présente dans le 5'UTR du gène. Des piRNA spécifiques à cet ET se lient au transcrit dans le cytoplasme via une protéine Argonaute et déclenchent la dégradation du transcrit

Les ET peuvent directement influencer l'expression des gènes par la disruption de motifs de régulation répresseurs (« silencer ») ou amplificateurs (« enhancer ») dans les régions promotrices, ou encore par l'insertion de nouveaux éléments de régulation à proximité du gène.

Chez *Drosophila melanogaster*, il a par exemple été montré que des insertions indépendantes d'ETs variés (LINE, LTR) dans des lignées distinctes ont ciblé une même région dans la séquence promotrice du gène Hsp70 (Lerman et al., 2003). Les auteurs notent que la présence de ces ET à cette position réduit l'expression de ce gène et qu'elle peut être associée à une altération de la thermo-tolérance ainsi que du succès reproductif des femelles. Les auteurs en concluent que « la transposition peut créer une variation génétique quantitative de l'expression des gènes au sein des populations, sur laquelle la sélection naturelle peut agir ». L'insertion d'un ET dans la région promotrice d'un gène peut aussi induire des régulations ciblées telles que l'expression tissu-spécifique d'un gène. Chez *Arabidopsis thaliana,* il a été montré que l'expression spécifique dans la racine d'un gène codant pour une protéine associée aux défenses de la plante (Jacalin Lectin Family Protein) nécessitait la présence d'un transposon à ADN TIR (hAT) dans le promoteur du gène (Wu et al., 2018).

Les ETs peuvent également apporter leurs propres séquences régulatrices à côté des séquences codantes et ainsi influencer l'expression des gènes. Par exemple, il a été démontré que l'insertion d'un rétrotransposon LTR en amont du gène Ruby, un gène impliqué dans la production d'anthocyane (pigments végétaux) dans l'orange sanguine fournit un nouveau promoteur contrôlant l'expression des gènes en réponse au froid et entraînant la coloration rouge de la chair du fruit (Butelli et al., 2012). De même, chez *Drosophila melanogaster* l'insertion d'un rétrotransposon LTR (Accord) en amont du promoteur du gène CYP6G1 codant pour un cytochrome P450, un gène de résistance aux insecticides comme le DDT, a permis d'augmenter le niveau d'expression de ce gène et de conférer une résistance adaptative (Schmidt et al., 2010).

Enfin, comme le font remarquer Bonchev & Parisod, il est important de noter que l'expression des gènes n'est pas seulement contrôlée par des promoteurs proximaux, mais qu'elle peut également être influencée par l'insertion d'ET à l'extrémité 3′ des gènes ou bien à des loci relativement éloignés (Bonchev and Parisod, 2013). Les auteurs rapportent ainsi que la variation du temps de floraison du maïs est étroitement associée à l'insertion d'un MITE perturbant une région non codante conservée (Vgt1) située à 70 kb du facteur de transcription AP2 qui régule effectivement ce caractère. De même, la surexpression de tb1, un gène réprimant la ramification dans le maïs cultivé, est contrôlée par l'activité amplificatrice d'une insertion de rétrotransposon LTR située à 60 kb du gène (Bonchev and Parisod, 2013)

L'expression des gènes peut aussi être indirectement influencée par des mécanismes de répression épigénétiques dirigés contre les ETs situés à proximité. Les gènes situés à proximité des insertions des ETs peuvent par exemple être méthylés suite à l'action de petits ARNs ciblant et réprimant

l'expression des ETs. Ainsi, bien que la structure de la région codant pour les protéines reste inchangée, les ETs proches de gènes peuvent être à l'origine d'(épi) allèles (Bonchev and Parisod, 2013). Il a par exemple été montré que l'insertion d'un rétrotransposon LTR en amont du promoteur du gène Agouti chez la souris, un gène responsable entre autre de la couleur du pelage, induit une variation de l'état de la chromatine dans cette région via la méthylation de l'ADN (Morgan et al., 1999). Cette variation de l'état de la chromatine module la capacité du gène Aguti à être transcrit et donc son expression. Le niveau de méthylation du transposon contrôle donc la variation de couleur du pelage de cette souris. De même, chez *Arabidopsis thaliana*, il a été démontré que la floraison précoce de l'écotype Ler est contrôlée par l'insertion d'un transposon dans le premier intron du gène FLC (pour « Flowering Locus C » ; un gène répresseur de la floraison) . L'ET inséré est en effet ciblé par des petits ARNs interférents (siRNAs pour « Small Interfering RNAs ») dérivés d'ETs homologues situés ailleurs dans le génome, ce qui entraîne la répression de l'expression du gène. Les écotypes dépourvus de cette insertion présentent une expression normale du gène FLC et une floraison tardive (Liu, 2004).

Il est intéressant de noter que chez certaines espèces, ce type de régulation pourrait constituer un phénomène de grande ampleur. Par exemple les MITEs représentent 13.83% du génome du mûrier (*Morus notabilis*), et il a pu être montré que ~16% des séquences de MITE (45 577 copies) s'alignent avec des séquences de petits ARNs (Xin et al., 2019), suggérant un impact potentiel à grande échelle sur la régulation des gènes à proximité des ETs concernés (Xin et al., 2019).

A l'opposé, le recrutement des ETs pourrait aussi constituer un mécanisme privilégié pour participer à la régulation épigénétique de l'expression de gènes spécifiques. En utilisant les lignées d'*Arabidopsis thaliana* chez lesquels ils ont provoqué une accumulation d'ETs, Quadrana et collaborateurs ont mis en évidence le rôle essentiel du variant d'histone H2A.Z dans l'intégration préférentielle des rétrotransposons LTR Ty1/copia au sein des gènes sensibles à l'environnement et dans leur maintien à l'écart des gènes essentiels (Quadrana et al., 2019). Ils montrent en outre que la répression épigénétique des copies nouvellement insérées peut moduler leur effet sur des traits liés au succès reproducteur comme le temps de floraison. Les auteurs en concluent que « les ETs sont de puissants (épi)mutagènes épisodiques qui, grâce à des tropismes chromatiniens marqués, limitent la charge de mutation et augmentent le potentiel d'adaptation rapide ».

3 - Impact sur la structure et la taille des génomes

Les ETs participent activement et passivement à la réorganisation des génomes. L'impact des ETs est bien sûr relatif à leur activité (i.e. mouvement) mais aussi au fait qu'ils puissent agir comme substrat de recombinaisons ectopiques. En effet, du fait de leur nature répétée, les ETs peuvent induire des recombinaisons se faisant par homologie entre deux séquences situées sur un même chromosome ou sur des chromosomes différents, et ce, même lorsqu'ils ont perdu leur capacité à être mobilisé. Qu'ils

soient actifs ou passifs, les effets des ETs sur les génomes sont variés puisqu'il sont à l'origine de multiples variations structurales telles que des duplications, délétions, inversions ou encore translocation de régions génomiques plus ou moins grandes, certaines contenant parfois des séquences (ou parties de séquence) codantes (voir (Bourque et al., 2018; Goodier and Kazazian, 2008; Krasileva, 2019; Munoz-Lopez and Garcia-Perez, 2010) pour une revue complète).

Le potentiel impact structural d'un ET est entre autres lié à son mécanisme de transposition (réplicatif e.g. copier-coller et peler-coller, ou non e.g. couper-coller) et à sa séquence (en particulier de ses terminaisons qui peuvent faciliter les recombinaisons ectopiques).
Les transposons à ADN TIR (transposition par « couper-coller ») ont par exemple la particularité de pouvoir être excisés et réintégrés à l'identique à une autre position du génome et cette caractéristique leur a valu d'être utilisé comme outil de transgénèse. Le transposon TIR (Tol2) est par exemple largement utilisé, notamment chez le poisson zèbre, un organisme modèle, comme vecteur de transfert de gènes pouvant efficacement transférer jusqu'à 11 kb de matériel génétique (Kawakami, 2007).

Néanmoins, il est aussi important de noter qu'un même type d'ET peut avoir plusieurs effets possibles. Dans le génome de l'orge (*Hordeum vulgare*), il a été montré qu'un événement de recombinaison entre deux copies de rétrotransposons LTR (BARE) flanquantes du gène HvRA2 (un gène architecte impliqué dans l'inflorescence) était à l'origine de la délétion de ce gène chez la lignée mutante (prbs) (Shang et al., 2017).
A l'inverse, il a été montré que l'augmentation du nombre de copies du gène NLR dans le génome du poivre était due à plusieurs évènements de rétroduplication induit par des rétrotransposons LTR (gypsy) (Kim et al., 2017). Schématiquement, des gènes situés à proximité de rétrotransposons peuvent être capturés lors de l'étape de transcription puis réintégrés ailleurs dans le génome lors de la transcription inverse (Krasileva, 2019). Il est intéressant de noter que ce mécanisme de « trans-duplication » de gènes induit par les rétrotransposons semble être commun dans le vivant puisqu'il est aussi décrit chez les métazoaires dont l'humain chez lequel il est estimé qu'un individu sur 6000 est porteur d'une nouvelle insertion de rétrogène (Bourque et al., 2018).

De plus, des ETs différents peuvent *in fine* produire le même effet. En effet, la duplication de séquences est aussi possible via l'intermédiaire de transposons à ADN, en particulier via des recombinaisons ectopiques.
Chez le Cyprès d'été (*Kochia scoparia*), l'augmentation du nombre de copies du gène EPSPS (pour « 5-énolpyruvylshikimate-3-phosphate synthétase ») confère une résistance au glyphosate, l'un des

herbicides le plus utilisé au monde. Généralement, chez cette espèce, les plantes résistantes au glyphosate possèdent trois à huit copies de l'EPSPS disposées en tandem dans le génome. La comparaison de plantes résistantes et sensibles a permis de retracer l'histoire de ces duplications. Les auteurs proposent que l'insertion d'un élément mobile dérivé de transposons à ADN TIR (MULE, pour « Mutator-Like transposable Elements ») autour du gène EPSPS aurait permis la duplication en tandem de cette région via des enjambements inégaux des chromosomes (« unequal crossing over »), aboutissant ainsi au phénotype résistant (Patterson et al., 2019).

Ainsi, les ETs peuvent être à l'origine de variations du nombre de copies (VNC) de gènes entre populations telles que celles présentées en première partie de l'introduction.

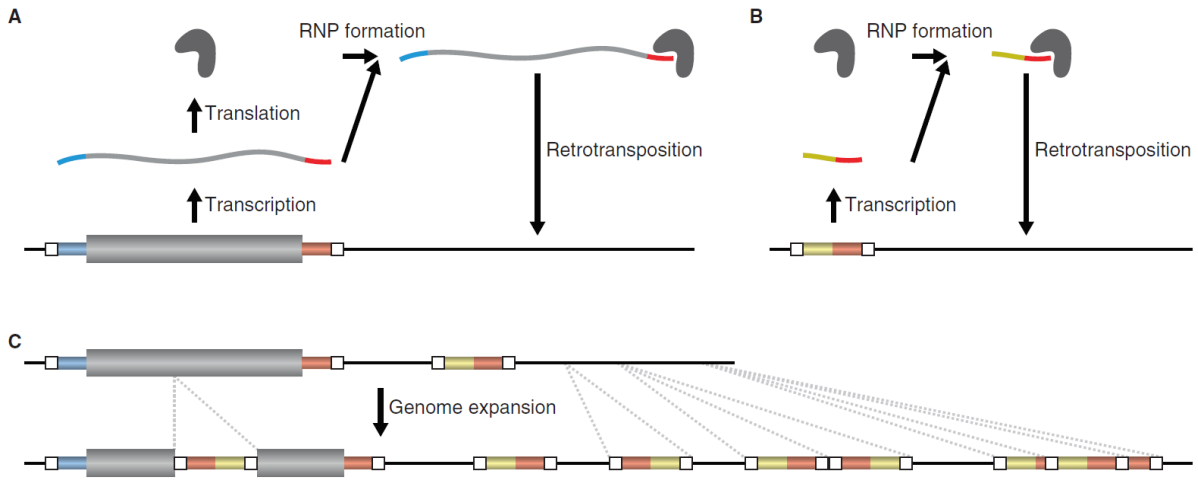


Figure 2.B.3.1 : amplification de la taille des génomes sous l'action d'ETs (rétrotransposons)

Figure issue de (Suh, 2019) (Figure 2)

A - Expansion directe du génome par rétrotransposition d'ETs autonomes (LINE).

B - Expansion indirecte du génome via la rétrotransposition d'ETs non-autonomes (SINE) après détournement de la machinerie enzymatique d'ETs autonomes.

C - Illustration schématique de l'expansion du génome due à l'expansion s'un élement SINE, qui pourrait in fine conduire à l'inactivation de l'éléments LINE fournissant la machinerie enzymatique.

Les lignes pointillées indiquent les événements d'insertion du SINE dans le génome élargi (en-dessous) par rapport à la situation avant l'expansion (au-dessus). RNP : ribonucléoprotéine.

Les queues poly(A) des LINEs /SINEs ne sont pas représentées pour des raisons de simplicité.

Outre les effets ponctuels illustrés par les exemples précédents, les ETs peuvent aussi avoir des impacts structuraux à grande échelle.

La transposition, par exemple, est un moteur majeur de l'évolution de la taille du génome chez les eucaryotes (Suh, 2019). S'ils sont actifs, les ETs ayant un mode de transposition réplicatif (e.g. rétrotransposons et transposons à ADN des ordres Helitron et Maverick) se multiplient et peuvent en effet substantiellement participer à l'amplification de la taille des génomes (Figure 2.B.3.1). Le séquençage de différentes espèces de la classe des tuniciers, des urochordés, a révélé d'importantes variations inter-espèce de la taille du génome (jusqu'à 12 fois la taille du plus petit génome) (Naville et al., 2019). De manière intéressante, aucun indice de duplication du génome entier n'a été trouvé au sein du groupe d'espèces analysées. En revanche, la quantité globale d'ETs était quant à elle fortement corrélée avec la taille du génome. Les auteurs montrent que les rétrotransposons non autonomes SINE

ont massivement contribué à la variation de la taille du génome par des amplifications indépendantes spécifiques à l'espèce, allant de 3 % dans le plus petit génome à 49 % dans le plus grand. Les variations de l'abondance des SINE expliquent jusqu'à 83% de la variation de la taille du génome entre ces espèces. Ces résultats suggèrent que l'expansion des génomes observés chez certaines espèces dans ce groupe d'organismes est due à l'accumulation de rétrotransposons non-autonomes SINEs.
Dans une étude similaire réalisée au sein du genre *Panax*, un groupe de plantes vivaces incluant le ginseng (*Panax ginseng*), il a été mis en évidence que le génome de *Panax quinquefolius* aurait subi une expansion en raison de l'amplification rapide d'un rétrotransposon LTR (PgDel1) au cours du dernier million d'années, suite à une adaptation environnementale consécutive à sa migration de l'Asie vers l'Amérique du Nord (Lee et al., 2017).
Bien que le ou les facteurs déclencheurs d'expansion des génomes induite par les ET ne soient pas clairement connus, les stress engendrés par des variations des conditions du milieu pourraient être l'un des éléments clefs (voir ci-dessous) (Belyayev, 2014).

Néanmoins, l'expansion du génome peut, au fil du temps, être contrecarrée par délétion d'ADN. Comme montré précédemment, les transposons « couper-coller » tels que les TIRs peuvent activement induire l'excision de régions génomiques. Mais certains transposons usuellement réplicatifs peuvent aussi avoir le même effet, en particulier via des recombinaisons ectopiques. En comparant les génomes de l'humain et du chimpanzé, il a ainsi été montré que des événements de recombinaison entre rétrotransposons LINE (L1) avaient mené à l'excision d'environ 450 kb du génome humain, dont 64 kb en une seule délétion (Han et al., 2008). Les auteurs notent néanmoins qu'environ 60 % des séquences d'ADN supprimées sont constituées de séquences L1 qui étaient soit directement impliquées dans les événements de recombinaison, soit situées dans la séquence intermédiaire entre les L1 recombinantes. Toujours en comparant l'humain et le chimpanzé, il a été mis en évidence que 19% des délétions génomiques de 200-500 pb qui se sont produites depuis que ces deux espèces ont divergé sont associées à des répétitions identiques flanquantes d'au moins 10 pb. Un grand nombre de délétions internes aux éléments Alu ont également été trouvées flanquées d'homologies. Sur la base de ces résultats, les auteurs suggèrent que la recombinaison illégitime entre des rétrotransposons SINE a joué un rôle important dans l'évolution du génome humain. Selon eux, « cette étude met en perspective l'idée que les insertions de rétroéléments représentent des événements génétiques unidirectionnels », ce qui est en accord avec l'idée émise plus haut sur les multiples rôles, parfois opposés, que peut avoir un même type d'ET.

Par ailleurs, les ETs participent au remaniement à grande échelle des génomes via la capture et la réorganisation massives de fragments de gènes. Dans le génome du riz, une analyse réalisée sur les transposons à ADN TIR (MULE) a montré qu'environ 3000 d'entre eux comportaient des fragments

d'ADN issus de plus de 1000 gènes non-tranposon. L'analyse de ces éléments chimériques (appelés Pack-MULEs) a montré que les fragments de gènes contenus provenaient souvent de chromosomes différents et avaient parfois fusionnés pour créer de nouvelles séquences codantes ; certaines séquences chimériques étant même transcrites. Ceci suggère que les transposons TIR (MULE) auraient capturé, réarrangé et amplifié quantité de fragments d'ADN dans le génome du riz, et ce vraisemblablement sur une période de temps importante estimée à plusieurs millions d'années (Jiang et al., 2004). Par ailleurs, une étude de génomique comparative menée entre des populations de *Coxiela burnetii* (la bactérie responsable de la fièvre Q) ayant des phénotypes différents a montré que la recombinaison entre d'abondants éléments IS (pour « Insertion Sequence », des transposons bactériens à ADN) a entraîné des réarrangements chromosomiques de blocs synténiques (blocs de gènes dont l'ordre est conservé entre populations/espèces) et des insertions/délétions d'ADN (Beare et al., 2009). A l'issue de cette analyse, les auteurs concluent que la perte de gènes via la formation de pseudogènes a été facilité par les réarrangements chromosomiques médiés par les ETs IS, et que ce mécanisme semble être la principale source de diversité génomique parmi les isolats de *Coxiela burnetii*.

Enfin, les ETs peuvent aussi être à l'origine d'inversions chromosomiques à large échelle. Bien que ce type de réarrangement n'entraîne pas de gain ou de perte de séquence génomique, il contribue à la variation génomique et peut avoir une importance fonctionnelle - par exemple, en provoquant des inversions d'exons (Cordaux and Batzer, 2009). Cordaux & Batzer rapportent ainsi que près de la moitié des inversions qui ont eu lieu dans les génomes de l'homme et du chimpanzé depuis leur divergence ont impliqué les rétrotransposons LINE (L1) et SINE (Alu), et qu'environ 20 % de toutes les inversions peuvent être clairement identifiées comme des produits des événements de recombinaison L1-L1 ou Alu-Alu (Cordaux et Batzer, 2009).

Les ETs, par leur mobilisation ou même juste par leur présence, ont un impact certain sur la dynamique et plasticité des génomes, qu'il s'agisse d'effets fonctionnels, régulateurs ou encore structuraux.

Les ETs sont donc impliqués plus ou moins directement dans la majorité des facteurs de plasticité génomique énoncés en fin de première partie de l'introduction. Leur action sur les génomes est variable et dépend de multiples facteurs comme leur nature, leur mécanisme de transposition ou encore leur charge (e.g. leur nombre), et leur diversité dans les génomes. Certaines de ces innovations, lorsqu'elles interviennent dans les lignées germinales, ont une probabilité variable (entre autres dépendante du mode de reproduction) de devenir un caractère héritable. A ce titre, elles sont soumises aux mécanismes évolutifs de sélection et de divergence qui vont diriger leur devenir dans les génomes.

C- Les ETs dans le vivant : charges et compositions variables

A quelques exceptions près, les ETs sont ubiquitaires dans le vivant. Cependant leur charge, *i.e.* la proportion de génome qu'ils occupent, peut être extrêmement variable d'un organisme à l'autre, et ce parfois même au sein d'une lignée donnée. Par exemple, parmi les eucaryotes unicellulaires parasites, les ETs sont absents du génome de *Plasmodium falciparum* (un des parasites causant le paludisme), alors que le génome de *Trichomonas vaginalis* (parasite humain sexuellement transmissible) est composé à 40% d'ETs. Chez les plantes, ~85% du génome du maïs est composé d'ET, alors que cette valeur n'est que de ~10% chez *Arabidopsis thaliana*. Chez les vertébrés, la charge d'ET varie de ~6% chez le fugu à plus de 50% chez le poisson zèbre ainsi que certains mammifères (Bourgeois and Boissinot, 2019). Ainsi, bien que variable, la part de génome que les ETs représentent est donc non négligeable.

La composition en ETs, i.e. la diversité des ETs rencontrés ainsi que leurs proportions relatives au sein d'un génome, varient également de manière considérable au sein du vivant. Par exemple, Bourgeois & Boissinot rapportent que les génomes des vertébrés non mammifères (poissons, amphibiens, reptiles) contiennent typiquement une grande diversité d'ETs représentés par de nombreuses familles d'éléments de classe I et de classe II (Bourgeois et Boissinot, 2019).
La même conclusion peut être tirée chez les invertébrés de l'embranchement des arthropodes pour lesquels deux études concomitantes, réalisées respectivement sur 14 et 65 espèces, ont décrit une grande diversité d'ETs, la quasi-totalité des ordres d'ETs étant représentés en proportion variables en fonction des espèces (Petersen et al., 2019; Wu and Lu, 2019).
A l'inverse, certains organismes ou lignées présentent une diversité d'ETs extrêmement faible. Ainsi Platt et collaborateurs rapportent que si un tiers à la moitié des génomes des mammifères sont dérivés d'ETs, l'extrême majorité sont des rétrotransposons LINEs et SINES ; les transposons à ADN étant rares et/ou anciens (Platt et al., 2018). Chez l'humain par exemple, pour lequel 45% du génome est occupé par des ETs, seuls 6.22 % sont des transposons à ADN (2.8% du génome). Les ETs restant illustrent bien la faible diversité précédemment évoquée puisque à elles seules, deux superfamilles (les éléments L1 (LINE) (16.9% du génome) et les éléments Alu (SINE) (10.6% du génome)) comptent pour 63% des ETs restants (Cordaux and Batzer, 2009).

Si la composition en ET des génomes varie grandement au sein du vivant, il semble en revanche fréquent qu'elle suive la phylogénie des espèces, créant ainsi des profils de répartition des ETs propres à des lignées ou à des sous-groupes d'espèces.

Chez les vertébrés, Chalopin et collaborateurs rapportent par exemple que la lignée des *Actinopterygii* (les poissons rayonnés comme le fugu ou encore le poisson zèbre) constitue le seul groupe d'espèces vertébrées pour lesquelles les transposons à ADN représentent la majorité de la charge en ETs (Chalopin et al., 2015). Toutes les autres lignées de vertébrés recensées dans cette étude, à l'exception des amphibiens (une seule espèce étudiée), présentent une majorité de rétrotransposons dans leur génome. L'un des cas les plus extrêmes mis en avant par cette étude est le requin éléphant, le vertébré recensé évoluant le moins rapidement (Venkatesh et al., 2014), dont plus de 30 % du génome est composé d'ETs mais chez qui on note une quasi absence de transposons à ADN.

Chez la plupart des arthropodes et en particulier les insectes, la diversité des ETs est aussi majoritairement lignée-spécifique (Petersen et al., 2019). Par exemple, si la superfamille de rétrotransposon Odin (LINE) est absente de tous les hyménoptères étudiés, la superfamille d'éléments à ADN Harbinger (TIR) est quant à elle retrouvée chez tous les lépidoptères, à l'exception du ver à soie *Bombyx mori*. Les auteurs suggèrent que ces absences clades-spécifiques d'une superfamille d'ETs peuvent être le résultat d'événements d'extinction spécifiques à la lignée au cours de l'évolution des différents ordres d'insectes.

Chez les nématodes, une étude menée sur 42 espèces a ici aussi montré que la charge en ET est variable entre groupes d'espèces (Szitenberg et al., 2016). Par exemple, les transposons à ADN Tc1-Mariner (TIR) sont rares chez *Dorylaimida* (Clade I) et abondantes chez *Rhabditina* (Clade V). Néanmoins, on peut aussi noter certains profils d'espèces spécifiques. Par exemple, Onchocerca *volvulus* (*Spirurina* ; Clade III), contrairement à ses parents du clade III, présente une charge élevée en transposons à ADN Helitron et pratiquement aucun autre ET. Malgré ces variations, on peut néanmoins observer que les transposons à ADN sont quasiment toujours majoritaires dans les génomes des nématodes analysés.

Il est intéressant de noter que si chez les eucaryotes il est généralement admis que la charge d'ETs est fortement corrélé à la taille du génome (Bourgeois and Boissinot, 2019), aucune relation simple n'a en revanche été trouvée entre la diversité en ETs et la taille du génome (Elliott and Gregory, 2015). En comparant 257 espèces d'eucaryotes (animaux, plantes, champignons et protistes), Elliott & Gregory ont en effet pu constater que la diversité des ETs (au niveau de la superfamille) augmente en fonction de la taille du génome mais seulement jusqu'à un certain point (environ 500Mbp) ; taille à laquelle ils observent la plus grande diversité d'ETs. Au-delà, ils notent soit une absence de relation chez les animaux, soit une corrélation négative chez les plantes. Les auteurs en concluent qu'il n'y a pas de relation directe entre la taille des génomes eucaryotes et la diversité des ETs au niveau de la superfamille (Elliott and Gregory, 2015).

En résumé, la composition et le pourcentage de génome occupé par les ETs sont très variables au sein du vivant. On peut néanmoins noter que la diversité des ETs ainsi que leurs proportions relatives dans les génomes semblent le plus souvent suivre l'histoire évolutive des espèces.

D- Rôle adaptatif et impact évolutif sur les organismes

Comme ils peuvent se déplacer et se multiplier dans les génomes mais aussi se recombiner et générer différents types de réarrangements, les ETs sont par nature une source importante de variabilité génomique que ce soit entre différentes espèces ou encore entre individus d'une même espèce voir d'une même population. On estime par exemple que chez l'humain, chaque individu porte environ une centaine de retrotransposons L1 (LINEs) actifs, la plupart provenant d'insertions récentes qui se diffusent encore actuellement au sein des populations (Bourque et al., 2018). Il est généralement admis que la plupart des insertions d'ETs sont délétères pour l'hôte, en particulier lorsqu'elles perturbent des gènes essentiels, des régions régulatrices ou des structures chromosomiques, entraînant des effets négatifs allant d'une légère diminution de l'aptitude de l'hôte à des mutations mortelles. Lorsque l'insertion d'un ET est associée à un tel désavantage, elle est généralement contre-sélectionnée et finalement perdue. Le processus de perte peut cependant être modulé par plusieurs facteurs, notamment le coefficient de sélection de l'insertion, son potentiel déséquilibre de liaison avec un allèle avantageux (e.g association préférentielle/non-aléatoire entre allèles au sein d'une population), le taux de recombinaison de la région d'insertion et la taille efficace de la population de l'hôte. Certaines insertions en revanche peuvent être neutres, par exemple si elles se produisent dans des régions génomiques qui n'ont pas d'impact crucial sur l'aptitude de l'hôte (comme les régions pauvres en gènes). Cependant, il est difficile de classer une insertion comme "neutre" une fois pour toutes car elle peut toujours induire des réarrangements chromosomiques par des recombinaisons ectopiques (illégitimes). Enfin, certaines insertions d'ET peuvent être impliquées dans des processus adaptatifs et augmenter en fréquence sur des temps évolutifs plus ou moins long en raison d'une pression de sélection positive plus ou moins forte (Bourgeois and Boissinot, 2019; Dechaud et al., 2019).

Sur des temps évolutifs courts, l'activité des ETs peut entraîner des modifications phénotypiques importantes qui, si elles apportent un avantage sélectif, peuvent rapidement se répandre dans une population. L'exemple le plus emblématique (et sans doute l'un des plus « visuel ») concerne un cas de microévolution chez un lépidoptère (Figure 2.D.1), la phalène du bouleau (*Biston betularia*). Durant la seconde révolution industrielle britannique, la forme claire commune (typica) de la phalène du bouleau a rapidement été supplantée par une forme sombre (carbonaria) jusqu'alors inconnue. La pollution importante à cette période à cause de l'usage massif de charbon dans l'industrie aurait diminué l'efficacité de camouflage de la forme claire et augmenté sa prédation à la faveur de la forme sombre (Cook, 2003). Les mécanismes adaptatifs impliqués dans cet exemple de microévolution sont longtemps restés inconnus mais une récente étude a pu montrer que l'apparition de la forme sombre était due à

l'insertion d'un transposon à ADN dans le premier intron du gène cortex impliqué dans le mélanisme de cet organisme (Hof et al., 2016). Les auteurs ont estimé que l'insertion de cet ET se serait produite autour de 1819, ce qui est cohérent avec les données historiques.

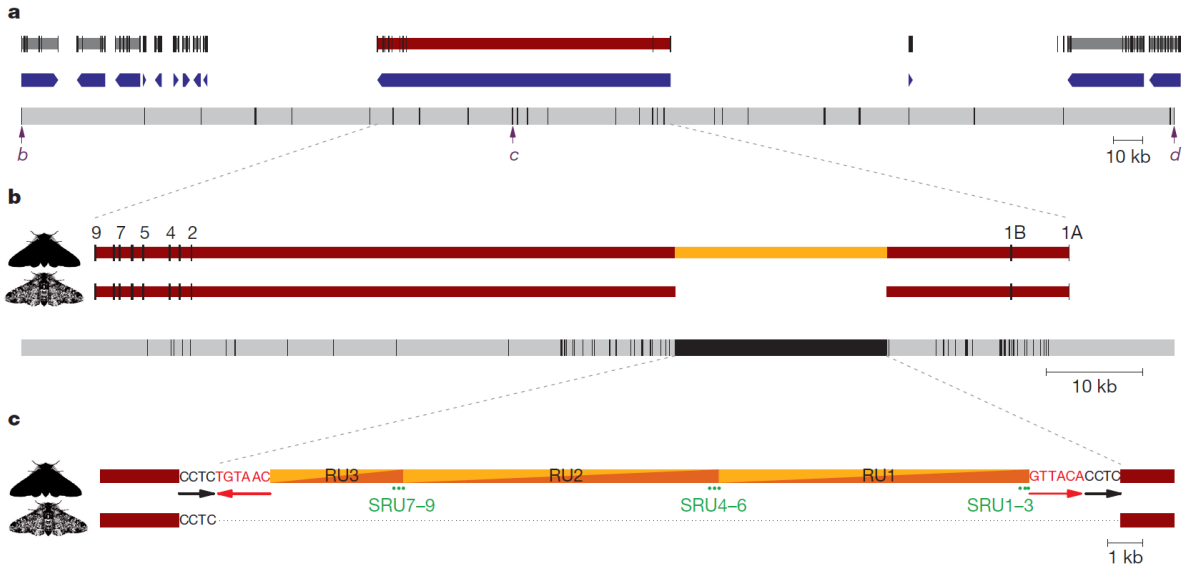


Figure 2.D.1 : la phalène du bouleau *(Biston betularia)*, cas de microévolution due à l'activité d'un ET.

Figure issue de (Hof et al., 2016) (Figure 1)

A - région candidate d'environ 400 kb (délimitée par les loci marqueurs b et d (réf. 2)) indiquant le contenu du gène et les positions du génotype (lignes verticales dans la barre grise continue). La structure et l'orientation de l'intron-exon sont illustrées séparément pour chaque gène (numéro d'accession GenBank : KT182637).

B - Région candidate raffinée comprenant les polymorphismes candidats (lignes sur la barre grise). La structure intron-exon du gene Cortex est illustrée pour la forme carbonaria (papillon noir) et la typica (papillon tacheté), en soulignant la présence d'un grand (22 kb) indel (orange) dans le premier intron. Les exons 1A et 1B sont des débuts de transcription alternatifs suivis par les exons partagés 2-9.

C - Polymorphisme d'ET carbonaria-typica dans la région candidate. La structure de l'insert, représentée dans la séquence de carbonaria, correspond à un transposon à ADN, avec des répétitions directes résultant de la duplication du site cible (TSD pour « Target Site Duplication », représentés par les nucléotides en noirs) à côté des répétitions inversées (nucléotides rouges). Les TSD, les répétitions inversées et la séquence du transposon sont absents de l'haplotype de tipica. Le transposon est une séquence de ~9 kb répétés en tandem 2.3 fois (unité de répétition (RU)1-RU3), avec trois courtes unités de sous-répétition en tandem (points verts, SRU1-SRU9) à l'intérieur de chaque unité de répétition. Les images des deux formes de lépidoptères ont été créées à partir de photographies prises par A.E.v.H.

Cet aspect mutagène des ETs peut aussi se révéler particulièrement utile chez certains organismes dont le mode de vie requiert une course à l'armement incessante avec leur environnement.

Par exemple chez le nématode phytoparasite *M. javanica* (espèce à reproduction strictement asexuée), la comparaison entre une lignée avirulente (e.g. incapable d'infecter les plants de tomates portant un gène de résistance aux nématodes) et une autre lignée virulente (qui a surmonté cette résistance), a conduit à l'identification d'un gène présent chez les nématodes avirulents mais absent chez les virulents. Ce gène (Cg-1) se trouve être inclus dans un transposon à ADN de type TIR (Tm1). Le complexe Tm1/Cg1 étant absent dans la lignée virulente, il a été fait l'hypothèse que cela était dû à l'excision du transposon et donc que cet ET avait joué un rôle dans l'acquisition du pouvoir pathogène de cet organisme (Gross and Williamson, 2011).

Outre cet aspect mutagène, les ETs peuvent aussi constituer un « réservoir » de diversité génétique mobilisable lors de variations des conditions du milieu, la sélection s'appliquant donc sur un paysage d'ET existant.

Il a ainsi été fait l'hypothèse que les ETs pourraient participer à l'adaptation rapide d'espèces nouvellement introduites (invasives) à de nouvelles conditions, et ce malgré une faible diversité génétique intra-populationnelle (Stapley et al., 2015). En effet, à la suite d'une migration (évènement fondateur), les populations invasives présentent le plus souvent une diversité génétique réduite en raison d'une diminution drastique de la taille de la population (goulot d'étranglement génétique). Il est généralement admis que cette perte de variation génétique, ainsi que l'augmentation de la probabilité de consanguinité et d'extinction, limitent la capacité d'une population à s'adapter à de nouveaux environnements. La quantité et la nature de la variation génétique disponible pour la sélection lors des invasions déterminent donc largement le potentiel d'adaptation de la population (Schrader and Schmitz, 2019; Stapley et al., 2015).

Chez *Drosophila melanogaster*, une étude de génomique comparative a permis de montrer que la diversité génétique apportée par quelques ETs a participé à la colonisation de l'Amérique du Nord par cette mouche depuis sa sortie d'Afrique liée à l'activité humaine (quelques centaines d'années donc) (González et al., 2008). En effet, il a pu être montré que 13 ETs, déjà présents dans le génome, avaient vu leur fréquence très fortement augmenter suite à la migration depuis l'Afrique. Par ailleurs, les auteurs ont pu montrer qu'une grande partie de ces loci (ayant un rôle putatif dans la régulation de gène) étaient sous sélection positive, et auraient donc eu un rôle adaptatif en participant notamment à l'adaptation au climat tempéré nord-américain.

Toujours en ce qui concerne les espèces invasives, il semble même que chez certains organismes ce pouvoir adaptatif des ETs soit potentialisé par une structure génomique caractéristique appelée « génome à deux vitesses ».

Dans ces génomes, les ETs (en évolution rapide) sont majoritairement regroupés dans des « îlots », pauvres en gènes, distincts du reste du génome (Schrader and Schmitz, 2019).

Chez la fourmi *Cardiocondyla obscurior*, ces « îlots d'ETs » couvrent environ 7 % du génome et abritent des gènes soupçonnés d'être particulièrement importants lors des adaptations des populations fondatrices à des environnements nouveaux (Schrader et al., 2014). Les auteurs font l'hypothèse que le stress environnemental consécutif aux événements fondateurs induit une brusque augmentation de l'activité des ETs qui génèrent relativement rapidement une variation génétique héréditaire sur quelques générations, facilitant ainsi l'évolution de phénotypes adaptés localement (Schrader et al., 2014).

Une explication serait que, dans ce type de génome, les compartiments à évolution rapide sont généralement enrichis en ETs qui favorisent les changements génomiques en provoquant des ruptures d'ADN lors de leur excision, ou encore en agissant comme substrat pour des réarrangements (Seidl and Thomma, 2014).

Les observations faites par Faino et collaborateurs chez le champignon phytopathogène *Verticillium dahliae* (organisme présumé asexué) semblent aller dans ce sens puisqu'ils ont pu mettre en évidence que les transposons sont la principale force motrice de l'évolution adaptative de son génome (Faino et al., 2016). Dans cette analyse, les auteurs montrent en effet que des régions très variables lignée-spécifiques (LS) ont évolué par des réarrangements génomiques intervenus lors de réparation erronées de cassure de l'ADN, vraisemblablement dues à des mouvements de transposons. De plus, ils ont pu montrer que les régions LS sont enrichies en transposons actifs, ce qui contribue à la plasticité locale du génome. Enfin ils montrent que la totalité des gènes effecteurs fonctionnellement caractérisés (gènes a priori impliqués dans le parasitisme) sont concentrés dans une seule de ces régions LS enrichie en ETs (∼2-Mb). Les auteurs en concluent que les ETs ont modelé le génome de ce champignon de manière active et passive, et que cette dynamique induite par les ETs impacte la virulence de ce pathogène.

De manière très intéressante il a été montré que le génome de *A. thaliana*, une plante infectée par le phytopathogène précédemment décrit, présente lui aussi des régions riches en ETsréa comportant des groupes de gènes impliqués dans le système immunitaire de la plante (rapporté dans (Seidl and Thomma, 2017)). Selon les auteurs les ETs seraient donc au cœur d'une coévolution entre les plantes et les pathogènes en étant des acteurs essentiels d'une course à l'armement pour le compte de l'ensemble des belligérants.

On remarquera que le manque de diversité génétique est une caractéristique commune entre les espèces invasives lors de l'évènement fondateur et les espèces à reproduction asexuée. La quantité de données exploitable étant de plus en plus importante, mettre en relation ces deux champs d'étude et la théorie qui les accompagne pourrait se révéler très instructif.

L'impact des ETs sur les génomes se fait aussi ressentir sur des temps évolutifs intermédiaires, les ETs ayant fréquemment été associés avec des adaptations environnementales, et ce des bactéries aux mammifères (Casacuberta and González, 2013).

Chez les bactéries, plusieurs liens clairs ont été établis entre des éléments IS et l'adaptation à des conditions extrêmes comme une haute osmolarité ou encore la résistance à des solvants toxiques (Casacuberta and González, 2013). Chez les plantes, plusieurs adaptations locales ont aussi pu être directement associées avec l'activité d'ETs. Chez le soja par exemple, deux copies du gène Gmphy codent pour la phytochrome A, une protéine responsable de la sensibilité au photopériodisme (rapport jour/nuit) impliquée, entre autres, dans le processus de floraison (période de l'année). Dans certaines lignées de soja, il a pu être montré qu'un rétrotransposon LTR s'était inséré dans l'une des deux copies (GmphyA2), provoquant une perte de sensibilité au photopériodisme (Kanazawa et al., 2009). Les auteurs de cette étude ont pu mettre évidence que cette insertion n'était détectée que dans les lignées de soja cultivées dans les régions du nord du Japon, ce qui suggère que l'insensibilité à la photopériode causée par le dysfonctionnement de GmphyA2 est l'une des modifications génétiques qui ont permis la culture du soja à des latitudes élevées. Enfin chez les mammifères, il a été montré que POMC, un gène impliqué dans la réponse au stress, la régulation de la prise de nourriture, et la gestion de la balance énergétique possède deux enhancers fonctionnels provenant de deux évènements indépendants d'insertion d'un ET au cours de l'évolution (Franchini et al., 2011). Il a été fait l'hypothèse que la présence de ces deux enhancers a été un élément clef de l'évolution des mammifères, par exemple lors de brusques changements climatiques (Casacuberta and González, 2013; Franchini et al., 2011).

Sur des temps évolutifs longs, là encore les ETs ont joué un rôle prépondérant dans l'évolution du vivant. Comme nous l'avons vu dans les exemples précédemment cités, les ETs peuvent induire des variations génomiques et phénotypiques ponctuelles. Cependant, ainsi que le rapporte exhaustivement A. Belyayev dans sa revue, une explosion d'ETs (burst d'ETs), i.e. une multiplication rapide et conséquente d'un ou plusieurs ETs, peut quant à elle induire un remodelage radical des génomes et avoir des conséquences évolutives massives (Belyayev, 2014; Bourgeois and Boissinot, 2019; Schrader and Schmitz, 2019).

Au cours de l'évolution, les burst d'ETs ont plusieurs fois été associés avec la formation de groupes taxonomiques et d'espèces (Belyayev, 2014). La radiation précoce des *Vespertilionidae*, la famille de chauves-souris la plus riche en espèces (> 400 espèces), coïncide ainsi avec une explosion de l'activité des ETs et en particulier d'Helitrons, des transposons à ADN (Feschotte and Pritham, 2007; Platt et al., 2014). Par ailleurs, il a pu être noté que durant la diversification des espèces de lézards du genre *Anolis*, une accumulation d'ETs avait eu lieu dans les gènes HOX; des gènes étant impliqués dans le développement et l'adaptation morphologique à de nouveaux habitats chez ces lézards (Feiner, 2016). Enfin, Pace & Feschotte rapportent qu'une forte activité des ETs (en particulier l'insertion massive

d'éléments SINE) aurait eu lieu chez les mammifères entre -85 et -63 Ma, et serait donc concomitante avec l'apparition des premiers primates (Pace and Feschotte, 2007). Néanmoins, même s'il est clair que des burst d'ETs ont eu lieu lors de transitions évolutives, le lien causal entre les burst d'ETs et spéciation reste débattu. En effet, bien qu'il ait été montré que des burst d'ETs interviennent en condition de stress environnementaux (idée initialement suggérée par Barbara McClintock), aucun lien direct de causalité n'a pu être fait entre ces explosion d'ETs, une innovation adaptative clef et divergente, et enfin la spéciation (voir (Belyayev, 2014; Serrato-Capuchina and Matute, 2018) pour une revue complète).

Toujours à long terme, la présence et l'activité d'ETs dans un génome peut profondément modifier sa structure. Il a ainsi été rapporté que les ETs influent sur la structure des chromosomes sexuels et pourraient être impliqués dans leur différenciation et leur évolution (Dechaud et al., 2019).
Une étude comparative de la distribution des ETs entre les chromosomes sexuels (gonosomes) et non-sexuels (autosomes) chez 7 espèces de poissons a montré une accumulation de transposons suite à un burst récent dans certaines régions des gonosomes et en particulier autour de loci récents de détermination du sexe (Chalopin et al., 2015). Par ailleurs, cette étude montre aussi une activité récente et spécifique de certains ETs dans ces régions ce qui suggérerait que le processus d'accumulation des ETs sur ces chromosomes est toujours en cours. Il a été proposé que l'accumulation des ETs peut parfois se révéler être un moyen efficace pour les chromosomes sexuels naissants d'acquérir des différences structurelles qui interfèrent ensuite avec l'appariement des chromosomes homologues durant la méiose, contribuant ainsi à supprimer la recombinaison entre eux ce qui favorise leur divergence (Dechaud et al., 2019; Schartl et al., 2016). Selon ce modèle, des accumulations massives d'ETs font que les chromosomes sexuels augmentent en taille durant une phase initiale puis, sous l'influence de la dégénérescence génétique, se contractent et deviennent petits.
Par ailleurs, il a été fait l'hypothèse que l'accumulation d'ETs (et d'autres séquences répétitives) sur le chromosome Y de la drosophile aurait pu avoir un impact global sur le paysage chromatinien de son génome (Dechaud et al., 2019). En effet, il a été montré que des variations de la charge d'éléments répétés entre les chromosomes Y de différentes populations de drosophiles étaient associées avec des modifications épigénétiques induisant des variations du niveau de compaction de l'ADN à différents loci sur des autosomes (chromosomes non sexuels) (Lemos et al., 2010).

Il a aussi été fait l'hypothèse que les ETs pourraient participer au maintien de la reproduction sexuée à cause de l'importance du mécanisme de recombinaison génétique (Dechaud et al., 2019). La reproduction sexuée favorise la transmission verticale d'ETs car, si une insertion survient dans le gamète d'un des deux parents, il existe une probabilité pour que cette mutation soit transmise à la génération suivante lors de la fusion du matériel génétique. Mais la reproduction sexuée implique aussi la recombinaison des chromosomes homologues lors de la méiose. Or, en permettant l'échange

d'informations génétiques entre les chromosomes homologues, la recombinaison a un impact antagoniste sur le taux de fixation des ETs en favorisant l'élimination des insertions d'ETs délétères. Ainsi, la recombinaison entrainée par la reproduction sexuée peut être considérée comme un mécanisme de défense contre l'insertion d'ETs délétères et il est donc probable qu'il existe une forte pression de sélection pour maintenir les mécanismes de recombinaison en place. Selon ce raisonnement, des taux élevés de mutations délétères (dont l'insertion d'ETs) pourraient donc favoriser le maintien de la reproduction sexuée comme moyen efficace de maintenir ces mutations à des niveaux compatibles avec la vie (Dechaud et al., 2019).

Enfin, il est largement admis aujourd'hui que l'évolution de divers mécanismes cellulaires fondamentaux chez de nombreuses espèces est à l'origine due à la domestication de certains ETs (Schrader and Schmitz, 2019).

Ainsi, les ETs, en fournissant des éléments régulateurs prêts à l'emploi, participent par exemple à la régulation de l'expression de gènes du développement sexuel (exhaustivement résumé dans (Dechaud et al., 2019)). Chez les mouches drosophiles, les gènes du chromosome X sont régulés de manière coordonnée par le complexe létal spécifique masculin (MSL), ce qui permet d'augmenter l'expression des gènes voisins (du MSL) chez les mâles XY et ainsi compenser l'absence d'un chromosome X par rapport aux femelles XX (compensation de dosage). Au sein du genre *Drosophila*, *miranda* est une espèce qui possède plusieurs chromosomes sexuels (dont XR et néo-X) apparus au fil de son évolution. Ellison & Bachtrog ont pu montrer que l'acquisition de douzaines de sites de liaison MSL (MRE) sur les nouveaux chromosomes X a été facilitée par des évènements d'insertion indépendants d'un Helitron qui attire le complexe MSL. Le chromosome sexuel néo-X récemment formé (~1 Ma) recrute des ETs Helitrons qui fournissent des dizaines de sites de liaison MSL fonctionnels, mais sous-optimaux, alors que l'ancien chromosome XR a cessé d'acquérir de nouveaux sites mais semble avoir amélioré l'affinité des sites existants pour le complexe MSL (Ellison and Bachtrog, 2013). Cet exemple illustre l'efficacité des ETs dans le « re-câblage » des réseaux de régulation des gènes, car ils peuvent répandre des sites de liaison, des facteurs de transcription, ou d'autres types de séquence régulatrices qui peuvent à leur tour co-réguler plusieurs gènes.

Par ailleurs, il existe de nombreux exemples prouvant que les éléments mobiles peuvent servir de réservoir dynamique pour de nouvelles fonctions cellulaires (Volff, 2006).

En effet, il n'est pas rare en particulier que la machinerie de transposition codée par des éléments mobiles soit recrutée par des organismes et ce parfois de manière indépendante dans différentes lignées (Schrader and Schmitz, 2019). L'exemple le plus classique concerne les mouches du genre *drosophila* chez qui les télomères sont maintenus non pas par des télomérases comme c'est majoritairement le cas dans le monde du vivant mais par trois rétrotransposons LINE domestiqués (Pardue et al., 2005). HeT-A (pour

Healing Transposon), TART (pour Telomere Associated Retrotransposon) et THARE (pour Telomere Associated and HeT-A Related) ajoutent activement et en tandem leurs longues répétitions aux régions terminales des chromosomes pour compenser la perte des nucléotides terminaux pendant la réplication de l'ADN (Casacuberta, 2017). De même, RAG 1 & 2, deux gènes essentiels du système immunitaire chez l'homme et d'autres vertébrés à mâchoires (impliqués dans la recombinaison somatique du V(D)J lors du développement des lymphocytes B et T), ont évolué grâce à la domestication d'un ancien transposon à ADN de l'ordre des TIR, produisant un gène chimérique combinant des séquences codant pour l'hôte et l'ET (Huang et al., 2016; Kapitonov and Jurka, 2005). Néanmoins, tous les cas de domestication d'éléments mobiles ne se limitent pas à la cooptation de leur machinerie de transposition. La syncytine par exemple, qui a un rôle important dans la fusion cellule-cellule pendant le développement placentaire de l'hôte a été domestiquée à partir du gène d'enveloppe d'un rétrovirus endogène (ERV) (Figure 2.D.2). Il est intéressant de noter que, chez les mammifères vivipares, la syncytine a été domestiquée plusieurs fois indépendamment au cours des 150 derniers millions d'années (revue par (Kaneko-Ishino and Ishino, 2012))

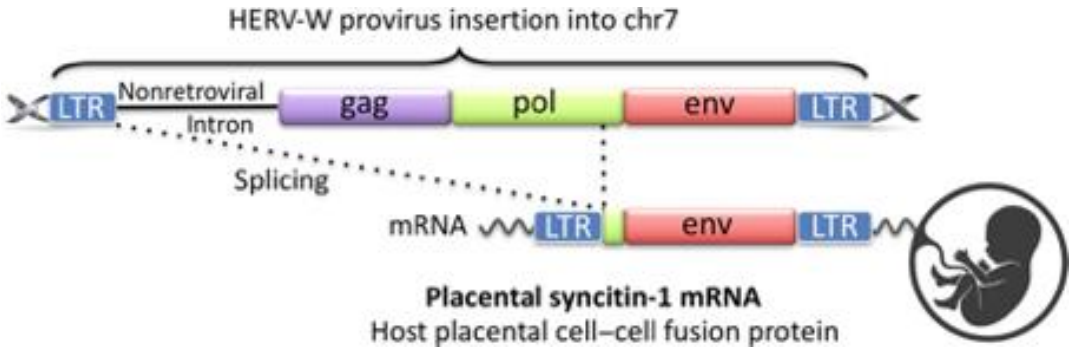


Figure 2.D.2 : Acquisition d'un élément génétique fonctionnel dérivé d'un ET.

Figure issue de (Schrader and Schmitz, 2019) (Figure 1)

Un provirus HERV-W s'est inséré il y a environ 35 millions d'années dans le chromosome 7 de la lignée germinale de l'ancêtre commun de la lignée des Catarrhiniens (« singes de l'Ancien Monde »). L'ORF rétroviral originel du gène d'enveloppe épissé a été domestiqué sous la forme d'un gène à un seul exon. Chez les singes, y compris l'homme, l'ARNm exprimé sert de médiateur dans la fusion cellule à cellules des trophoblaste, fonction essentielle au développement normal du placenta (symbolisé par l'embryon humain)

De par leur capacité à se déplacer et à se multiplier dans les génomes, les ETs sont des acteurs prépondérants de la plasticité génomique. Qu'ils agissent comme une nouveauté génétique ou bien comme une source de diversité pour répondre à un environnement changeant, l'apport génétique des ETs, bien que le plus souvent éliminés des génomes, peut parfois conférer un avantage sélectif et participer à l'adaptation des espèces et ce même en un très court laps de temps. A ce titre les ETs représentent un mécanisme de choix pour expliquer la remarquable adaptabilité des espèces de *Meloidogyne* en absence de recombinaison sexuée et de brassage génétique.

III – Objectifs de la thèse

Les *Meloidogyne* sont des phyto-parasites constituant un groupe d'espèces diversifié en terme de traits biologiques (ploïdie, origine hybride de certaines espèces, mode de reproduction) et de traits de vie (gamme d'hôte, aire de répartition). Les *Meloidogyne* présentent une vaste gamme de modes de reproduction et, de manière intéressante, il a pu être observé que les espèces les plus nuisibles se reproduisent de manière strictement asexuée. Par ailleurs, il a été constaté que certaines de ces espèces à reproduction asexuée arrivent à contourner les défenses de plantes hôtes (gène de résistance au parasitisme) en un nombre de génération restreint. En l'absence de brassage génétique et de recombinaison, le pouvoir adaptatif d'un organisme est en théorie restreint et l'adaptabilité de ces *Meloidogyne* apparaît donc comme un paradoxe évolutif. Des études récentes ont montré qu'il y avait très peu de variations au niveau de la séquence nucléotidique à l'échelle du génome complet entre isolats différents chez *M. incognita* (Koutsovoulos et al., 2020). Ceci confirme les observations précédentes menées sur plusieurs espèces de *Meloidogyne* à partir de quelques marqueurs génétiques. A l'heure actuelle, plusieurs facteurs générant de la plasticité génomique et de la diversité génétique nécessaires à cette adaptabilité ont été étudiés mais aucun ne résout entièrement ce paradoxe.

Les ETs, de par leur capacité à se déplacer et à se multiplier dans les génomes, sont des acteurs reconnus de la dynamique des génomes et peuvent en outre directement être mis en relation avec plusieurs des facteurs de plasticité déjà étudiés pour ces espèces. Il est donc important d'estimer si les ETs jouent un rôle dans la plasticité des génomes dans le genre *Meloidogyne* et si cette plasticité peut ensuite passer le crible de la sélection et constituer un moteur de l'adaptation dans ces espèces.

L'objectif général de cette thèse est d'évaluer le rôle des Éléments Transposables dans la dynamique des génomes des espèces du genre *Meloidogyne*. Trois angles de vues distincts correspondant en réalité à une déclinaison de cette question sur trois temps évolutifs différents ont été exploré :

- A l'échelle du genre *Meloidogyne*: Le paysage en Éléments Transposables (charge et diversité) des génomes varie-t-il entre espèces en fonction de traits biologiques et / ou de leur histoire évolutive? (Chapitre V)

- A l'échelle de l'espèce au sein du genre : Quelle est la contribution des Éléments Transposables à la plasticité génomique et quel est son impact sur l'adaptabilité de cette espèce ? (Chapitre VI)

- A l'échelle de l'isolat au sein d'une espèce : Quel est l'impact des Éléments Transposables dans la diversité génétique d'un organisme supposé clonal ? (Chapitre VII)

Au préalable, pour explorer l'ensemble de ces questions et tenter d'y répondre, j'ai mis en place une méthodologie rigoureuse et dédiée aux spécificités de mon modèle d'étude (Chapitre IV). Méthodologie que j'ai pu ensuite appliquer à d'autres espèces dans le cadre de collaborations.

IV – Mise en place d'une méthodologie

A – Prédiction de-novo et annotation des ETs dans les génomes

1 - Contexte

La technologie se démocratisant, le nombre de génomes séquencés croît de jour en jour. Ceci ouvre de nombreuses perspectives de recherche, en particulier en ce qui concerne les organismes non-modèles. Dans ce contexte, définir le contenu en ET dans les génomes est nécessaire, qu'il s'agisse d'une finalité, ou bien d'une étape intermédiaire dans le cadre d'une analyse bio-informatique plus vaste.

Le contenu en ET d'un génome se décrit via deux caractéristiques : sa charge et sa composition. La charge d'ET correspond à la proportion du génome occupée par les ETs. La notion de composition regroupe quant à elle la diversité des ETs rencontrés mais aussi la proportion relative de ces différents ETs au sein d'un génome. La description de la composition et de la charge du contenu en ET d'un génome peut être réalisée selon différentes méthodes mais l'approche la plus répandue consiste à réaliser une annotation par homologie de séquence. L'annotation par homologie de séquence consiste à identifier l'ensemble des positions d'un génome assemblé correspondant à des copies (ou occurrences) d'un ET donné sur la base de la similitude entre les séquences de régions génomiques et celle de l'ET (séquence de référence). Les séquences de référence d'ETs, aussi appelées séquences consensus, sont regroupées en librairies qui se doivent d'être aussi représentatives que possible de la diversité des ETs rencontrés dans le génome. Lors d'une approche par homologie de séquence, la qualité de la librairie de séquence consensus est le point critique qui va conditionner la qualité de l'annotation.

L'annotation en ET d'un génome peut être réalisée à partir de bases de données existantes et potentiellement curées ou bien à partir d'une librairie nouvellement créée via une étape de détection/prédiction *de novo*. Le choix d'effectuer l'annotation d'un génome à partir d'une librairie existante sera souvent conditionné par la complétude de ladite librairie, mais aussi par la divergence entre le génome à annoter et ceux desquels sont issus les séquences consensus d'ETs présentes dans la librairie. En effet, dans le cas où les séquences de référence d'ETs sont trop divergentes du génome à annoter, seules les parties les plus conservées des séquences pourront correctement être alignées sur le génome ; créant ainsi une annotation fragmentaire et incomplète. Par ailleurs, cette approche rend

impossible l'annotation d'ETs / types d'ETs spécifiques à une lignée ou à un groupe d'espèces si ces séquences ne sont pas disponibles dans la base de données. Pour résumer, dans une approche où la similitude entre des séquences de référence d'ETs et les séquences génomiques sont la clef, le fait d'avoir des séquences de référence d'ETs trop divergentes du génome à annoter pose un problème majeur. Ainsi, dans le cadre de l'étude d'organismes « non-modèles », il est fréquent de devoir réaliser une étape de prédiction *de novo* avant de pouvoir réaliser l'annotation du génome. On notera tout de même que la constitution d'une librairie de séquence *de novo* nécessite le plus souvent d'avoir recours à des ETs déjà caractérisés et décrits afin de pouvoir classifier les séquences nouvellement découvertes.

Il existe de nombreux outils dédiés à la caractérisation du contenu en ETs dans les génomes et ce nombre ne cesse de croître (voir Table 1 (Lerat, 2010), Table 1 (Ewing, 2015) et Table 3 (Makałowski et al., 2019) pour avoir une idée de l'évolution rapide de ce domaine ainsi que de la richesse des outils). Ces outils peuvent pour certains être utilisés seuls, en étant par exemple dédiés à la détection d'un type d'ET en particulier, ou bien être regroupés sous forme de « pipeline » i.e une suite d'outils dont les traitements sont réalisés selon un ordre donné. En ce qui concerne les pipelines d'analyse, certains sont capables de réaliser la prédiction de novo d'une librairie d'ETs consensus ainsi que leur annotation et d'autres sont spécifiques de l'un ou l'autre. Face à cette diversité, il est nécessaire d'envisager plusieurs options afin de choisir la méthodologie correspondant le mieux aux besoins de l'analyse à réaliser.

2 - Méthodologie

Choix de l'outil

Dans le but premier de décrire le contenu en ETs des génomes de *Meloidogyne,* j'ai décidé de concentrer mon attention sur 3 pipelines généralistes (capable de détecter la plupart des types d'ETs) : RepeatModeler/RepeatMasker (http://www.repeatmasker.org/RepeatModeler/ ; http://www.repeatmasker.org/RepeatMasker/), CARP (Zeng et al., 2018), et REPET (Flutre et al., 2011; Quesneville et al., 2005) afin de les tester et d'en choisir un qui soit adapté à mes besoins. Le choix de tester ces outils en particulier s'est basé sur différents critères : i) leur capacité à réaliser la prédiction *de novo* d'une librairie d'ETs propre au génome, ainsi que leur annotation, ii) leur popularité (RepeatMasker et REPET sont les plus utilisés) et iii) de manière plus subjective, sur leur "philosophie" (CARP et REPET).

Je tiens dès à présent à préciser que, faute d'avoir mis en place un protocole de comparaison rigoureux à l'époque à laquelle j'ai réalisé les tests présentés ci-dessous, certaines des conclusions que j'en ai tirées sont *de facto* subjectives et seront discutées plus loin.

Parmi les outils évoqués, RepeatMasker est sans aucun conteste le plus utilisé. RepeatMasker est un pipeline d'annotation par homologie de séquence uniquement. RepeatMasker repose principalement sur l'utilisation de la base de données RepBase (Bao et al., 2015), même s' il est aussi possible de fournir une autre librairie de séquences consensus. RepBase est une librairie de séquences consensus d'ETs qui, le plus souvent, sont issues d'organismes modèles tels que l'humain, la souris, ou encore la drosophile. Cette approche comporte des avantages et des inconvénients. Le principal avantage est que comme les organismes cités sont largement étudiés, une part importante des séquences représentées dans RepBase sont curées et donc fiables. Le revers de la médaille concerne le manque de diversité des organismes représentés. Dans sa version de 2015 (soit sensiblement la version utilisée à l'époque de ces tests), 90% des séquences présentes dans cette librairie étaient issues de 134 espèces seulement (Bao et al., 2015). Par ailleurs, seules 2 espèces de nématodes étaient représentées dont l'organisme modèle *Caenorhabditis elegans*. Or on estime que *C. elegans* et les espèces du genre *Meloidogyne* sont séparées par plus de 200 millions d'années d'évolution (203 Ma entre *Caenorhabditis* et *Meloidogyne*; calculé avec TimeTree (Kumar et al., 2017)). Malgré cela, j'ai néanmoins souhaité tester RepeatMasker en réalisant l'annotation du génome de *M. incognita* (PRJEB8714, (Blanc-Mathieu et al., 2017)). J'ai pu constater que l'utilisation de cet outil, en se basant uniquement sur les séquences recensées dans RepBase, aboutissait à une annotation (très) fragmentée des ETs sur le génome, la plupart des copies annotées étant courtes du fait plusieurs parties d'un même ET étaient annotées à proximité mais pas reliées entre elles. Bien que non surprenant, ce résultat est symptomatique de l'utilisation de séquences référence (ETs) trop divergentes du génome à annoter. J'en ai conclu que l'analyse du contenu en ETs des génomes de *Meloidogyne* nécessitait une étape préalable de prédiction *de* novo afin de constituer une librairie d'ET consensus propre à cette espèce. J'ai donc utilisé RepeatModeler, un outil développé par le même groupe que RepeatMasker, pour créer une librairie de consensus spécifique à *M. incognita* à partir de son génome. J'ai ensuite concaténé cette librairie avec les séquences d'ETs contenues dans RepBase puis j'ai réalisé l'annotation du génome de *M. incognita* à partir de cette librairie hybride. En analysant les sorties de RepeatModeler, j'ai pu constater que peu de séquences consensus créées étaient classées en ordre ou familles. J'ai aussi pu remarquer que les séquences classifiées dans un groupe d'ETs étaient généralement courtes par rapport aux longueurs canoniques attendues pour ce groupe d'ETs. L'annotation du génome à partir de la librairie hybride, était là encore fragmentaire avec beaucoup d'annotations (très) courtes. L'utilisation de RepeatModeler pour venir enrichir la librairie d'ETs n'a donc pas apporté de réelle amélioration. Ainsi, j'en ai conclu que l'utilisation de RepeatMasker, même couplée à une étape préalable de prédiction *de novo* via RepeatModeler, ne me fournirait pas le niveau de détail suffisant pour réaliser une analyse fiable du contenu en ET chez les *Meloidogyne*.

CARP, acronyme pour « Comprehensive Ab initio Repeat Pipeline » a été conçu pour l'annotation et la caractérisation des ETs ainsi que des duplications segmentales (Zeng et al., 2018). Malheureusement étant encore en cours de développement actif à cette période, ce programme avait de nombreux bugs et j'ai rapidement abandonné l'idée de l'utiliser pour réaliser mes analyses.

Mon choix s'est donc porté sur le méta-pipeline REPET (Flutre et al., 2011; Quesneville et al., 2005). REPET est un méta-pipeline regroupant deux composantes (sous-pipelines) : TEdenovo et TEannot; chaque pipeline pouvant être utilisé indépendamment de l'autre. TEdenovo est un pipeline modulaire dédié à la création d'une librairie de séquences consensus à partir de l'analyse d'un génome assemblé et à la classification de ces séquences consensus (voir Figure 4.A.2.1).

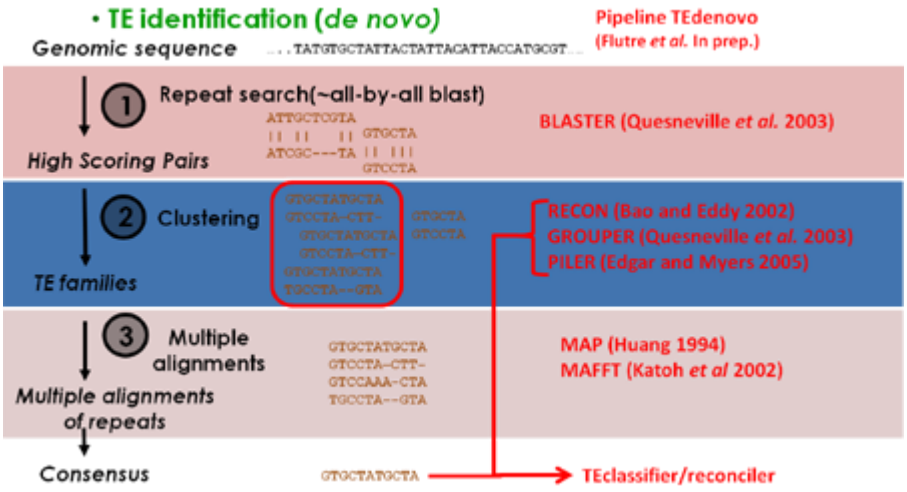


Figure 4.A.2.1: vue générale du pipeline de prédiction de novo TEdenovo (REPET).

Figure issue de https://urgi.versailles.inra.fr/Tools/REPET

TEdenovo commence par réaliser un alignement du génome avec lui-même en utilisant BLASTER (Quesneville et al., 2003) afin d'identifier des paires de séquences similaires (HSP pour « High scoring pairs ») correspondant à des répétitions. Les HSPs sont ensuite regroupées par ensemble de séquences similaires avec RECON (Bao, 2002), GROUPER (Flutre et al., 2011) et PILER (Edgar and Myers, 2005). Pour chaque groupe de séquences, un alignement multiple est réalisé avec MAP (Huang, 1994) ou MAFFT (Katoh, 2002) et une séquence consensus est créée. Cette séquence consensus est assignée/classifiée à un type d'ET par PASTEC (Hoede et al., 2014) . PASTEC classifie une

séquence d'ET en fonction de ses caractéristiques structurelles (présence de longue répétions terminales (LTR), de répétitions terminales inversées (TIR), de queues polyA, de queues de type SSR, etc.), et de ses caractéristiques de « codage » (correspondances avec des ETs connus, des gènes hôtes, des profils ADNr ou HMM, etc.). En fonction des différents indices trouvés, PASTEC classifie les ETs en « classe », « ordre », et parfois en « superfamille » selon la classification de Wicker (Wicker et al., 2007). Les « classe », « ordre », et « superfamille » assignées à une séquence consensus sont encodés dans son nom par un trigramme, la première lettre étant un R pour un rétrotransposon ou un D pour un transposon à ADN. En cas d'indécision ou de caractéristique incomplète, la lettre X est assignée. Par exemple, une séquence consensus dont le trigramme est DTX, correspond à un transposon à ADN de l'ordre des TIR mais dont la superfamille est inconnue. PASTEC permet en outre de caractériser d'autres éléments répétés tels que les SSR (pour « Short Simple Repeats » e.g. courtes répétitions simples), de potentiel gènes dupliqués ou encore de l'ADN ribosomique. A la fin de l'exécution du pipeline, une librairie non redondante de séquences consensus est créée et exportée au format fasta.

TEannot est un pipeline modulaire dédié à l'annotation en ETs d'un génome à partir d'une librairie de séquences consensus d'ET pouvant provenir d'une prédiction *de novo* préliminaire ou bien d'une source extérieure (voir Figure 4.A.2.2).

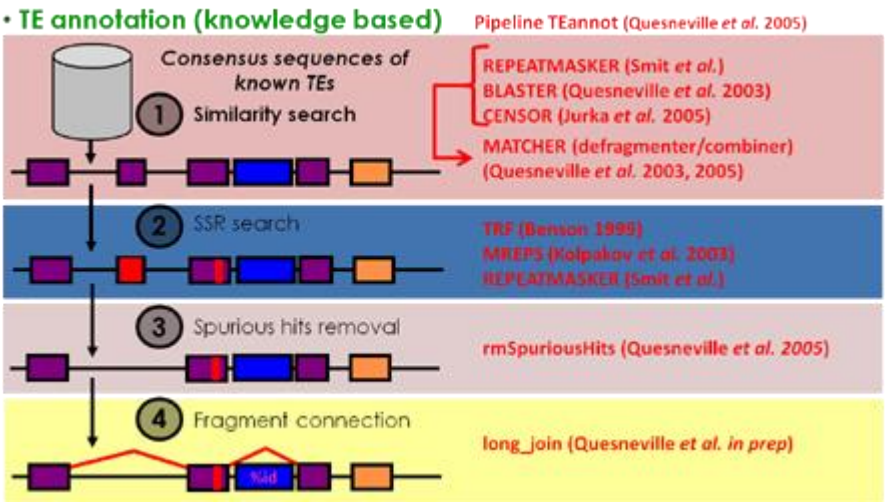


Figure 4.A.2.2: vue générale du pipeline d'annotation TEannot (REPET).

Figure issue de https://urgi.versailles.inra.fr/Tools/REPET

TEannot regroupe divers outils dont BLASTER, RepeatMasker et CENSOR (Jurka et al., 1996) qui lui permettent de réaliser des annotations indépendantes du génome à partir de la librairie de séquences fournie. Un filtre statistique est ensuite appliqué pour écarter les correspondances erronées. Les SSR (pour « Short Simple Repeats » e.g. courtes répétitions simples) sont annotées séparément avec TRF (Benson, 1999), RepeatMasker et MREPS (Kolpakov, 2003). Le programme MATCHER est ensuite exécuté afin d'éliminer des annotations redondantes, mais aussi de défragmenter l'annotation en reconnectant les fragments appartenant à une même copie, et ce même pour des fragments distants (« long join procedure »). Enfin, les annotations sont exportées dans différents formats dont le gff3.

Mise en place d'un protocole de détection *de novo* et d'annotation des ETs avec REPET

La prédiction *de novo* d'une librairie de séquences consensus constitue le point critique de la description du contenu en ET des génomes puisque c'est à partir des séquences contenues dans cette base de données que seront annotées les positions du génome correspondant à des ETs. Le gros du travail a donc consisté à optimiser cette étape de l'analyse. Les phases d'optimisations successives de ce protocole sont le fruit de mon expérience (empirique) personnelle, mais aussi d'une formation réalisée à l'URGI (équipe à la tête du développement et du maintien de REPET) et d'échanges avec plusieurs membres de cette même équipe. A ce titre, je remercie sincérement Hadi Quesneville (https://orcid.org/0000-0003-3001-4908), Véronique Jamilloux et Joëlle Amselem (https://orcid.org/0000-0001-7124-3454) pour leur écoute et leurs conseils avisés.

Le protocole de prédiction *de novo* et d'annotation des ETs qui a découlé de ces évolutions successives est représenté en version simplifiée en Figure 4.A.2.3 et est décrit en détail dans (Koutsovoulos et al., 2019; Kozlowski et al., 2020). Je ne reviendrai donc ici que sur les points clefs ayant mené à sa forme actuelle.

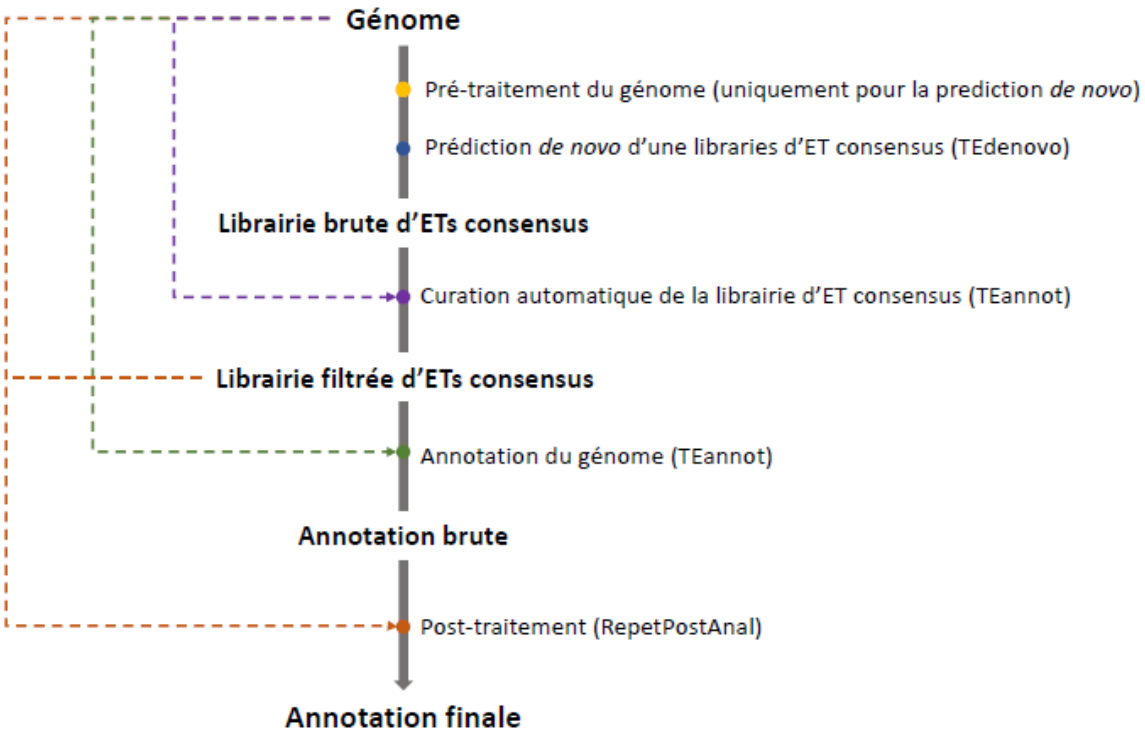


Figure 4.A.2.3 : vue d'ensemble du protocole mis en place de prédiction *de novo* et d'annotation du contenu en ETs à partir d'un génome assemblé.

Les points de couleurs sur la frise chronologique verticale représentent les différentes étapes du protocole. Les flèches de couleurs indiquent les fichiers supplémentaires nécessaires pour réaliser l'étape en question.

Importance de la qualité des données (génome assemblé)

Lors de la prédiction *de novo* d'une librairie de séquences consensus d'ET à partir d'un génome assemblé, la contiguïté des séquences en entrée influe sur la qualité de l'annotation. En effet, les fragments/contigs les plus courts comportent souvent des fragments d'ETs à leurs extrémités et peuvent même être exclusivement composés d'ETs. Le fait de conserver ces séquences peut induire la prédiction de séquences consensus partielles/fragmentaires. Ainsi, il est nécessaire de réaliser la prédiction *de novo* à partir de matériel génétique qui soit le moins fragmentaire possible. Cette exigence peut être partiellement surmontée en supprimant les contigs les plus courts mais pose néanmoins un autre problème. En effet, le fait de supprimer de l'information en entrée a un impact sur le nombre et la diversité des ETs détectés. J'ai donc réalisé plusieurs analyses afin de mesurer l'impact de ces paramètres sur la qualité de la librairie de séquence consensus (mesurée via le nombre, la taille et le

niveau de classification des séquence identifiées). Pour ce faire, j'ai progressivement éliminé les séquences les plus courtes jusqu'à atteindre le L50, i.e. à ne retenir que les séquences dont la longueur (ordonnée) cumulée était égale à 50 % de la taille du génome. J'ai pu noter qu'une réduction aussi drastique du matériel en entrée diminue en conséquence le temps d'exécution de l'outil (ce qui est un point positif) mais aussi la quantité de séquences consensus identifiées ; les consensus identifiés correspondant aux ETs les plus répétés dans le génome. Empiriquement, j'ai donc conclu que le meilleur compromis entre contiguïté et quantité de séquences conservées en entrée se situait entre le L90 et le L99 suivant la qualité des génomes, ce qui correspond à ne conserver que les séquences dont la taille (ordonnée) cumulée est supérieure à 90-99% de la taille du génome.

La présence de certains types de séquences (SSR, répétitions en tandem, microsatellites et régions de faible complexité, etc) dans les génomes est connue pour poser problème lors des alignements de séquences car elles génèrent un score de similarité biaisé. Or la première étape de TEdenovo consiste précisément à réaliser un alignement du génome par rapport à lui-même afin de détecter des couples de séquences hautement similaires. J'ai donc aussi évalué l'impact de ces séquences sur la qualité des ETs consensus produits en masquant ces régions grâce à RepeatMasker (avec l'option -int) puis en les éliminant du génome. Je n'ai pas retenu cette piste car je n'ai pas noté d'impact bénéfique sur la qualité de la librairie de séquences consensus produite. Au contraire, le fait de supprimer ces régions avait pour effet de fragmenter le génome, les plus petites séquences devant ensuite être retirées comme présenté ci-dessus.

Importance de la prise en compte des caractéristiques biologiques de l'organisme.

Les espèces de *Meloidogyne* que j'ai eu l'occasion d'étudier ont pour la plupart des génomes polyploïdes avec un fort pourcentage de divergence entre copies homéologues (6-8% de divergence entre copies pour *M. incognita, M. javanica* et *M. arenaria*; probablement due à un ou plusieurs évènement(s) d'hybridation (Blanc-Mathieu et al., 2017)). Ces espèces possèdent donc « plusieurs génomes en un ». Or la première étape de TEdenovo consiste à rechercher des séquences correspondant potentiellement à des ETs sur la base de leur répétitivité; seules les séquences répétées un nombre de fois donnée pouvant ensuite être considérées pour les étapes suivantes. Compte tenu de la structure de leur génome il a donc été nécessaire pour ces espèces d'augmenter le nombre de répétitions minimales afin de limiter le nombre de faux positifs comme des régions dupliquées (codantes ou non). Pour les génomes polyploïdes avec un fort pourcentage de divergence (> 5%), j'ai placé le nombre de répétitions minimal à 2*p+1 ; p correspondant au niveau de ploïdie de l'organisme.

Mise en place d'une étape de filtrage des séquences consensus d'ETs.

Afin de limiter les séquences faux-positifs ainsi que la redondance dans les librairies d'ETs consensus, j'ai introduit une étape intermédiaire de curation automatique des séquences consensus entre l'étape de prédiction de novo et l'étape d'annotation. Cette curation automatique consiste à effectuer une annotation superficielle du génome grâce à TEannot (étapes 1, 2, 3, 7) à partir de la librairie « brute» précédemment crée par TEdenovo. Seules les séquence consensus ayant été annotés au moins une fois (e.g. au moins une copie) en pleine longueur sont alors retenues pour créer la librairie d'ETs consensus définitive qui servira à réaliser l'annotation finale du génome. L'idée derrière ce choix est que si une séquence consensus ne peut être alignée (annotée) en pleine longueur au moins une fois sur le génome à partir duquel elle a été générée, elle est probablement erronée et doit donc être éliminée.

Mise en place d'un protocole de post traitement adapté à la recherche d'ET « canoniques ».

Bien que l'étape de curation automatique des séquences consensus ait grandement amélioré la qualité globale de l'annotation, un nombre important d'annotations plus dégénérées (courtes et/ou fragmentaires) demeuraient. Ceci vient en partie de la philosophie de REPET dont l'une des forces est qu'il permet de retrouver des ETs anciens et dégénérés faisant partie de la « matière noire » du génome. Mes analyses étant plus concentrées sur la recherche d'ET susceptibles d'avoir été récemment actifs, j'ai dû mettre en place un protocole de filtrage du repeatome brut produit par REPET afin d'en extraire les ET « canoniques ». C'est dans cet objectif que j'ai progressivement développé RepetPostAnal (version 1.0.5 actuellement) (Kozlowski, 2020).

A partir du fichier d'annotation du repeatome, de la librairie de séquences consensus et du génome, RepetPostAnal applique une série de filtres afin de dissocier les ETs « canoniques » et récents de la « matière noire ». En outre, RepetPostAnal génère automatiquement toute une série de statistiques globales sur l'annotation du génome (charge, composition) mais aussi de statistiques spécifiques à chaque séquence consensus d'ET représentée (nombre de copies annotées, identité moyenne des copies, longueur moyenne des copies, etc) et à chaque copie annotée (longueur, proportion de consensus couvert, identité avec le consensus, etc). La plupart des filtres implémentés dans RepetPostAnal sont des modules indépendants que l'on peut choisir d'appliquer ou non et dont la stringence est paramétrable. A titre personnel, je retiens uniquement comme annotation canonique les annotations i) classifiées comme retrotransposons ou transposons à ADN, ii) dont la longueur est supérieure à 250 pb, iii) qui couvrent au moins un tiers de la séquence de l'ET consensus à partir duquel elles ont été annotées, et iv) qui partagent plus de 85% d'identité avec cette même séquence consensus. En outre, je ne retiens que les annotations qui s'alignent en priorité sur leur séquence consensus lors du blast contre la librairie

de séquence consensus. Et enfin, dans le cas ou deux annotations viendraient à être chevauchantes (sur un brin donné), les deux annotations seront supprimées de l'annotation finale.

3 - Résultats

La mise en place de la méthodologie précédemment décrite m'a donné l'opportunité de collaborer avec différents groupes et de la mettre en application sur des organismes variés. Certaines de ces analyses sont actuellement déjà publiées. D'autres sont en cours d'écriture ou en preprint, comme détaillé ci-dessous: Tout d'abord, j'ai réalisé une nouvelle prédiction et annotation du contenu en ET de l'actuel génome de référence de *M. incognita* (PRJEB8714 ; (Blanc-Mathieu et al., 2017)) dans le cadre d'une étude ayant pour objectif de juger de la variabilité du contenu en ET au sein de 12 isolats de cette espèce (voir chapitre VI et (Kozlowski et al., 2020)). Dans cette même étude, j'ai réalisé le même type d'analyse sur le génome du nématode modèle *Caenorhabditis elegans,* ce qui a permis de valider notre méthodologie en comparant nos prédictions avec les données expertes disponibles.

En outre, au sein du genre *Meloidogyne*, j'ai également réalisé la caractérisation du contenu en ET dans les génomes de *M. enterolobii* (Koutsovoulos et al., 2020), une espèce émergente en Europe et non contrôlée par les gènes de résistance habituellement utilisés contre les autres Meloidogyne ; et de *M. graminicola* (Phan et al., 2020), le nématode le plus problématique sur les cultures de riz en Asie.

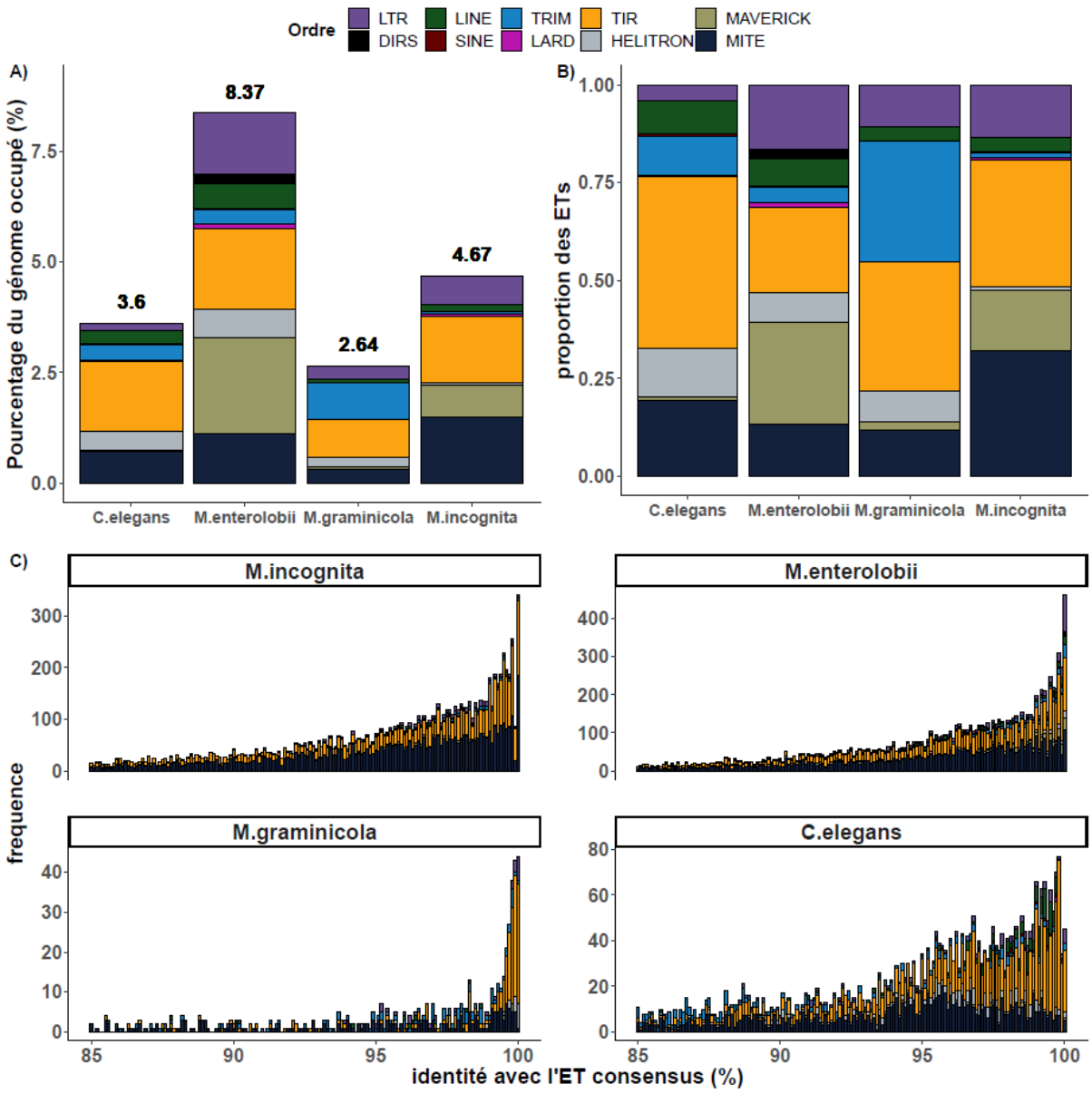


Figure 4.A.3.1 : récapitulatif de la charge et du contenu en ETs de 4 génomes de nématodes.

A - Pourcentage du génome occupé par ordre d'ET et par espèce.

B - Proportion relative des ordres d'ETs par espèce.

C - Age des ETs dans les génomes (fréquences cumulées)

Les estimations de la charge et de la composition en ETs de ces 4 génomes de nématodes sont représentées en Figure 4.A.3.1-A & B. La composition en ET du genome de *M. incognita* est cohérente avec les prédictions précédemment réalisées pour cette espèce (Abad et al., 2008 ; Blanc-Mathieu et al., 2017). De manière générale, on constate que les génomes de ces nématodes sont plus riches en transposons à ADN qu'en rétro-transposons. Pour résumer ces analyses en quelques lignes, on peut noter

que *M. enterolobii* est la seule des espèces de *Meloidogyne* étudiées avec cette méthodologie dans laquelle des rétrotransposons de l'ordre des DIRSs ont été identifiés. En proportion, les *Meloidogyne* possèdent plus de rétrotransposons (en particulier de LTR) que *C. elegans*. Globalement, la charge d'ETs et la proportion relative des ordres d'ETs sont variables entre les 4 espèces. Si l'on compare avec *C. elegans,* mise à part une différence dans la charge en rétrotransposons LTR, il ne semble pas y avoir de patron de charge ou de composition en ETs propre aux espèces de *Meloidogyne*. Néanmoins, il serait nécessaire dans le futur d'inclure plus d'espèces des genres *Caenorhabditis* et *Meloidogyne* afin d'en tirer des conclusions plus globales.

Il est usuel d'utiliser le pourcentage d'identité entre les copies d'ETs et leur séquence consensus comme estimateur de l'âge des ETs dans un génome. Plus les copies d'ETs présentent un pourcentage d'identité élevé vis -à -vis de leur consensus, plus elles sont estimées récentes. Les profils d'âge des ETs dans les génomes de nématodes étudiés sont représentés en Figure 4.A.3.1-C. On peut constater que le profil d'âge des ETs de *C. elegans* diffère de ceux des espèces de *Meloidogyne*. En effet, *C. elegans* présente un profil d'âge dont l'évolution est moins abrupte et qui pourrait indiquer une activité plus régulière des ETs au cours du temps chez ce nématode que chez les *Meloidogyne*. Chez les *Meloidogyne,* la majorité des copies présente une identité importante avec leur consensus, synonyme d'une activité récente dans le génome de ces espèces. Chez *M. graminicola,* en comparaison des deux autres espèces de Meloidogyne, cette tendance est encore exacerbée avec un pic plus marqué de copies fortement similaires à leur consensus et une déplétion des copies ayant des similitudes plus modérées. Cela pourrait indiquer une activité plus récente et soudaine que pour les deux autres espèces. Cependant, le pourcentage du génome occupé par des ETs étant plus faible chez *M. graminicola,* ce profil pourrait plutôt résulter d'une élimination des ETs plus anciens au cours du temps.

En dehors des nématodes du genre *Meloidogyne*, j'ai également annoté les ETs dans les génomes de deux autres ravageurs de cultures: le coléoptère *Anthonomus grandis* (charançon du cotonnier) et l'oomycète à large gamme d'hôtes *Phytophtora parasitica* . Ces résultats n'étant pas encore publiés, je ne donnerai pas plus de détails ici mais la même méthodologie a été utilisée et les résultats de ces analyses seront inclus dans les futurs articles décrivant ces génomes.

Enfin, j'ai aussi eu l'occasion de travailler sur le nouveau génome de référence du « scandale évolutif » *Adineta vaga* (preprint : (Simion et al., 2020)), une espèce de rotifère se reproduisant de manière assexuée depuis des dizaines de millions d'années. La prédiction et l'annotation du contenu en

ET du génome de *A. vaga* ont nécessité des ajustements méthodologiques. En effet, une précédente version du génome (Flot et al., 2013) avait montré que cet organisme contenait relativement peu d'ETs, mais surtout que les ETs représentés, bien que divers, étaient présents en un faible nombre de copies. Or, TEdenovo du meta-pipeline REPET, comme la plupart des outils de prédiction *de novo* reposant sur la répétitivité de séquence pour détecter de potentiels ETs, a tendance à ne pas détecter les ETs présents en un nombre de copies trop faible dans le génome. Afin de contourner cette limitation, nous avons modifié notre protocole de prédiction de-novo. Cette modification a consisté à réaliser une prédiction de-novo indépendante du contenu en ETs de *A. vaga* en utilisant EDTA (Ou et al., 2019) (prédiction réalisée par Alessandro Derzelle ; Université de Namur, Belgique) en plus de celle que j'avais réalisée avec TEdenovo. EDTA (pour « Extensive *De novo* Te Annotator ») est un pipeline regroupant plusieurs programmes tels que LTRharvest ou encore HelitronScanner, chacun étant dédié à la détection d'un type d'ET en particulier (dont principalement les LTRs, TIRs et Helitrons) sans avoir à passer par une phase préliminaire d'identification des répétitions dans le génome. Cet outil est donc théoriquement plus adapté à la détection d'ETs en faible nombre de copies dans les génomes. J'ai ensuite concaténé les librairies prédites indépendamment par EDTA et TEdenovo (REPET), puis j'ai repris le cours normal du protocole en réalisant i) la curation automatique de cette librairie hybride, ii) l'annotation du génome et iii) le post traitement de cette annotation selon la même méthodologie que celle présentée dans les méthodes.

Brièvement, les ETs « canoniques » couvrent 3,3 % du génome d'*A. vaga*. La plupart des ETs sont présents en un faible nombre de copies (< 6 copies en moyenne). A titre de comparaison, chez *M. incognita*, les ETs consensus présentent en moyenne plus de 20 copies. Les transposons à ADN, et en particulier les TIR sont majoritaires dans le génome de *A. vaga* (53% des annotations sont des copies de TIR). L'ensemble de ces résultats est en accord avec ceux précédemment décrits dans (Flot et al., 2013).

4 - Discussions et perspectives

Il existe une grande diversité d'outils permettant la prédiction et l'annotation du contenu en ETs à partir de génomes assemblés et ce nombre ne cesse de croître. Néanmoins, l'outil parfait n'existe pas et il est nécessaire de prendre en considération divers paramètres afin de choisir celui qui sera le plus adapté à l'analyse que l'on souhaite réaliser, et d'interpréter au mieux les résultats obtenus. Avant toute analyse, il est donc nécessaire d'avoir une bonne connaissance des forces et faiblesses de l'outil, des spécificités des organismes étudiés, ou encore de la nature et de la qualité des données, mais aussi d'avoir décidé du niveau de précision souhaité.

Mon expérience personnelle m'a par exemple mené à la conclusion que l'utilisation de RepeatMasker n'est pas adaptée à l'analyse fine des ETs chez des espèces « non-modèles » si celles-ci

sont trop distantes des espèces représentées dans RepBase, et ce, même si cette annotation est couplée à une phase de prédiction *de novo* préliminaire grâce à RepeatModeler. Ceci peut en partie être expliqué par la simplicité de ces deux pipelines qui limite leur efficacité et leur spécificité. Néanmoins, on notera que cette simplicité se révèle aussi être une force puisqu'elle permet en contrepartie de faciliter le déploiement de ces outils et leur utilisation, mais aussi de leur permettre d'être rapides d'exécution. Ce dernier point peut se révéler crucial, en particulier dans le cadre d'analyses de génomes de grande taille ou encore d'analyses de génomique comparative comportant beaucoup de génomes, ce qui explique en partie l'utilisation massive de ces outils.

L'utilisation de REPET m'a permis de caractériser de manière exhaustive le contenu en ETs de plusieurs espèces modèles et non-modèles, en particulier au sein des nématodes du genre *Meloidogyne*. L'analyse du génome de *C. elegans* et la comparaison des résultats obtenus avec les données expertes disponibles dans (Bessereau, 2006) ont permis de valider la méthodologie mise en place (Kozlowski et al., 2020). REPET constitue donc une suite d'outils complète et cohérente, mais comporte aussi son lot de spécificités d'utilisation et de biais qu'il est nécessaire de prendre en considération afin d'interpréter au mieux les résultats obtenus. Irina Arkhipova rapporte par exemple que PASTEC, l'outil de classification de TEannot, a tendance à facilement classifier les ETs en LARDs et TRIMs sur la base de trop peu d'indices (Arkhipova, 2017). Selon elle, une part non négligeable des consensus classés dans ces ordres pourrait être des faux positifs de classification. Il est donc nécessaire de garder en tête ce biais potentiel. J'ai par ailleurs pu noter que TEdenovo était sensible au niveau de ploïdie de l'organismes et à la divergence entre copies homéologues du génome. Il a donc été nécessaire de faire varier pour chaque analyse le seuil de répétitivité requis pour la détection d'éléments répétés afin de limiter la création de séquences consensus « faux positif ». Enfin, j'ai aussi pu constater que TEdenovo était aussi sensible à la fragmentation des génomes, celle-ci se faisant ressentir sur la qualité des séquences consensus créés. Néanmoins, les technologies de séquençage étant de plus en plus performantes et produisant des génomes toujours plus contigus, ce biais technique devrait disparaître à l'avenir. En attendant, on notera aussi qu'un certain nombre de méthodes existent actuellement pour détecter et/ou quantifier les ETs directement à partir des librairies de séquençage plutôt que des génomes assemblés. On peut par exemple citer RepeatExplorer (Novak et al., 2013) dont l'algorithme est centré sur le clustering de reads, RepARK (Koch et al., 2014) qui repose principalement sur l'analyse des K-mer, ou encore dnaPipeTE (Goubert et al., 2015) qui réalise une analyse de la charge et la composition en ETs d'un génome en reconstituant des séquences d'ETs à partir d'un sous-échantillonnage des données. Ce type d'outils constitue donc une alternative possible pour étudier le contenu en ETs chez des organismes pour lesquels aucun génome assemblé n'existe ou si cet assemblage est de mauvaise qualité (voir Chapitre V).

Outre le fait de devoir choisir un outil le plus adapté possible à l'analyse à réaliser, il est aussi important de garder à l'esprit que l'utilisation de cet outil nécessitera parfois certains ajustements méthodologiques supplémentaires.

Par exemple, REPET a en partie été développé pour fouiller les génomes à la recherche de la « matiere noire », c'est-à-dire des ETs très dégénérés, vestiges d'une activité ancienne dans les génomes (Hadi Quesneville). Mes analyses étant plus concentrées sur la recherche d'ETs susceptibles d'avoir été récemment actifs, j'ai dû mettre en place un protocole de filtrage du repeatome pour dissocier les ETs « canoniques » et récents de la « matière noire ».

Par ailleurs, REPET, comme la plupart des outils reposant sur la répétitivité des séquences lors des phases de prédiction de-novo, a tendance à ne pas détecter les ETs présents dans le génome en un faible nombre de copies alors que ces copies d'ETs pourraient par exemple être détectées via des signatures structurales (Hoen and Bureau, 2015). C'est dans cette optique que lors de l'analyse du génome d'*A. vaga*, nous avons couplé les prédictions *de novo* d'EDTA à celle de TEdenovo ce qui a permis de constituer une librairie de séquence plus complète (Simion et al., 2020). L'annotation du génome à partir de cette librairie étant cohérente avec les résultats manuellement curés précédemment décrits pour cette espèce (Flot et al., 2013), ce type d'approche constitue donc une piste méthodologique intéressante qu'il faudrait évaluer plus en détails.

Il existe une grande diversité d'outils permettant de décrire le contenu en ETs d'un organisme et ce nombre ne cesse de croître. Cette diversité, bien que bénéfique en soit, n'en demeure pas moins une source d'incertitude lorsqu'une méthode doit être choisie ; chaque outil ayant ses forces, ses faiblesses, et ses spécificités dont l'identification reste parfois complexe. C'est pour cette raison que Hoen et collaborateurs, (auxquels je me joins) appellent à établir d'urgence une base de comparaison fiable et réaliste permettant de tester et comparer de manière objective les différents outils disponibles (Hoen et al., 2015). Une telle base de comparaison permettrait en outre d'effectuer une réévaluation des résultats obtenus avec des méthodes différentes, facilitant ainsi la réalisation de méta-analyses tout en réduisant les biais d'interprétation du type « c'est l'organisme que j'étudie qui a la plus grosse (proportion de génome occupée par des ETs) ».

B – Estimation de la variabilité des paysages d'ETs dans les génomes à l'échelle populationnelle

1 - Contexte et objectifs

La prédiction et/ou l'annotation du contenu en ET d'un génome, via les méthodes précédemment présentées par exemple (RepeatMasker, REPET, etc), permet de faire l'état des lieux de la charge et de la composition en ETs de ce génome à un instant "t" et pour une population / lignée / isolat donnée. Bien que très informative en soit, ce type d'analyse ne donne qu'une idée partielle des ETs actuellement / récemment actifs dans ledit génome. En effet, s'il est possible (et usuel) d'utiliser la divergence entre les copies d'ETs et leur séquence consensus comme estimateur de l'âge des ETs dans un génome, cette méthode n'en donne qu'une évaluation globale et ne permet pas de détecter des mouvement d'ETs, élément fondamental de l'activité des ETs. Ainsi, si l'on souhaite évaluer dans le détail si les ETs jouent/ont récemment joué un rôle dans la plasticité génomique d'un organisme et évaluer dans quelles proportions, il est nécessaire d'entreprendre d'autres types d'approches.

L'activité de transposition génère des insertions/délétions d'ETs qui diffèrent d'un individu/population à l'autre. La détection de ces loci d'ET polymorphes nécessite l'utilisation d'un autre type d'outils qui reposent sur l'analyse de données de (re)séquençage.

Généralement, en plus de données de (re)séquençage, ces outils requièrent deux types d'informations : un génome de référence et de l'information sur le contenu en ET du génome (séquences consensus d'ETs et/ou la position des ETs dans le génome e.g. un fichier d'annotation). Dans un premier temps, les données de séquençage sont alignées sur le génome de référence (étape de « mapping »). L'emplacement, la quantité de reads alignés à cette position (RD pour « reads depth »), ainsi que la nature de l'alignement sont ensuite analysés afin de prédire et de caractériser des positions (loci) d'ETs polymorphes (présence/absence, estimation de la fréquence, sens d'insertion, etc). La détection des sites polymorphes repose principalement sur deux types d'approches algorithmiques : l'analyse des reads discordants (DP pour « discordant pairs »), l'analyse des reads tronqués (SR pour « Split Reads ») (voir Figure 4.A.1.1); ces deux approches pouvant parfois être combinées pour certains outils.

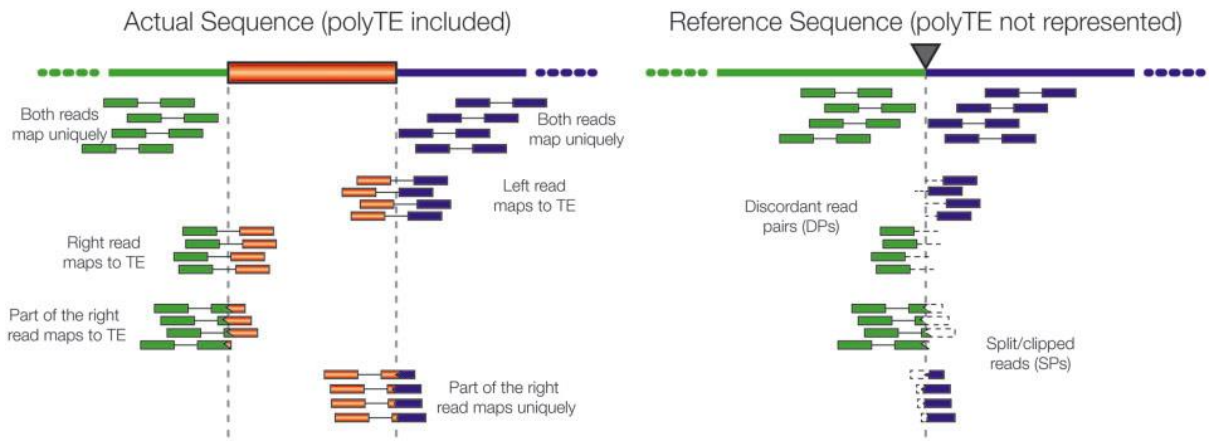


Figure 4.A.1.1 : détection de la présence d'ET via des approches DP et SR.

Figure issue de (Rishishwar et al., 2016) (Figure 1)

Les deux schémas présentés illustrent comment les informations issues de l'alignement de reads pairés sont utilisées pour la détection des sites d'insertion d'ETs polymorphess. Le schéma de gauche représente le cas où un ET (en orange) est inséré (présent) à une position donnée tandis que le schéma de gauche représente le cas où l'ET est absent à cette position (séquence de référence). La manière dont les données de séquençage d'une population donnée s'alignent sur la séquence de référence révèle la présence ou non d'un ET pour cette population. Il existe 3 classes d'alignement : (1) les deux fragments d'une même paire s'alignent sur le génome de manière unique, (2) l'alignement des fragments est discordant (DP pour "discordant pair") : un fragment s'aligne sur le génome et l'autre s'aligne sur un ET, et (3) Une partie d'un des fragments s'aligne sur le génome et l'autre sur l'ET (SP pour "split read"). La présence de DP et de SR, ainsi que la distance de correspondance entre leurs lectures appariées, est utilisée dans la prédiction des sites polymorphes d'insertion d'ETs.

De nombreux outils de détection de polymorphismes d'ET à partir de données de séquençage existent (voir Tableau 1 (Ewing, 2015) et Tableau 1 (Rishishwar et al., 2016) pour un listing non exhaustif, voir (Rishishwar et al., 2016) pour une comparaison de plusieurs outils). J'ai décidé de concentrer mon attention sur l'évaluation de deux outils recouvrant les deux types d'approches algorithmiques : TEPID (approche DP + SR) (Stuart et al., 2016) et PopoolationTE2 (approche DP) (Kofler et al., 2016). Les algorithmes de TEPID et PopoolationTE2 sont schématisés et décrits en Figures 4.B.1.2 & 4.B.1.3 respectivement. Deux critères m'ont poussé à utiliser ces outils plutôt que ceux décrits et comparés dans (Rishishwar et al., 2016). Le premier concerne leurs approches algorithmiques générales. TEPID réalise à la fois des approches DP et SR ce qui intuitivement pourrait accroitre la précision et l'efficacité de la détection. PopoolationTE2 introduit quant à lui le concept de « couverture en information physique » (« physical coverage ») à chaque position du génome. Pour chaque nucléotide sont ainsi encodées les informations sur i) les reads « concordant » (en opposition à discordant) couvrant un site - ce qui indique l'absence d'insertion d'un ET, ii) les reads soutenant une insertion d'ET et iii) les reads soutenant des réarrangements structurels. L'utilisation combinée de ces informations permet de calculer des fréquences de présence d'un ET dans une population donnée plutôt

que de décrire une insertion/deletion d'ETs par rapport à une référence. Le deuxième critère, plus objectif, concerne l'adéquation entre les analyses réalisées avec les outils et celles que je souhaitais faire chez *M. incognita* (voir chapitre VI).

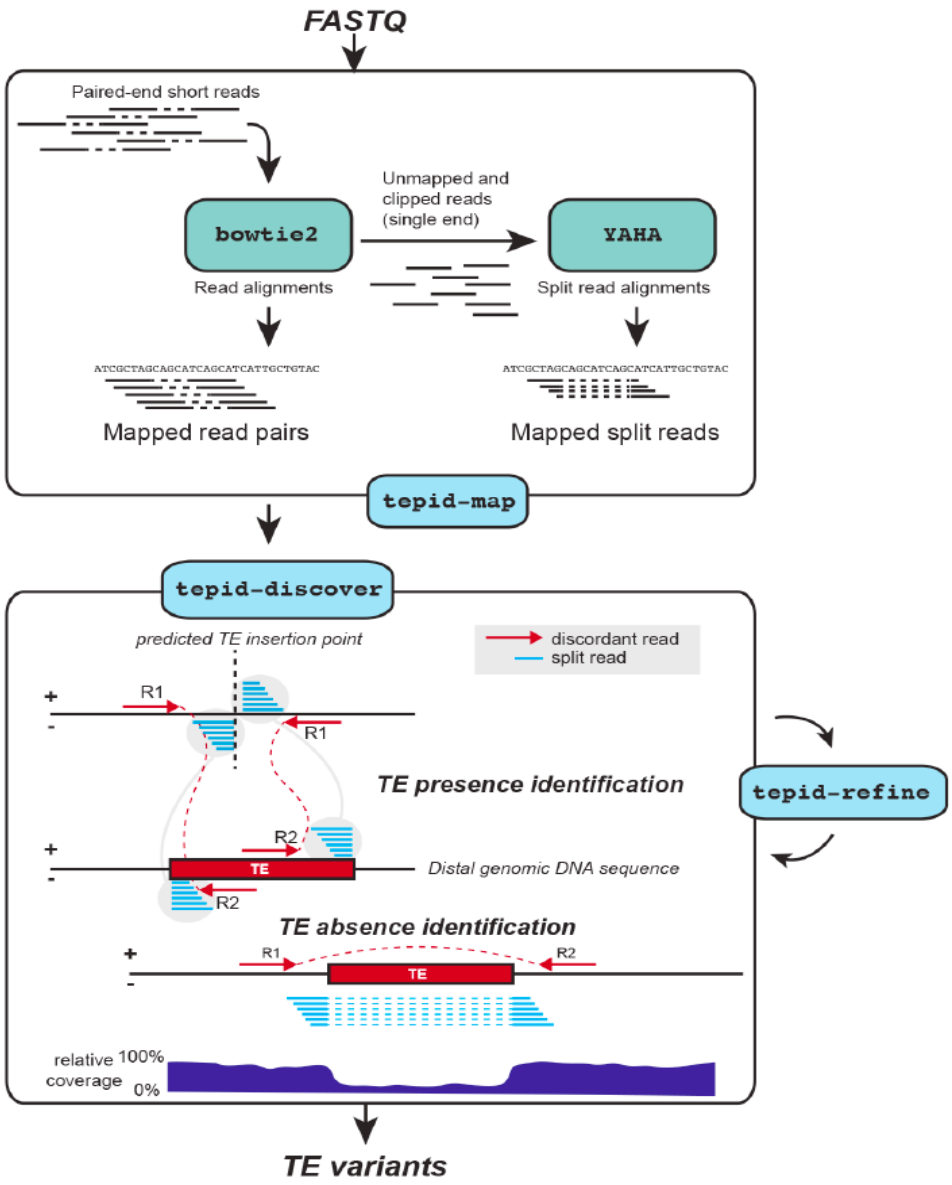


Figure 4.B.1.2 : présentation du pipeline d'analyse TEPID

Figure issue de (Stuart et al., 2016) (Figure 1)

Principe de la découverte de variants d'ETs en utilisant des informations d'alignement de lecture SR et DP. Les reads pairés sont d'abord alignés sur le génome de référence en utilisant Bowtie2. Les reads non-alignés ou dont l'alignement est incomplet sont ensuite extraits et réalignés à l'aide de Yaha, un aligneur spécialisé pour les reads SR. L'ensemble des reads alignés sont ensuite utilisés par TEPID pour découvrir des variants d'ET (polymorphismes d'insertion) par rapport au génome de référence, dans l'étape "tepid-discover". Lors de l'analyse de groupes d'échantillons apparentés, cette analyse peut être affinée en utilisant l'étape "tepid-refine", qui examine plus en détail les régions génomiques dans lesquelles un variant d'ET a été identifié dans un autre échantillon, et tente d'identifier le même variant pour l'échantillon en question en réduisant le nombre de lecture nécessaire à la détection par rapport à l'étape "tepid-discover". Ceci permet de réduire le nombre de faux négatifs dans un groupe d'échantillons apparentés.

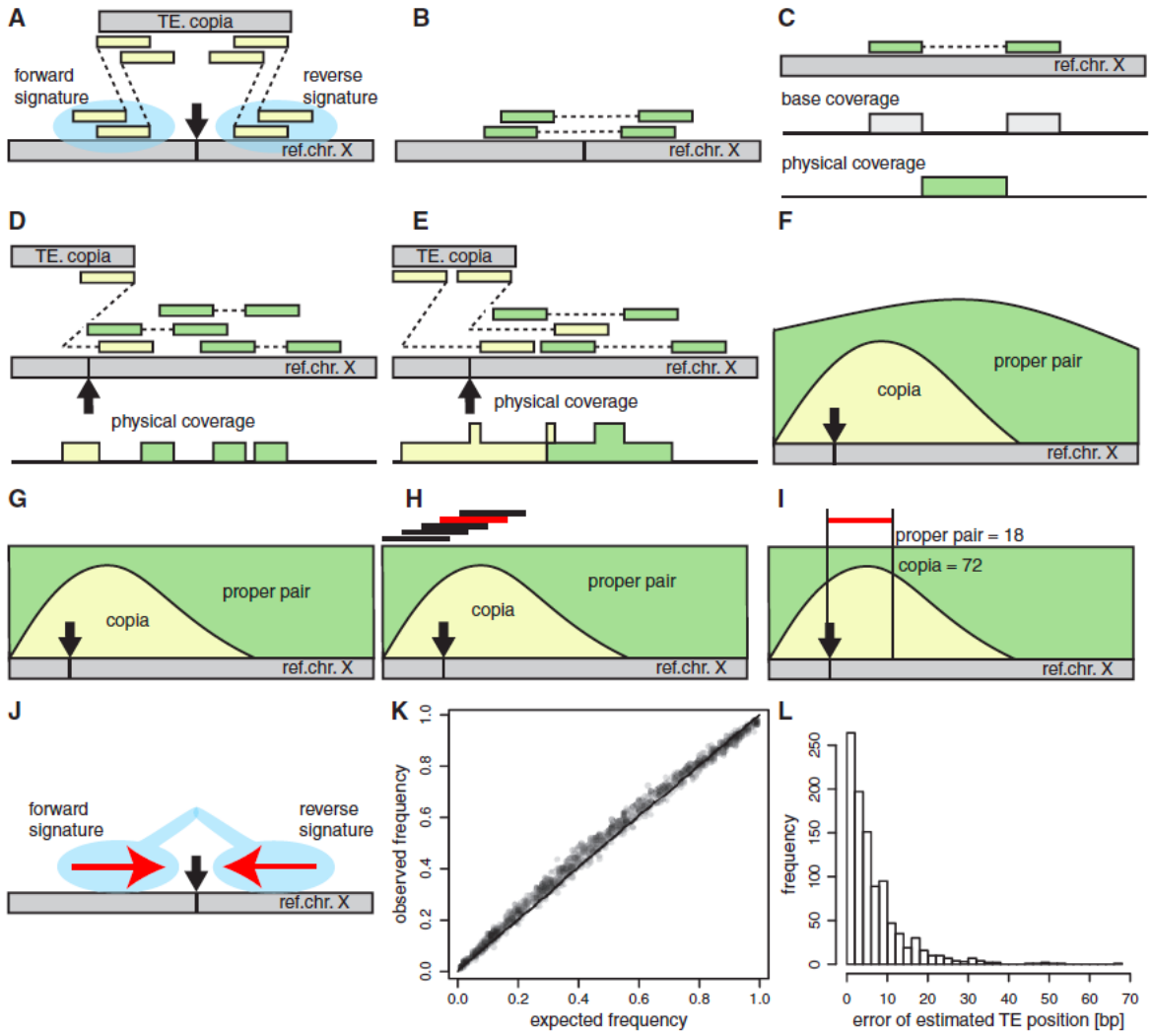


Figure 4.B.1.3 : présentation du pipeline d'analyse popoolationTE2

Figure issue de (Kofler et al., 2016) (Figure 1)

A - L'insertion d' un ET (flèche noire) fait qu'un fragment du read (jaune) est aligné sur le chromosome de référence (X) alors que l'autre fragment s'aligne sur l'ET (copia). Un groupe de ces reads discordants se situe à gauche du point d'insertion (signature "sens" ou F; e.g. "forward signature") et un autre se situe à droite du point d'insertion (signature "antisens" ou R; e.g. "reverse signature").

B - L'absence d'insertions d'ET donne lieu à un alignement cohérent des reads pairés de part et d'autre du site d'insertion supposé (vert).

C - L'information sur la nature de l'alignement des reads est utilisé pour calculer la couverture de reads (gris) et la couverture physique (vert) à une position donnée. Pour la couverture de base, la position des reads est prise en compte alors que pour la couverture physique, c'est la région entre les reads (insert) qui est prise en compte.

D - L'insertion d'un ET induit l'alignement de reads discordants qui peut etre traduit en un autre type de couverture physique. La distance médiane entre les reads alignés de manière cohérente est utilisée pour estimer la distance entre les fragments de ces reads discordants.

E - L'augmentation de la distance interne entre les reads par rapport au panneau D se traduit par un plus grand nombre de lectures supportant une insertion d'ET (copia) et une couverture physique plus élevée. Si les reads discordants et cohérents se chevauchent, la couverture physique des reads est additionnée, ce qui contribue à la

hauteur totale de la courbe de couverture physique. La couverture physique soutenant la présence (jaune) et l'absence (vert) d'un ET peut se chevaucher (région centrale).

F - En combinant les informations de l'ensemble des reads pairés pour chaque position génomique, on obtient une piste de couverture physique.

G - Pour homogénéiser le pouvoir d'identification des ETs, la couverture physique est échantillonnée de façon aléatoire à des niveaux équivalents pour chaque position génomique.

H - La position des signatures des insertions d'ETs est déterminée en utilisant une approche de fenêtre glissante (lignes noires en haut). La fenêtre avec la couverture physique maximale supportant un ET (la ligne rouge indique la fenêtre avec la couverture de copia la plus élevée) est conservée pour une analyse plus approfondie.

I - La fréquence de la signature d'insertion de l'ET dans la population est estimée à partir du rapport entre la couverture physique moyenne soutenant la présence d'un ET et la couverture physique totale dans une fenêtre (copia¼ 72=ð72 þ 18Þ ¼ 0:8).

J - Les paires de signatures d'insertion d'un même ET (F et R) situées à une distance inférieure à une valeur donnée sont jointes, ce qui aboutit à la création d'un set d'insertions d'ETs. Les estimations de la fréquence et de la position de l'insertion pour une population sont obtenues en faisant la moyenne des estimations réalisées pour les signatures F et R.

K - Précision des estimations de la fréquence populationnelle des ETs pour 1 000 ET à partir de données simulées (pool-seq). PoPoolationTE2 présente un léger biais à la hausse pour les ETs présents en fréquence intermédiaire et un léger biais à la baisse pour les ETs présents en haute fréquence.

L - Précision de l'estimations de la position d'insertion pour 1 000 ETs à partir de données simulées (pool-seq).

Dans cette analyse, j'ai simulé des données « jouets » (génomes de références et librairies de données de séquençage). J'ai ensuite utilisé ces données afin de comparer deux outils concurrents de détection de polymorphismes (installation, prise en main, limitations, sensibilité, spécificité). Enfin, après avoir sélectionné un outil sur la base de ces comparaisons, j'ai effectué d'autres séries de tests afin d'évaluer plus en détail les spécificités et limites de l'outil retenu.

2 - Matériel

J'ai simulé des données génomiques (génomes et librairies de données de séquençage) correspondant à des paysages d'ETs à partir d'un fragment de 100 kb du chromosome 2 de la drosophile d'ADN vierge d'ET (châssis) sur lequel j'ai inséré des ETs issus d'une bibliothèque composée de 123 séquences identifiées par SanMiguel et collaborateurs sur le locus Adh-1 du génome de la drosophile (SanMiguel et al., 1996). Ces deux fichiers ont été téléchargés depuis le site suivant: (https://sourceforge.net/p/simulates/wiki/TheClassic_SanMiguel_TELandscape/).

3 - Méthodes

Définition du paysage d'ET

J'ai généré des données pour deux sets de paysages d'ETs, chaque set décrivant en réalité le paysage d'ETs de 3 génomes haploïdes correspondant chacun à 3 populations homogènes (A, B, C pour le premier set et D, E, F pour le deuxième set).

Pour chacun des 2 sets, j'ai généré un fichier au format dédié (pgd) en utilisant le script define-landscape_template.py de la suite simulaTE v-1.10.04 (Kofler, 2018). Ce script prend en entrée les fichiers fasta du châssis et de la librairie d'ET précédemment décrits. Pour chaque set, j'ai spécifié que je souhaitais générer 3 génomes haploïdes (-N 3) sur lesquels je souhaitais répartir 10 points d'insertion vide. Pour chacun des deux fichiers générés, j'ai ensuite manuellement rempli pour chaque locus vide généré i) la nature de l'ET à insérer (e.g. quel ET), et ii) dans quelle population insérer l'ET à cette position. La distribution et la composition en ET des 2 sets sont représentés graphiquement en Figures 4.B.4.1 & 4.B.4.2. Le set 1 (populations A, B, et C) représente un paysage d'ET de forte complexité en terme de diversité d'ETs présents. Le set 2 (populations D, E, et F) représente un paysage de faible complexité.

Création de séquences génomiques

A partir du châssis, de la librairie d'ETs et du fichier .pdg précédemment décrits, j'ai généré pour chaque set les 3 génomes haploïdes en utilisant le script build-population-genome.py de la suite simulaTE.

Evaluation de la position des ETs dans les génomes créés

Le paysage des ETs varie entre les 3 génomes d'un même set, que ce soit en terme de composition (tailles variables des ETs) ou en terme de nombre d'ETs insérés. Ainsi, bien que le châssis de départ soit le même pour chaque génome, la position finale des ETs insérés est relative à chaque génome créé. De ce fait, pour chaque génome généré, j'ai réalisé un blastn (evalue = 0, min identity 100%) contre la librairie d'ETs afin de connaître les positions de début et de fin de chacun des ETs insérés.

Cette étape n'est nécessaire que pour pouvoir évaluer la précision avec laquelle chaque outil testé détecte les ETs sur chacun des génomes si l'on vient à changer de génome de référence.

Simulations de données de séquençages (reads) de population homogène

A partir de chaque génome, j'ai simulé des données de séquençage correspondant à une population homogènes grâce à read_pool-seq_illumina-PE.py de la suite SimulaTE. A l'intérieur d'une population, 100 % des individus présentent strictement le même paysage d'ET. Afin d'être le plus proche possible des caractéristiques des données réelles à venir, j'ai généré ~80X de reads pairés de 145 pb pour chaque génome, avec une taille d'insert de 400bp (+-94) (--read-length 145 --inner-distance 400 --std-dev 94). Seul un taux d'erreur de 2%, similaire à celui rencontré avec la technologie Illumina, a été introduit par rapport aux 6 séquences génomiques A, B, C, D, E, et F (--error-rate 0.02)

Simulations de données de séquençages (reads) de population hétérogène

J'ai créé une population hétérogène en combinant les données de séquençage des 3 populations homogènes A, B et C du premier set de simulation. La population hétérogène ainsi formée est en fait composée de 3 « individus ». Les fréquences de présence d'ETs attendues sont donc de 33%, 66% ou 100% en fonction des cas de figures. Le paysage d'ETs de cette population hétérogène est détaillé en Figure 4.B.4.3.

Détection des polymorphismes

J'ai testé deux outils de détection de polymorphisme d'ETs à partir de données populationnelles de séquençage : TEPID et PopoolationTE2.

Détection de polymorphisme d'ETs avec TEPID

J'ai réalisé l'analyse du premier set de données simulées avec TEPID en suivant le protocole décrit dans le manuel d'utilisation (https://github.com/ListerLab/TEPID). Brièvement : j'ai aligné indépendamment les données de séquençage de chaque population A, B, et C sur le génome de référence A en utilisant le script tepid-map en mode « pairé » (-s 400). Tepid-map automatise l'utilisation de bowtie2 (Langmead and Salzberg, 2012) et de yaha (Faust and Hall, 2012) afin d'identifier respectivement les reads discordant et les reads « splités ». J'ai ensuite utilisé tepid-discover afin d'identifier les sites d'ET polymorphes (insertions et délétions par rapport au génome de référence) à partir des fichiers d'alignement précédemment générés et du fichier d'annotation de référence des ETs du génome A.

J'ai analysé les trois sets de données simulées (les deux sets correspondants à des populations homogènes et le set correspondant à une population hétérogène) avec popoolationTE2 en suivant le protocole « joint analysis » décrit dans le manuel d'utilisation (https://sourceforge.net/p/popoolation-te2/wiki/Manual/).

Pour pouvoir réaliser une analyse, popoolationTE2 requiert au préalable de créer 2 fichiers. Le premier est un fichier de séquence dans lequel sont concaténés i) le génome de référence (masqué pour les régions correspondant aux annotations d'ETs), ii) les séquences correspondant aux annotations d'ETs et iii) la librairie de séquences consensus (123 séquences ; optionnel). Le deuxième fichier (« TE-hierarchy ») est un fichier tabulaire dans lequel sont renseignés un identifiant unique, l'identifiant de la séquence consensus, et l'ordre de la séquence consensus pour chaque séquence de copie d'ET renseignée dans le précédent fichier. J'ai généré les fichiers pour chaque set en utilisant le génome A et son annotation en ETs comme référence pour l'analyse des sets de données 1 et 3 et j'ai utilisé le génome D et son annotation en ETs comme référence pour l'analyse du set de données 2.

Brièvement : pour chaque set de simulation, j'ai aligné indépendamment pour chaque population les fragments gauches et droits de chaque read sur le génome de référence en utilisant bwa bwasw v-0.7.17-r1188 (Li and Durbin, 2010) (paramètres par défaut). J'ai ensuite restauré l'information de pairage des reads pour chaque librairie en utilisant l'outil « sep2pe » de popoolationTE2 (--sort). J'ai ensuite identifié les signatures de polymorphismes avec l'outil « identifySignature » de popoolationTE2 (--mode joint; --min-count 2; --signature-window minimumSampleMedian; --min-valley minimumSampleMedian). Puis, pour chaque site identifié, j'ai estimé la fréquence des différents ETs dans les populations à l'aide de l'outil « frequency » (paramètres par défaut). Enfin, pour chaque site, j'ai mis en relation les signatures de polymorphismes avec l'outil « pairupSignatures » (--min-distance -200; --max-distance -- 300 comme recommandé par R. Kofler), aboutissant ainsi à la liste des polymorphismes d'ET, de leur position et de leur fréquence pour chaque population analysée.

PopoolationTE2 surpasse TEPID

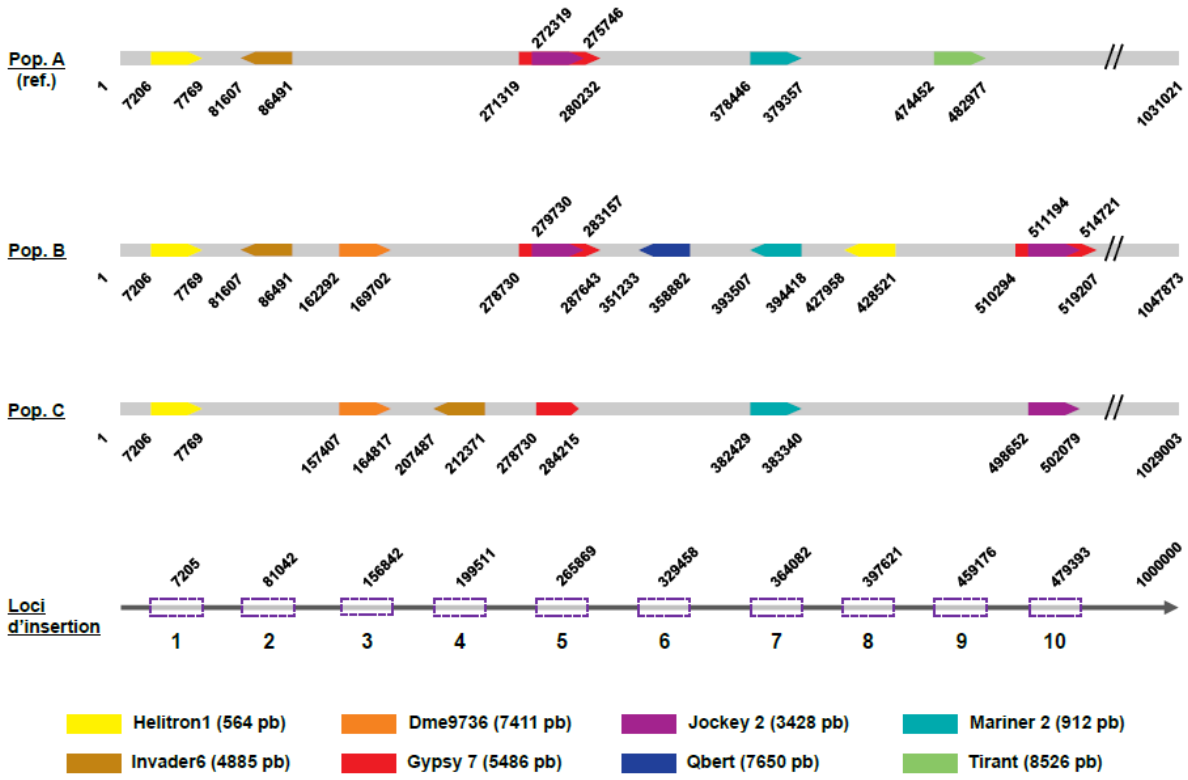


Figure 4.B.4.1 : paysage complexe d'ETs simulé pour 3 populations homogènes.

Les points d'insertions des ETs sur le châssis sont représentés sur la frise du bas par des rectangles en pointillés (10 points d'insertion au total). Les valeurs données au-dessus de chaque locus correspondent à la position de ces loci d'insertion sur le châssis vide. Les 3 frises supérieurs représentent le paysage d'ETs dans le génome de chaque population (A, B, et C). Chaque ET est représenté par une flèche de couleur, le sens de la flèche indiquant le sens d'insertion. La nature des ETs, ainsi que leur taille sont décrites en légende. Les valeurs en dessous de chaque frise correspondent aux positions de début et de fin de chaque ET sur le génome de la population en question.

Chaque population (A, B, et C) est homogène, i.e. 100% des individus à l'intérieur d'une population présentent strictement le même paysage d'ET.

Dans cette première série de simulation, j'ai souhaité tester la capacité des deux outils sélectionnés à identifier des insertions et délétions d'ETs dans le cadre d'un paysage complexe d'ETs présentant une diversité importante tant par le type d'ETs rencontrés que par leurs tailles (voir Figure 4.B.4.1). Nous nous concentrerons ici sur la capacité des deux outils à correctement détecter la présence/absence des différents ETs dans les populations et à la précision de l'estimation de leurs positions.

PopoolationTE calcule la fréquence de présence d'un ET dans une population, c'est-à-dire le pourcentage de la population pour lequel l'ET est présent à cette position. Pour chaque locus détecté, toutes les populations ont une valeur de fréquence associée. Les notions de néo-insertion ou de délétions nécessitent donc de choisir une population qui servira de référentiel. J'ai choisi d'utiliser la population A comme référentiel. Les populations simulées ici (A, B et C) étant homogènes, l'ensemble des individus d'une population donnée partagent donc strictement le même paysage d'ETs. Ainsi, la fréquence d'un ET donné devrait donc être de 100 % s'il est présent dans la population et de 0 % s'il est absent. Dans les faits, les fréquences estimées dans cette analyse sont toutes > 88% ou < 0.05%. J'ai donc utilisé ces valeurs comme seuil haut et bas afin de binariser les fréquences en « présence » et « absence ». Par ailleurs cette première analyse m'a montré que seul les loci dont des signatures d'insertion étaient détectés aux deux extrémités de l'ET (i.e. signatures « FR ») devaient être prises en compte, les détections présentant uniquement une signature gauche ou droite correspondent très majoritairement à du bruit de fond.

Tableau 4.B.4.1 : nature et position des ETs détectés dans des population homogènes par PopoolationTE2 (paysage d'ET « complexe »)

Chaque sous-tableau renferme les résultats de détection pour une population donnée (A, B et C). Les paysages d'ETs (nature et position des ETs) de chaque population sont détaillés en Figure 4.B.4.1.

PopoolationTE2 calcule la position d'un ET par rapport au génome de référence, qui ici est celui de la population A. Lorsqu'un locus d'ET est décrit dans le génome de référence, la position donnée (attendue et observée) correspond au point central de la copie sur le génome (e.g (position de fin + position de début) /2), qu'il s'agisse de la population de référence ou bien d'une autre population. Lorsqu'un ET n'est pas présent dans le génome de référence, la position de cette copie correspond à la position du point d'insertion (frise du bas) à laquelle on ajoute la somme des longueurs des copies d'ETs présentes en amont sur le génome de référence.

Les tirets dans les colonnes « attendu » indiquent qu'aucun ET n'est présent à cette position pour cette population. Les tirets dans les colonnes « observé » indiquent que la fréquence calculée à cette position pour cette population a conduit à le considérer l'ET comme absent (absent si fréquence < 0.05%).

Les distances calculées (colonne distance) correspondent a la valeur absolue de la différence entre la position attendue et la position observée.

Les ET nommés Gypsy7[Jockey2] indiquent qu'un élément Jockey2 est imbriqué dans un élément Gypsy7 (« nested TE »).

	popA				
	ET		**position**		
locus	**attendu**	**observé**	**attendue**	**observée**	**distance (pb)**
1	Helitron1	Helitron1	7488	7488	0
2	Invader6	Invader6	84049	84050	1
3	-	-	-	-	-
4	-	-	-	-	-
5	Gypsy7[Jockey2]	Gypsy7	275776	275778	2
6	-	-	-	-	-
7	Mariner2	Mariner2	378902	378898	4
8	-	-	-	-	-
9	Tirant	Tirant	478715	478693	22
10	-	-	-	-	-

	popB				
	ET		position		
locus	**attendu**	**observé**	**attendue**	**observée**	**distance (pb)**
1	Helitron1	Helitron1	7488	7488	0
2	Invader6	Invader6	84049	84050	1
3	Dme9736	Dme9736	162291	162297	6
4	-	-	-	-	-
5	Gypsy7[Jockey2]	Gypsy7	275776	275778	2
6	Qbert	Qbert	343821	343817	5
7	Mariner2	Mariner2	378902	378898	4
8	Helitron1	Helitron1	412896	412909	13
9	-	-	-	-	-
10	Gypsy7[Jockey2]	Gypsy7	503194	503185	9

	popC				
	ET		position		
locus	**attendu**	**observé**	**attendue**	**observée**	**distance (pb)**
1	Helitron1	Helitron1	7488	7488	0
2	-	-	-	-	-
3	Dme9736	Dme9736	162291	162297	6
4	Invader6	Invader6	204990	204971	19
5	Gypsy7	Gypsy7	275776	275778	2
6	-	-	-	-	-
7	Mariner2	Mariner2	378902	378898	4
8	-	-	-	-	-
9	-	-	-	-	-
10	Jockey2	Jockey2	503194	503196	2

Comme détaillé dans le Tableau 1, l'ensemble des insertions par rapport au génome de référence ont correctement été détectées, même pour les ETs n'étant pas présents dans le génome de référence (population A), c'est-à-dire Dme9736 (locus 3, populations B et C) et Qbert (locus 6, population B). De même, l'ensemble des délétions par rapport au génome de référence ont correctement été détectées (loci 2 et 9). Seuls les ETs « imbriqués » (Jockey2 imbriqué dans Gypsy7) n'ont pas correctement été prédits puisque seul Gypsy7 est détecté ; et ce que l'ET imbriqué soit présent dans le génome de référence (locus5) ou non (locus 10). Ce résultat est néanmoins attendu car ce comportement est décrit dans la documentation de l'outil. On notera que les ETs « simples » présents à ces loci ont quant à eux correctement été détectés (Gypsy7, locus 5, population C ; Jockey2, locus 10, population C). L'inversion du sens d'insertion de Mariner2 dans la population B au locus 7 par rapport aux autres populations n'a pas posé de problème quant à sa détection. Pour cette simulation, popoolationTE2 présente donc une efficacité de 90 % (27/30 ET correctement détectés) si l'on considère la détection de Gypsy7 au lieu de Gypsy7[jockey2] comme un faux négatif ou une efficacité de 100% le cas contraire.

En ce qui concerne la précision de détection des éléments j'ai pu calculer que pour cette simulation, popoolationTE2 présente une variabilité moyenne de 5.37 pb (+/- 1.44 pb) entre les positions détectées et les positions attendues. Ce résultat est en accord avec les valeurs données dans la documentation de l'outil (~10 pb).

A l'inverse de popoolationTE2, TEPID prend comme référentiel absolu le génome de référence, par rapport auquel sont détectés les insertions ou délétions d'ET dans une population donnée. Ainsi, seules les variations concernant des ETs présents dans le génome de référence pourront être détectées dans les autres populations (B et C). La délétion de Tirant est correctement détectée pour les populations B et C au locus 9. De même, la délétion d'Invader6 pour la population C au locus 6 ainsi que la délétion de Jockey2 au locus 5 pour la population C sont correctement détectées. En revanche TEPID prédit que Mariner2 est absent de la population B au locus 7, ce qui correspond à un faux négatif. En ce qui concerne les insertions, seule l'insertion d'Helitron1 au locus 8 dans la population B est détectée. Ceci est dû au fait que les éléments Dme9736 (locus 3) et Qbert (locus6) ne sont pas présents dans le génome de référence. Néanmoins, même dans le cas où l'ET est bien présent dans le génome de référence, il arrive que son insertion dans une autre population à un autre locus ne soit pas détectée. C'est par exemple le cas d'invader6 au locus 4 qui n'est pas détecté dans la population C alors que cet ET est présent au locus 2 dans le génome de référence.

En ce qui concerne l'estimation des positions, seule l'insertion d'Helitron1 au locus 8 dans la population B peut être jugée puisque les délétions détectées concernent des positions décrites dans l'annotation de référence. Au locus 8, Helitron1 est prédit pour être inséré entre les positions 412824 et 412969 par rapport au génome de référence. Ainsi, si la position d'insertion observée correspond bien à celle attendue (point central d'insertion attendu : 412896 ; point central d'insertion prédit : 412897), la taille de l'élément inséré est quant à elle très différente de l'attendu (observé : 145 pb ; attendu : 564 pb).

L'efficacité de détection de TEPID est inférieure à celle de popoolationTE2. J'ai décidé de concentrer les prochains tests sur l'analyse des limitations de popoolationTE2.

PopoolationTE2 détecte efficacement les polymorphismes même en cas de paysage très faible en diversité.

Dans cette seconde série de simulation, j'ai souhaité évaluer la capacité de popoolationTE2 à discriminer des copies strictement identiques d'un ET (Gypsy7) entre 3 populations (homogènes, i.e. 100 % des individus d'une population ont l'ET). La répartition des copies dans les populations aux différents loci est décrite en Figure 4.B.4.2.

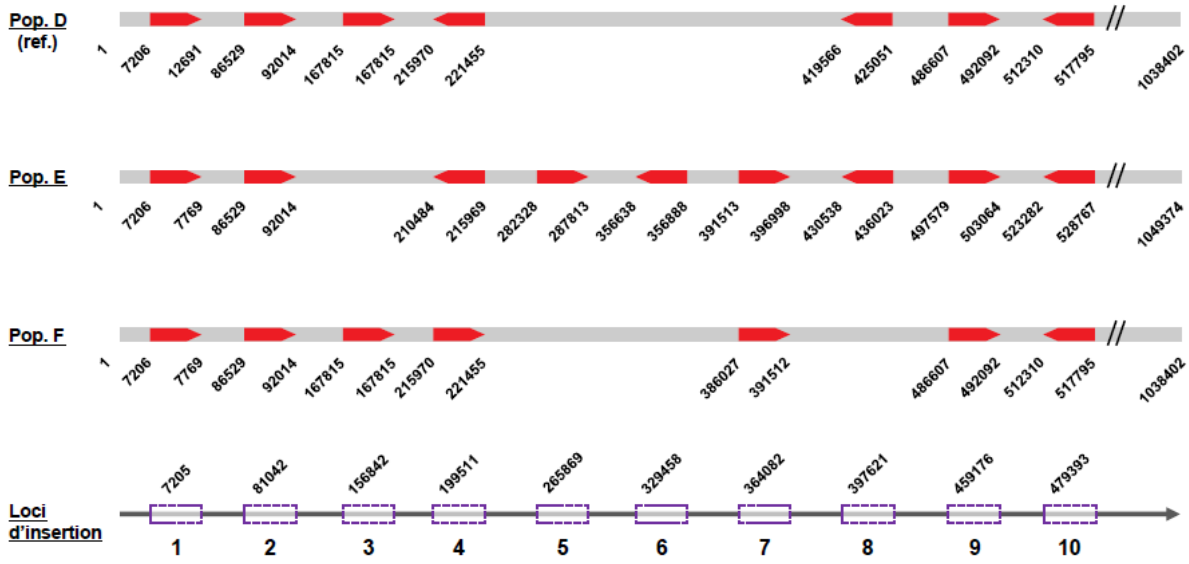


Figure 4.B.4.2 : paysage d'ETs de faible complexité simulée pour 3 populations homogènes.

Les points d'insertions des ETs sur le châssis sont représentés sur la frise du bas par des rectangles en pointillés (10 points d'insertion au total). Les valeurs données au-dessus de chaque point d'insertion correspondent à la position de chaque point d'insertion sur le châssis vide. Les 3 frises supérieurs représentent le paysage d'ET dans le génome de chaque population (D, E, et F). L'ET inséré est Gypsy7 (5486 pb). Chaque copie d'ET est représentée par une flèche, le sens de la flèche indiquant le sens d'insertion. Les valeurs en dessous de chaque frise correspondent aux positions de début et de fin de chaque ET sur le génome de la population en question.

Chaque population (D, E, et F) est homogène, i.e. 100% des individus d'une population présentent strictement le même paysage d'ETs.

Les résultats de cette analyse sont détaillés dans le Tableau 4.B.4.2. L'ensemble des ETs décrits pour chaque population ont été détectés avec succès. L'efficacité de détection des ETs est de 100 % pour cette simulation. En ce qui concerne la précision de détection des éléments, on observe une variabilité moyenne de 5.56 pb (+/- 0.87 pb) entre les positions détectées et les positions attendues. Ce

résultat est en accord avec les valeurs données dans la documentation de l'outil (~10 pb). PopoolationTE2 est donc capable de discriminer des copies identiques d'ETs avec une grande précision.

Tableau 4.B.4.2 : nature et position des ETs détectés dans des populations homogènes par PopoolationTE2 (paysage d'ET de faible complexité)

	popA				
	ET		position		
locus	**attendu**	**observé**	**attendue**	**observée**	**distance (pb)**
1	Gypsy7	Gypsy7	9949	9958	9
2	Gypsy7	Gypsy7	89272	89280	8
3	Gypsy7	Gypsy7	170558	170557	1
4	Gypsy7	Gypsy7	218713	218711	2
5	-	-	-	-	-
6	-	-	-	-	-
7	-	-	-	-	-
8	Gypsy7	Gypsy7	422309	422311	2
9	Gypsy7	Gypsy7	489350	489357	7
10	Gypsy7	Gypsy7	515053	515049	4

	popB				
	ET		position		
locus	**attendu**	**observé**	**attendue**	**observée**	**distance (pb)**
1	Gypsy7	Gypsy7	9949	9958	9
2	Gypsy7	Gypsy7	89272	89280	8
3	-	-	-	-	-
4	Gypsy7	Gypsy7	218713	218711	2
5	Gypsy7	Gypsy7	287813	287817	4
6	Gypsy7	Gypsy7	351402	351382	20
7	Gypsy7	Gypsy7	386026	386022	4
8	Gypsy7	Gypsy7	422309	422311	2
9	Gypsy7	Gypsy7	489350	489357	7
10	Gypsy7	Gypsy7	515053	515049	4

	popC				
	ET		position		
locus	**attendu**	**observé**	**attendue**	**observée**	**distance (pb)**
1	Gypsy7	Gypsy7	9949	9958	9
2	Gypsy7	Gypsy7	89272	89280	8
3	Gypsy7	Gypsy7	170558	170557	1
4	Gypsy7	Gypsy7	218713	218711	2
5	-	-	-	-	-
6	-	-	-	-	-
7	Gypsy7	Gypsy7	386026	386022	4
8	-	-	-	-	-
9	Gypsy7	Gypsy7	489350	489357	7
10	Gypsy7	Gypsy7	515053	515049	4

L'estimation des fréquences de présence d'ETs par popoolationTE2 est satisfaisante bien que légèrement sous-évaluée.

Dans cette simulation, j'ai cherché à évaluer l'efficacité de détection (nature et position de l'ET) de popoolationTE2 à partir d'une population hétérogène mais aussi à évaluer la précision avec laquelle cet outil estime la fréquence de présence de chaque ET, i.e. la part de la population pour laquelle un ET est détecté à une position donnée. J'ai créé la population hétérogène en combinant les données des 3 populations homogènes A, B et C de la première simulation. La population hétérogène ainsi formée est composée de 3 « individus ». Les fréquences de présence d'ETs attendues sont donc de 33%, 66% ou 100% en fonction des cas de figures. Le paysage d'ETs de cette population hétérogène est détaillé en Figure 4.B.4.3.

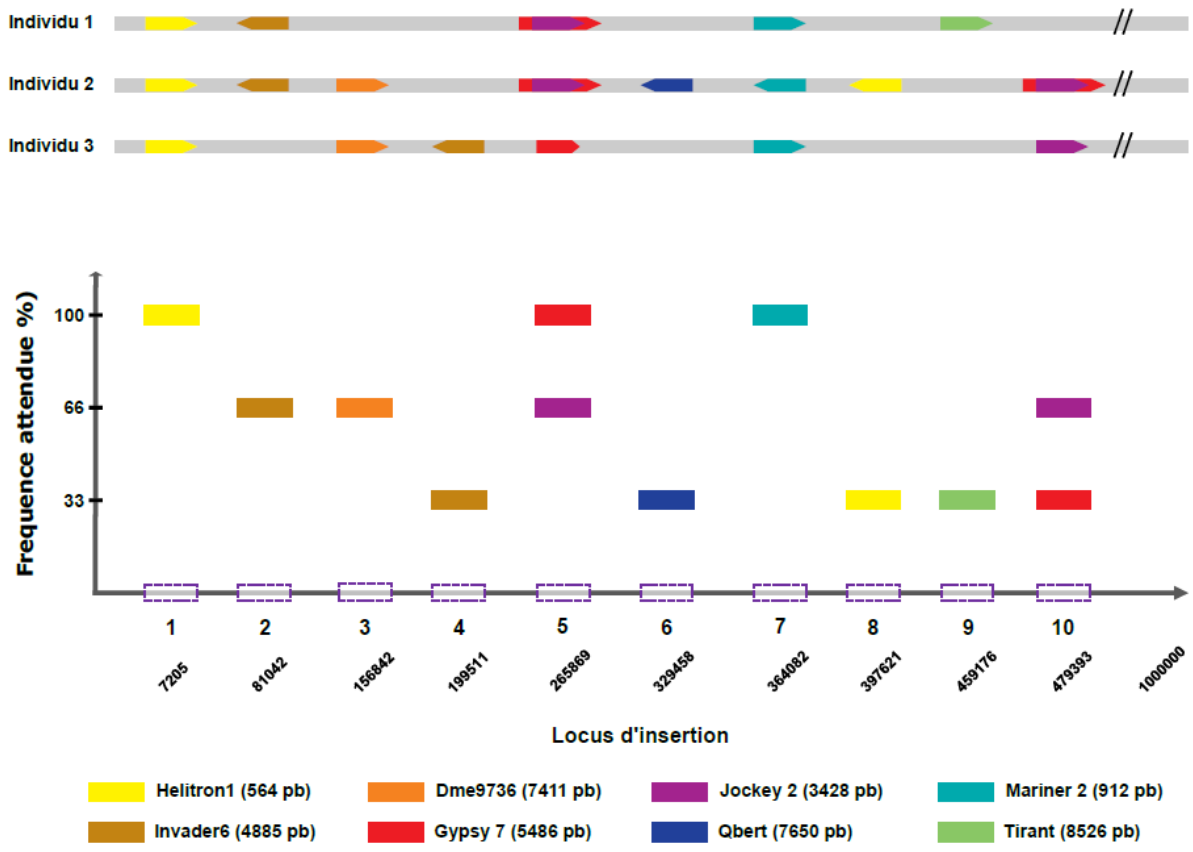


Figure 4.B.4.3 : paysage d'ETs d'une population hétérogène et fréquence attendues.

La population hétérogène est composée de trois « individus » présentant des paysages d'ETs différents. La fréquence d'un ET correspond à la part de la population pour laquelle l'ET est présent à cette position. Les fréquences d'ET attendues dans la population pour chaque locus sont représentées dans la sous-figure du bas. Les fréquences sont données sous forme de pourcentage. Dans cette population hétérogène simpliste, les fréquences d'ETs attendues sont 33%, 66% ou 100% en fonction des cas de figures.

Les ETs sont correctement détectés pour l'ensemble des loci à l'exception des loci 5 et 10 (voir Tableau 4.B.4.3). Au locus 5, seul Gypsy7 est détecté au lieu de l'ET Gypsy7[jockey2] qui devrait être présent pour 2/3 de la population. Au locus 10, ici encore seul Gypsy 7 est détecté au lieu de la construction Gypsy7[jockey2]. Jockey2 seul est quant à lui correctement détecté. En ce qui concerne la précision de détection des éléments, on observe une variabilité moyenne de 7.45 pb (+/- 2.25 pb) entre les positions détectées et les positions attendues. On notera que cette variabilité, bien que restreinte et toujours dans les valeurs attendues, est plus importante que pour les deux analyses précédentes.

Tableau 4.B.4.3 : détection et estimation de fréquence d'ETs au sein d'une population hétérogène

Le paysage d'ET de la population hétérogène analysé est détaillé en Figure 4.B.4.3. La fréquence d'un ET correspond à la part de la population pour laquelle l'ET est présent à cette position. Les valeurs de fréquences sont données sous forme de pourcentage de la population. Les différences de fréquences (colonne « différence »), exprimées sous forme de pourcentage, correspondent à la valeur absolue de la différence entre la fréquence attendue et la fréquence observée pour un ET à un locus donné.

	ET		**position**			**fréquence (%)**		
locus	**attendu**	**observé**	**attendue**	**observée**	**distance (pb)**	**attendue**	**observée**	**différence**
1	Helitron1	Helitron1	7488	7488	0	100,0	93,5	6,5
2	Invader6	Invader6	84049	84050	1	66,6	61,7	4,9
3	Dme9736	Dme9736	162291	162297	6	66,6	58,5	8,1
4	Invader6	Invader6	204990	204971	19	33,3	28,5	4,8
5	Gypsy7 [Jockey2]	Gypsy7	275776	275778	2	100,0	93,7	27,1
	Gypsy7	-	275776	-	-	0,666	-	-
6	Qbert	Qbert	343821	343817	4	33,3	28,3	5,0
7	Mariner2	Mariner2	378902	378898	4	1	93,7	6,3
8	Helitron1	Helitron1	412896	412909	13	33,3	28,5	4,8
9	Tirant	Tirant	478715	478693	22	33,3	31,7	1,6
10	Gypsy7 [Jockey2]	Gypsy7	503194	503185	9	33,3	29,5	3,8
	Jockey2	Jockey2	503194	503196	2	66,66	27,6	39,1

L’estimation des fréquences est cohérente avec celles attendues pour l’ensemble des loci à l’exception des loci 5 et 10. Au locus 5, la construction Gypsy7[jockey2] n’est pas détectée. En revanche Gypsy7 seul est détecté avec une fréquence de 93.7%. Au locus 10, seul Gypsy7 est détecté pour la construction Gypsy7[jockey2]. La fréquence prédite de Gypsy7 (29.5%) a cette position est cohérente avec celle attendue (33%). Jockey2 est uniquement détecté dans sa forme « simple » (27.6% ~1/3) alors que sa fréquence devrait être de 66% s’il avait été détecté dans la construction Gypsy7[jockey2]. Ainsi, dans le cas où un ET est inséré à l’intérieur d’un autre, seul l’ET « englobant » est correctement détecté et sa fréquence est correctement estimée. On notera néanmoins que de manière générale, les valeurs de fréquences prédites sont toujours sous-évaluées par rapport à celles attendues : 28-31% observé contre 33% attendu, 58-61% observé vs 66 % attendu et ~93% observé *vs*. 100% attendu. En moyenne, la valeur absolue de la variation de fréquence entre les valeurs attendues et observées est de 8,28% (+/- 3,11 %).

Les capacités de détection d’ET (nature, position) et d’estimation de leur fréquence par popoolationTE2 au sein d’une population hétérogène sont donc satisfaisantes bien que les fréquences estimées aient tendance à être légèrement sous évaluées.

5 - Discussion

Dans cette analyse, j'ai comparé deux outils de détection de polymorphismes à partir de données que j'avais préalablement simulées. Bien que ces données soient simplistes, elles m'ont permis de tester une variété importante de cas de figures possibles pouvant ensuite être rencontrés sur les données réelles. Par ailleurs, ces analyses m'ont permis de prendre en main ces outils et d'en saisir les spécificités d'utilisation. J'en ai par exemple retenu que seules les détections présentant des signatures d'insertion « FR » devaient être conservées dans les analyses réalisées avec popoolationTE2, les autres détections étant assimilables à du bruit de fond. L'ensemble des analyses réalisées m'ont finalement permis de déterminer que la solution la plus adaptée à mes analyses futures parmi les deux testées était popoolationTE2.

Dans les différentes simulations présentées ici, popoolationTE2 a démontré une très bonne efficacité de détection (la bonne copie d'ET détectée à la bonne position) et ce même dans des cas complexes comme lorsque i) deux ETs différents sont présents à la même position pour deux populations différentes, ii) plusieurs copies strictement identiques sont présentes dans une région génomique restreinte, ou encore iii) l'ET n'est pas décrit dans le génome de référence. Ce dernier point constitue un des avantages majeurs de popoolationTE2 par rapport à TEPID. En effet, l'analyse réalisée par TEPID est uniquement basée sur l'étude des ET décrits dans le génome de référence et son annotation en ET. La complétude de cette analyse est donc proportionnelle à cette de l'annotation de référence des ETs puisque seules les copies d'ETs décrites dans cette annotation seront analysées. A l'inverse, popoolationTE2 donne l'opportunité de fournir les séquences consensus des ETs (ou d'autres séquences) en plus des séquences des copies de ces ETs. Grâce à cela, il est possible de détecter des ETs non décrits dans le génome de référence (si leur séquence est fournie) ou encore de détecter des copies plus divergentes et donc même de compléter l'annotation de référence. Empiriquement, j'ai conclu que l'ajout de la librairie de séquences consensus en plus des séquences des copies d'ET décrites dans le fichier d'annotation améliore la qualité de la détection.

PopoolationTE2 présente aussi une précision d'estimation très satisfaisante de la position des sites polymorphes puisque dans chacune des simulations présentées j'ai pu observer une variabilité inférieure à 10 pb entre les positions détectées et les positions réelles des ETs. J'ai néanmoins pu noter que dans le cas d'une grande concentration d'ETs (proximité importante), on pouvait observer une diminution drastique de la précision d'estimation de la position des sites polymorphes mais aussi de la qualité de détection des ETs (données non présentées). Ce phénomène est très probablement dû à

l'algorithme de popoolationTE2 qui repose sur l'analyse de la couverture d'insert de reads discordants et nécessite donc une certaine distance entre les ETs. PopoolationTE2 ne semble donc pas être adapté à l'analyse de paysage d'ETs trop denses. Empiriquement, j'ai pu noter qu'une distance minimale de ~700 pb était nécessaire pour que 2 ETs puissent être correctement discriminés (positions et natures des ETs ainsi que fréquence dans la population) avec le paramétrage décrit en méthodes. Ce point mériterait de réaliser des analyses complémentaires.

Outre son efficacité de détection accrue par rapport à TEPID, popoolationTE2 présente l'avantage d'estimer la fréquence de présence des ETs dans chaque population et ce pour l'ensemble des sites polymorphes. Cette information supplémentaire est particulièrement intéressante puisqu'il est alors possible de comparer la fréquence d'un ET entre populations mais aussi d'avoir une idée de la variabilité du contenu en ET au sein d'une population. En contrepartie, si l'on souhaite effectuer une analyse binaire (e.g. détection d'insertion/délétions d'ETs), il est nécessaire i) de définir des seuils de fréquences à partir desquels on considère l'ET comme présent ou absent pour une population et une position donnée et ii) de choisir quelle population utiliser comme référentiel.

PopoolationTE2 est donc capable d'évaluer précisément la position d'un polymorphisme et la nature du / des ETs impliqué(s) ainsi que la fréquence de présence associée(s) dans les populations étudiées. Cet outil est adapté à l'étude des polymorphismes d'ETs au sein des espèces du genre *Meloidogyne* (voir chapitre VI).

V – Charge et composition en ETs au sein du genre *Meloidogyne*

Avant-propos

Cette analyse repose partiellement sur l'utilisation de données non-publiées à l'heure actuelle.

1 - Contexte

Les nématodes à galles (*Meloidogyne*) constituent un groupe d'espèces varié en termes de gamme de plantes hôtes, de répartition géographique, mais aussi en termes de traits biologiques comme le mode de reproduction. Il a été observé que plusieurs espèces de *Meloidogyne* sont capables de contourner les défenses de plantes initialement résistantes en l'absence de reproduction sexuée, et ce, en un nombre de générations restreint. Cette observation est surprenante car, comme présenté en introduction, en absence de recombinaison sexuée, le potentiel adaptatif d'une espèce est théoriquement très restreint. D'autres mécanismes que le brassage génétique dû à la recombinaison et pouvant rapidement créer de la plasticité génomique sont donc nécessairement à l'œuvre pour sous-tendre cette adaptabilité en l'absence de sexe. Nous faisons l'hypothèse que les ETs pourraient jouer / avoir joué un rôle dans la plasticité génomique des espèces à reproduction strictement asexuées de *Meloidogyne.*

Afin d'explorer cette hypothèse, il est dans un premier temps nécessaire de déterminer si la charge et la composition en ETs au sein des génomes des espèces de Meloidogyne sont liées à des traits biologiques et/ou à de l'héritage phylogénétique.

Deux études consécutives ont été réalisées sur des sujets similaires par le groupe de Dave Lunt, à l'Université de Hull en Angleterre. La première portant sur l'ensemble de l'embranchement Nematoda (42 espèces dont 4 de *Meloidogyne*) (Szitenberg A. et al., GBE, 2016) et la seconde portant sur les espèces apomictiques du genre *Meloidogyne* (6 espèces) (Szitenberg et al., 2017). Dans ces deux analyses, Szitenberg et collaborateurs ont conclu qu'il n'y avait pas de différences claires dans la teneur en éléments transposables entre les espèces analysées, ni en fonction de leurs traits d'histoire de vie (dont le mode de reproduction), ni en fonction de leur position phylogénétique dans l'arbre de vie des nématodes. On pourra néanmoins noter le manque d'espèces « réplicats » dans une analyse comme dans

l'autre concernant le mode de reproduction. Par ailleurs, ces deux articles reposent sur le même protocole d'analyse du contenu en ETs des génomes basés sur l'étude de génomes assemblés. Or, en utilisant ces mêmes données, Ranallo-Benavidez et collaborateurs ont montré que leur estimation de la taille des génomes de Meloidogyne était 1,65-2,69 fois plus grande que l'assemblage produit dans (Szitenberg A. et al., GBE, 2017) et ont suggéré que ces "assemblages ont partiellement collapsés les chromosomes homologues" (Ranallo-Benavidez et al., 2020). Or, comme précédemment évoqué (voir chapitre IV A), la qualité de la prédiction et de la détection des ETs basée sur l'utilisation de génomes assemblés dépend fortement de la qualité de l'assemblage (fragmentation, phasage, etc.). Il est donc probable que les estimations du contenu en ET réalisées par Szitenberg et collaborateurs aient pâti de ce biais.

Dans cette étude, j'ai évalué la charge et la composition en ETs au sein d'espèces du genre *Meloidogyne*. Par rapport aux précédentes études, des espèces supplémentaires ont été introduites ainsi que plusieurs réplicats de données pour certaines espèces. En outre, j'ai basé cette étude sur l'utilisation de l'outil dnaPipeTE (Goubert et al., 2015) qui repose sur l'analyse de données brutes de séquençage plutôt que sur des séquences génomiques et qui n'est donc pas soumis aux biais de qualité d'assemblage des génomes.

2 - Matériel

J'ai utilisé 16 librairies de données de séquençage (reads) pour un total de 8 espèces: *M. incognita, M. floridensis, M.arenaria, M. javanica, M. luci, M. enterolobii, M. graminicola, M. chitwoodi*. Des détails supplémentaires sur les caractéristiques de ces librairies de données de séquençage (nombre brut de reads par librairie, taille des reads, etc) sont disponibles dans le Tableau annexe 1. Les librairies dont les numéros d'accessions sont les suivants ont été téléchargées depuis la base de données publique ENA : SRR4242457, SRR4242458, SRR4242460, SRR4242472, SRR4242474, SRR2350716, MgVN18S1, ERR1212566, ERS3574357. Les librairies nommées "arenariaV3" , "incognitaV3" et "javanicaV3" sont disponibles dans les bases de données publiques sous les numéros d'accession suivants : ERS671128, ERS1696677 et ERS671129. Les librairies nommées "arenariaV4" , "incognitaV4" et "javanicaV4" ont été produites plus récemment par notre laboratoire et ne sont pas publiées à l'heure actuelle.

3 - Méthodes

Pré-traitement global des données de séquençage réelles

Pour chaque librairie, j'ai i) coupé les extrémités de mauvaise qualité des reads et ii) éliminé les reads ayant une valeur de qualité moyenne < 25 et dont la taille était < 95 pb à l'aide de trimmomatic v-0.39 (Bolger et al., 2014) (SE LEADING:25 TRAILING:25 SLIDINGWINDOW:4:25 MINLEN:95). J'ai ensuite effectué un contrôle qualité manuel des librairies grâce à fastqc v-0.11.9 (Andrews S., 2010) afin d'identifier des signes de contamination (via l'analyse de la distribution du contenu en GC) ou autres défauts dans les librairies qui pourraient interférer avec les analyses futures. Les librairies ERR1212566 (*M. chitwoodi*) et enterolobiiV3 présentaient des signes de contamination et ont été retirées de l'analyse.

Estimation de la ploïdie et de la taille des génomes à partir des données de séquençage réelles

Estimation de la ploïdie

J'ai estimé pour chaque librairie la ploïdie de l'organisme en suivant le protocole décrit dans la documentation de smudgeplot v-0.2.3 (https://github.com/KamilSJaron/smudgeplot). (Ranallo-Benavidez et al. 2020). Smudgeplot permet de visualiser et d'estimer la ploïdie et la structure d'un génome par l'analyse des paires de k-mer hétérozygotes.

Pour chaque librairie indépendamment, j'ai concaténé les fragments gauches et droits de chaque read dans un même fichier. J'ai ensuite utilisé la suite d'outil kmc v-3.1.1 (Kokot et al., 2017) (kmc -k21 -m200 -ci1 -cs10000 ; kmc_tools transform histogram -cx10000) afin de générer un histogramme de fréquence/couverture des k-mer (21 pb) pour chaque librairie. J'ai ensuite utilisé l'outil cutoff de smudgeplot afin de calculer les bornes de l'intervalle de fréquence « utile » de chaque histogramme. J'ai mis les valeurs calculées en entrée de l'outil transform (-ci L -cx U dump ; avec L étant la valeur de la borne minimale et U la valeur de la borne maximale) de kmc_tools afin d'éliminer les parties « non-utiles » de chaque histogramme (i.e. en dehors des bornes données). Enfin, j'ai utilisé les outils hetkmers et plot de smudgeplot afin de réaliser une estimation de la ploïdie à partir de chaque histogramme « recadré ».

Estimation de la taille des génomes

J'ai utilisé l'outil genomescope v-2.0 (Ranallo-Benavidez et al., 2020) pour estimer les tailles des génomes haploïdes correspondant à chaque librairie. J'ai réalisé cette analyse à partir des histogrammes complets/originaux de couverture en k-mer précédemment générés avec kmc. Pour chaque librairie analysée, j'ai placé la borne maximale (-m) aux valeurs maximales de couverture utile des k-mer précédemment calculées avec l'outil cutoff de smudgeplot.

Pour chaque librairie, j'ai ensuite calculé la taille du génome en multipliant la taille du génome haploïde par le niveau de ploïdie précédemment estimé.

Homogénéisation de la taille des données de séquençage

Homogénéisation intra-librairie

Comme recommandé dans la documentation de dnaPipeTE v-1.3 (Goubert et al., 2015) (https://github.com/clemgoub/dnaPipeTE/wiki/Running-dnaPIpeTE), il est nécessaire que la taille des reads soit homogène à l'intérieur de chaque librairie.

Pour chaque librairie, j'ai utilisé trimmomatic pour éliminer l'ensemble des reads dont la taille était inférieure de 5 pb à la taille maximale des reads rencontrés dans cette librairie (SE MINLEN:x ; avec x=95 pb pour les librairies arenariaV3, incognitaV3, javanicaV3, MgVN18S1 et SRR2350716 ; x=120 pb pour les librairies SRR4242457, SRR4242458, SRR4242460 et SRR4242472 ; x=145 pb pour les librairies SRR4242474 et ERS3574357 ; et x=245 pb pour arenariaV4, incognitaV4 et javanicaV4). Ainsi, chaque librairie de données réelles de séquençage est homogène en longueur de reads à 5 pb près. Il sera dorénavant fait référence à ces données sous l'appellation « reads homogènes intra-lib». On notera que la taille maximum étant la plus fréquente dans l'ensemble des librairies, cette étape n'a supprimé que peu de données.

Homogénéisation inter-librairies

J'ai homogénéisé la taille des reads entre l'ensemble des librairies en raccourcissant avec trimmomatic (MAXLEN :101) l'ensemble des reads de l'ensemble des données réelles de séquençage à une taille maximale de 101 pb. Chaque librairie générée renferme donc uniquement des reads dont la taille est homogène et est comprise entre 95 pb et 101 pb. Il sera dorénavant fait référence à ces données sous l'appellation « reads homogènes inter-lib ».

Création d'une gamme de taille de reads à partir de données réelles

J'ai utilisé les librairies de « reads homogènes intra-lib» arenariaV4, incognitaV4 et javanicaV4 (2*250 pb) afin de générer 4 autre librairies de reads de taille inférieures (variabilité de taille des reads au sein d'une librairie : +/- 1 pb). Pour ce faire, pour chaque librairie, j'ai raccourci les reads originaux aux longueurs suivantes : 200, 150, 125 et 101 pb avec trimmomatic (SE CROP:n ; avec n = 200, 150, 125, 101). Au total, les librairies arenariaV4, incognitaV4 et javanicaV4 ont donc été déclinées en 5 libraires de reads homogènes en taille : 250 pb, 200 pb, 150 pb, 125 pb et 101 pb. Il sera dorénavant fait référence à ces données sous l'appellation « gamme de taille de reads à partir de données réelles ».

Simulation de données de séquençage

J'ai simulé 3 librairies de reads illumina « Paired-End » (PE) 2*250 pb avec art (Huang et al., 2012) (art_illumina v-2.5.8 ; -ss MSv3 -p -na -qL 23 -l250 -f100 -m800 -s50) en me servant des génomes assemblés de référence de *M.arenaria* (GCA_900003985.1), *M. incognita* (GCA_900182535.1) et *M. javanica* (GCA_900003945.1) comme châssis. J'ai généré des reads jusqu'à atteindre 100X de couverture par espèces, soit 47 358 444 reads pour *M. arenaria*, 34 708 835 *reads pour M. incognita et* 42 358 747 *reads pour M. javanica.* J'ai ensuite filtré les reads simulés avec trimmomatic (SE LEADING:25 TRAILING:25 SLIDINGWINDOW:4:25 MINLEN:95) pour ne retenir que les reads avec une valeur de qualité >= 25. Enfin, j'ai raccourci les reads restants à différente taille (200, 150, 125 et 101 pb) avec trimmomatic (SE CROP:n ; avec n = 200, 150, 125, 101) afin de générer 4 autres librairies de reads homogènes en taille par espèce (variabilité de taille des reads au sein d'une librairie : +/- 1 pb). Au total, j'ai donc généré 5 libraires de reads homogènes en taille par espèce : 250 pb, 200 pb, 150 pb, 125 pb et 101 pb.

Prédiction de la charge et du contenu en ETs

dnaPipeTE v-1.3 (pour « de-novo assembly & annotation Pipeline for Transposable Elements »), est un pipeline d'analyse conçu pour trouver, annoter et quantifier les éléments transposables à partir de petits échantillons de données de séquençage (librairies de reads) (Goubert et al., 2015). La philosophie de cet outil est la suivante : les séquences d'ETs, de par leur nature répétée, sont surreprésentées dans les données de séquençage génomique. En sous-échantillonnant suffisamment ces données de séquençage, seuls les reads correspondant à des régions répétées devraient rester. dnaPipeTE utilise l'assembleur RNAseq Trinity (Grabherr et al., 2011) pour construire des contigs de séquence répétées à partir de petits échantillons génomiques (couverture du génome < 1X). Ces séquences sont

ensuite classifiées en utilisant RepeatMasker (Smit et al., 2013) puis leur proportion dans le génome est quantifiée partir de l'échantillon grâce à un blastn des reads sur les contigs créés.

Pour chacune des analyses réalisées avec dnaPipeTE présentées ci-dessous (i.e estimation du contenu en ET à partir d'une librairie de reads donnée), 3 passes indépendantes de sous-échantillonnage et d'assemblage ont été réalisées (-sample_number 3). Pour chaque analyse, la taille de génome fournie pour calculer le nombre de reads à sous-échantillonner pour atteindre la couverture cible (-genome size) correspond à la valeur précédemment estimée in-silico à l'aide de smudgeplot et genomescope2 pour la librairie en question. Les résultats produits ont été exploité sous R' v-3.6.

Test de l'importance du paramètre "couverture de sous-échantillonage"

J'ai testé l'influence de la couverture de sous-échantillonnage sur l'estimation du contenu en ET par dnaPipeTE en réalisant une série d'analyses sur une gamme de couvertures cibles pour chaque librairie de « reads homogènes intra-lib».

Pour chaque librairie, j'ai réalisé 4 analyses indépendantes avec dnaPipeTE à différente couverture de sous-échantillonnage (-genome_coverage) : 0.125X, 0.250X, 0.375X et 0.5X.

Test de l'importance de la longueur des reads

J'ai testé l'influence de la taille des reads (différence de longueur de reads entre librairies) sur l'estimation du contenu en ET par dnaPipeTE en réalisant une série d'analyse à partir de donnée simulées et réelles (« gamme de taille de reads à partir de données réelles ») de tailles variables : 250 pb, 200 pb, 150 pb, 125 pb et 101 pb ; la variabilité de taille des reads au sein d'une librairie étant de +/- 1 pb pour les données simulées comme pour les données réelles.

Pour chaque librairie, j'ai réalisé une analyse avec dnaPipeTE en fixant arbitrairement la couverture cible à 0.25X (-genome_coverage).

Analyse comparative du contenu en ETs au sein des espèces de *Meloidogyne*.

Afin de limiter les biais d'estimation liés à la couverture de sous échantillonnage et à la différence de longueur de reads entre librairies, j'ai réalisé l'analyse du contenu en ETs au sein des espèces de *Meloidogyne* à une couverture fixe et à partir de données pour lesquelles la taille des reads était homogène entre librairies (« reads homogènes inter-lib », 95-101 pb pour l'ensemble des librairies). Pour chaque librairie, j'ai réalisé une analyse avec dnaPipeTE en fixant arbitrairement la couverture cible à 0.25X (-genome_coverage).

4 - Résultats

L’estimation in silico de la ploïdie et de la taille des génomes est cohérente avec la réalité biologique.

J’ai estimé la ploïdie des espèces de *Meloidogyne* à partir de données de séquençages grâce à une méthode basée sur l’analyse de l’histogramme des k-mer (voir méthodes).

L’estimation de la ploïdie des espèces étudiées est cohérente avec les valeurs décrites dans la littérature, et ce pour l’ensemble des espèces à l’exception de *M. floridensis* (voir Tableau 4.4.1). Handoo et collaborateurs ont décrit que le génome de *M. floridensis* comportait 36 chromosomes dont 18 durant sa phase haploïde, ce qui signifie que cette espèce serait diploïde (Handoo et al., 2004). La ploïdie estimée *in silico* pour cette espèce (3n) est donc supérieure de 1 copie par rapport à la description cytogénétique (2n). On notera tout de même que bien que cette espèce soit décrite comme diploïde, son mode de reproduction exact (parthénogénétique obligatoire vs. facultatif) est inconnu à ce jour. Dans le seul article décrivant la cytogénétique de *M. floridensis*, il est dit que, durant l'ovogenèse cette espèce réalise une seule des deux divisions de la méiose et qu’il pourrait s’agir d’un intermédiaire entre les espèces mitotiques et méiotiques (Handoo 2004). Les espèces pour lesquelles je disposais de plusieurs librairies issues de séquençages indépendants réalisés à partir de lignées différentes (*e.g. M. incognita, javanica, arenaria,* et *graminicola)* présentent des estimations de ploïdie concordantes entre réplicats. D’une manière générale, la ploïdie estimée varie entre 2 et 4 au sein des espèces de *Meloidogyne*, les espèces parthénogénétiques obligatoires (mitotiques) ayant le plus haut niveau de ploïdie (3-4) alors que l’espèce parthénogénétique facultative (méiotique) *M. graminicola* présente un niveau de ploïdie inférieur (2).

Tableau 4.4.1 : estimation de la ploïdie à partir de librairies de reads et comparaison avec la littérature

(!) : estimation de la ploïdie réalisée selon la même méthodologie que dans cette analyse.

Espèce	numéro d'accession / nom	ploïdie estimée	ploïdie (littérature)
M. arenaria	arenariaV3	4	4 (Blanc-Mathieu et al., 2017; Szitenberg et al., 2017a)
	arenariaV4	4	
	SRR4242457	4	
M. javanica	javanicaV3	4	4 (Blanc-Mathieu et al., 2017; Szitenberg et al., 2017a)
	javanicaV4	4	
	SRR4242458	4	
M. incognita	incognitaV3	3	3 (Blanc-Mathieu et al., 2017; Szitenberg et al., 2017a)
	incognitaV4	3	
	SRR4242460	3	
M. enterolobii	SRR4242472	3	3 (Szitenberg et al., 2017a)
M. floridensis	SRR4242474	3	2 (Handoo et al., 2004; Szitenberg et al., 2017a)
M. graminicola	MgVN18S1	2	2 (Phan et al., 2020; Somvanshi et al., 2018)
	SRR2350716	2	
M. luci	ERS3574357	3	3 (!) (Susič et al., 2020)

De la même manière que pour le niveau de ploïdie, j'ai estimé la taille des génomes des espèces de *Meloidogyne* à partir de données de séquençage. Pour ce faire, j'ai calculé la taille du génome haploïde de chaque espèce à partir de données de séquençage grâce à une méthode basée sur l'analyse de l'histogramme des k-mer (voir méthodes). J'ai ensuite multiplié les valeurs obtenues par les valeurs de ploïdie précédemment estimées afin d'obtenir la taille des génomes.

Pour les espèces pour lesquelles plusieurs librairies sont disponibles, l'évaluation de la taille des génomes est homogène entre réplicats (voir Tableau 4.42). Pour ces espèces la taille moyenne estimée pour chaque génome est de 264.53 Mb (sd=7.80) pour *M. arenaria*, 253.80 Mb (sd = 16.06) pour *M. javanica* , 188.20 Mb (sd = 2.73) pour *M. incognita*, et enfin 69.54 Mb (sd = 6.55) pour *M. graminicola.* Pour *M. arenaria, javanica* et *incognita,* ces valeurs moyennes sont cohérentes avec les estimations de taille de génome réalisées en cytométrie en flux (Blanc-Mathieu et al. 2017). En ce qui concerne *M. graminicola*, l'évaluation *in silico* de la taille des génomes est légèrement inférieure à la valeur obtenue en cytométrie (81.5-83.8Mb, (Phan et al., 2020)), et ce pour les deux librairies analysées. Pour les espèces pour lesquelles je ne disposais d'aucune estimation expérimentale de la taille des génomes, j'ai comparé les tailles estimées avec la longueur de l'assemblage des génomes actuels. L'estimation *in silico* de la taille du génome de *M. luci* est cohérente avec la taille de l'assemblage du génome, ce qui n'est en revanche pas le cas pour *M. floridensis* et *M. enterolobii*. Néanmoins, il a été suggéré par Ranallo-Benavidez et collaborateurs que des régions homologues de ces deux génomes avaient probablement été collapsés (fusionnés) durant leurs assemblages à partir de ces librairies (Ranallo-Benavidez et al., 2020). Il en est de même pour les assemblages réalisés à partir des librairies SRR4242457 (*M. arenaria*), SRR4242458 (*M. javanica*) et SRR4242460 (*M. incognita*), tous faisant partie de la même analyse (Szitenberg et al., 2017b).

Ainsi, malgré quelques variations, l'estimation de la taille des génomes que j'ai réalisé *in silico* est cohérente avec la réalité biologique.

Tableau 4.4.2 : estimation de la taille des génomes à partir de librairies de reads et comparaison avec la littérature.

(*) taille haploïde. L'estimation de la taille du génome effectuée en cytométrie de flux chez *M. graminicola* (MgVN18S1) est cohérente avec la taille du génome diploïde (2*41.5 = 83 Mb)

Espèce	Numéro d'accession / nom	Estimation taille génome haploïde (Mb)	Estimation taille du génome (ploidie*taille haploïde)	Longueur du génome assemblé (littérature) (Mb)	Estimation taille du génome en cytométrie en flux (Mb)
M. arenaria	arenariaV3	66.453677	265.814708	258.07	304+/-9
	arenariaV4	64.039844	256.159376	NA	NA
	SRR4242457	67.903498	271.613992	163.770989	NA
M. javanica	javanicaV3	64.757560	259.030240	235.8	297+/-27
	javanicaV4	58.944735	235.778940	NA	NA
	SRR4242458	66.646268	266.585072	142.608877	NA
M. incognita	incognitaV3	63.783032	191.349096	183.53	189+/-15
	incognitaV4	62.213261	186.639783	NA	NA
	SRR4242460	62.203939	186.611817	122.043328	NA
M. enterolobii	SRR4242472	83.253990	249.761970	162.361678	NA
M. floridensis	SRR4242474	63.263110	189.789330	74.893904 (*)	NA
M. graminicola	MgVN18S1	32.452847	64.905694	41.5 (*)	86,9+/-9.54
	SRR2350716	37.083071	74.166142	38.18 (*)	NA
M. luci	ERS3574357	67.248938	201.746814	209.16	NA

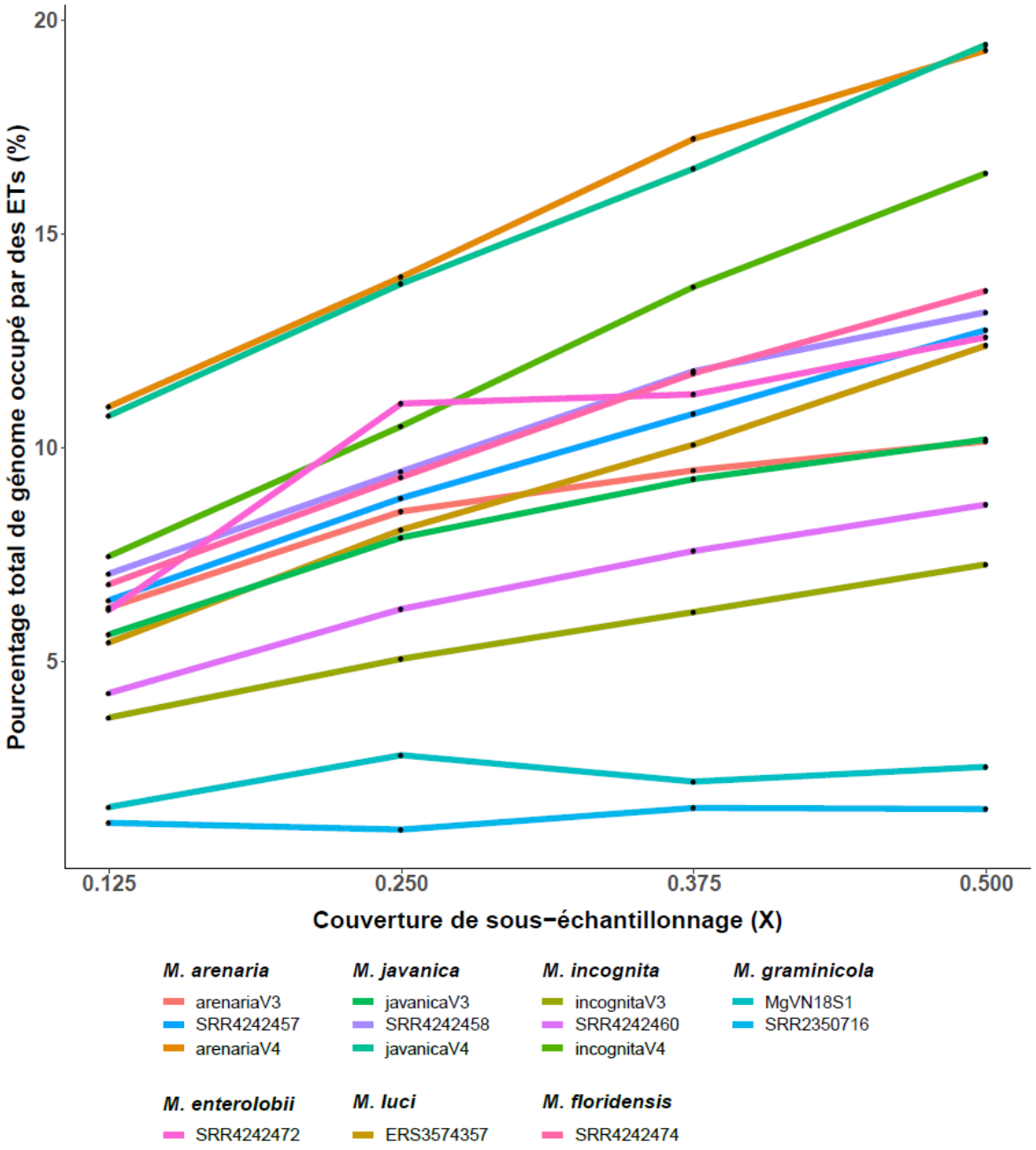


Figure 4.4.1 : distribution de la charge totale d'ET par librairie (% génome occupé) en fonction de la couverture de sous échantillonnage utilisée.

L'analyse de contenu et de charge en ET été réalisée sur 14 librairies de données de séquençage de 7 espèces de Meloidogyne et ce pour quatre valeurs de sous-échantillonnage (0.125X, 0.25X, 0.375X et 0.5X). Chaque point représente la charge totale d'ET estimée par librairie (% total de génome occupé) pour une couverture de sous-échantillonnage donné. La charge totale d'ET correspond à la somme cumulative des valeurs de % de génome occupé calculées pour les ordres LTR, LINE, SINE, DNA et Helitrons.

Couverture du sous-échantillonnage des reads

Dans un premier temps, j'ai souhaité estimer l'influence de la valeur de couverture de sous-échantillonnage sur l'estimation du pourcentage total de génome occupé par les ETs. La charge totale d'ETs correspond à la somme cumulative des valeurs calculées pour les ordres LTR, LINE, SINE, DNA et Helitrons. Pour chaque librairie, j'ai réalisé une analyse d'estimation du contenu en ET avec dnaPipeTE pour 4 couvertures cibles (<<1X) : 0.125X, 0.250X, 0.375X et 0.5X.
La Figure 4.4.1 montre que pour chacune des librairies, à l'exception de celles relatives à *M. graminicola* (MgVN18S1 et SRR2350716) et *M. enterolobii* (SRR4242472), il semble exister un lien entre la couverture cible de sous-échantillonnage et la charge totale d'ET détectée. Cette observation est confirmée par un test de corrélation sur l'ensemble des données (test de corrélation de Pearson; R= 0.44; p-valeur=5.9e-4). Ainsi la charge totale d'ETs prédite grâce à cette méthode *in silico* est partiellement dépendante de la couverture de sous-échantillonnage choisie.

Technologie de séquençage (*ie* longueurs des reads)

J'ai ensuite cherché à évaluer si le contenu en ETs était similaire entre plusieurs prédictions réalisées pour une même espèce à une même couverture (0.25X, valeur choisie arbitrairement) mais à partir de librairies Illumina différentes respectivement homogènes en taille de read. J'ai donc comparé la charge totale en ET prédite pour les 3 librairies disponibles de *M. arenaria*, *incognita* et *javanica* (voir Figure 4.4.2).

Bien que les tailles de génome prédites pour les librairies correspondant à une même espèce soient similaires (Table 2), on observe une différence notable en termes de pourcentage de génome occupé par les ETs d'après dnaPipeTE. La différence principale entre les libraires étant la taille de leurs reads, ce résultat semble indiquer qu'il existe un lien entre la taille des reads des libraires et la charge en ETs calculée par dnaPipeTE; la charge totale d'ETs croissant avec la taille des reads.

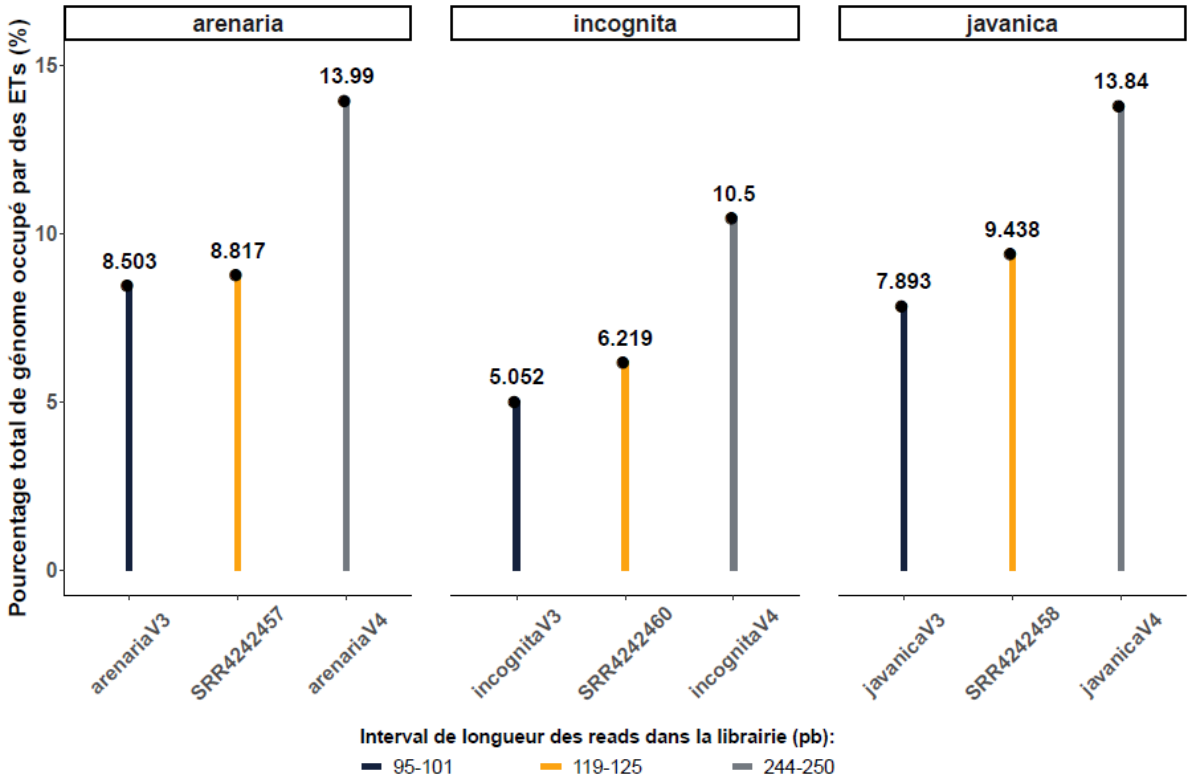


Figure 4.4.2 : charge totale d'ET par librairie de données de reads pour 3 espèces en "triplicat".

Les valeurs en gras représentent la charge en ETs (pourcentage de génome occupé par des ETs) pour chaque librairie de reads. Seuls les ETs des ordres LTR, LINE, SINE, DNA et Helitrons sont pris en compte. La couleur des bâtons des "sucettes" indique la longueur maximale des reads pour chaque librairie.

Afin de tester l'impact de la taille des reads sur l'estimation de la charge en ETs, j'ai créé deux séries de librairies de données de séquençage à partir de données réelles ou simulées pour former des gammes de tailles de reads (250 pb, 200 pb, 150 pb, 125 pb, 100 pb ; variabilité de taille de reads par librairie : +/- 1 pb ; voir méthodes). Pour les données réelles, je suis parti des librairies v4 de *M. incognita, arenaria et javanica* (2x250pb) que j'ai découpées comme indiqué dans les méthodes. Pour les données simulées, je suis parti des assemblages des génomes de référence de ces espèces pour générer des librairies de reads (2x250 pb) que j'ai ensuite découpées selon la même méthodologie que pour les données réelles. J'ai ensuite réalisé une analyse distincte avec dnaPipeTE a une couverture fixe (0.25X, valeur choisie arbitrairement) pour chacune de ces librairies (voir Figure 4.4.3).

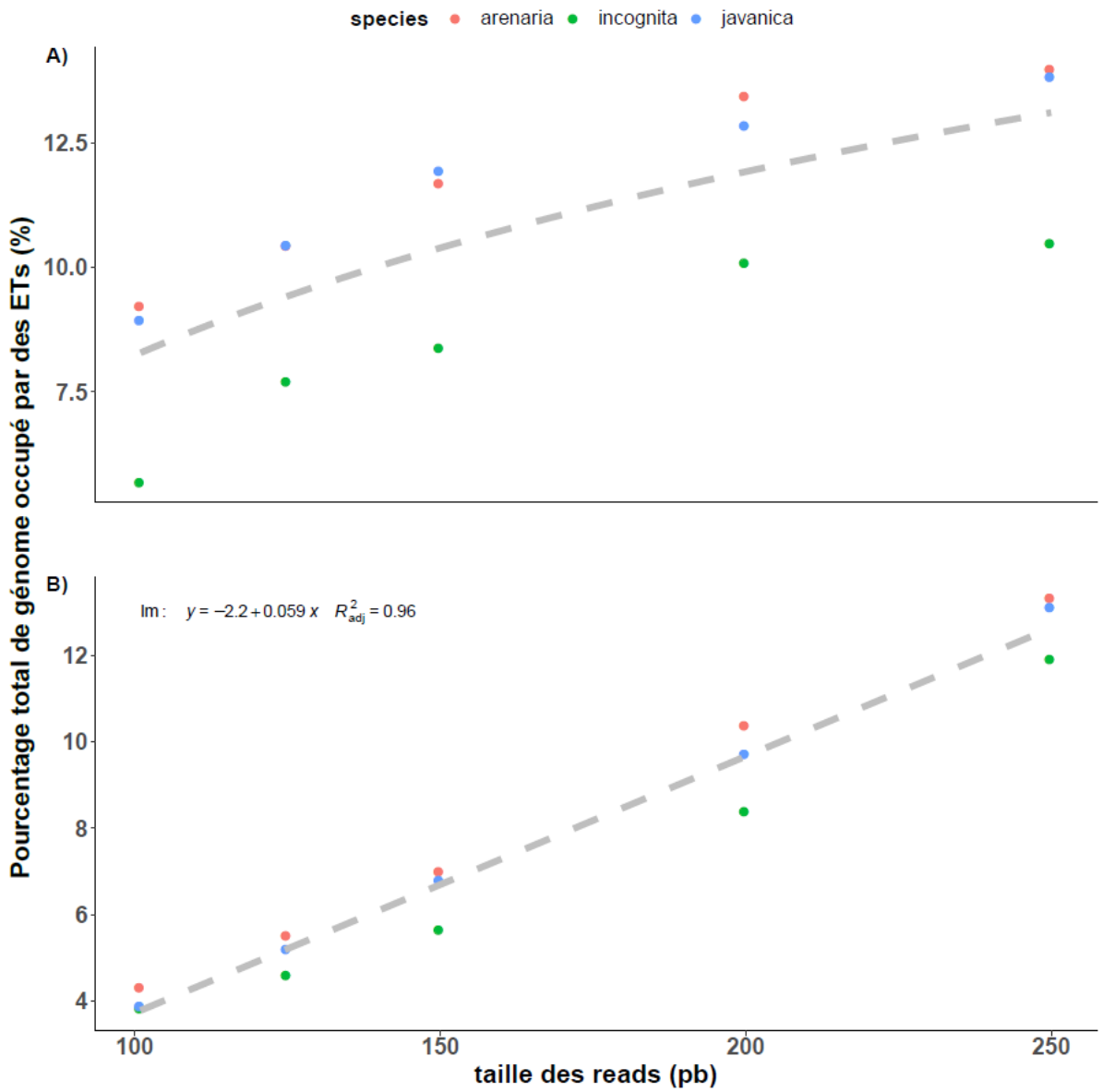


Figure 4.4.3 : lien entre charge totale d'ETs et longueur des reads par librairie

A - données réelles (librairies : arenariaV4, incognitaV4, javanicaV4).

La courbe grise représente la courbe d'équation y=log(x)

B - données simulées.

La courbe grise représente la droite d'équation y=0.059x 2.2.

Pour les deux équations, y représentant le pourcentage total de de génome occupé par des ETs dans le génome et x représentant la taille des reads contenus dans les librairies de données utilisées pour réaliser ces estimations.

Pour les données simulées (voir Figure 4.4.3-B), il existe une très forte corrélation positive entre la taille des reads et la charge totale en ETs (test de corrélation de Pearson ; R=0.983 & p-valeur=6.032e-11). La dynamique de cette relation peut être modélisée par l'équation linéaire y= 0.059x -2.2 où x correspond à la taille des reads dans une librairie donnée et y correspond au pourcentage d'ET dans le génome.

Pour les données réelles (voir Figure 4.4.3-A), il existe aussi un lien entre la taille des reads et la charge totale en ETs mais la nature de ce lien est logarithmique.

Que ce soit à partir de données réelles ou simulées, ces analyses révèlent que la prédiction de la charge d'ET par dnaPipeTE est fortement impactée par la taille des reads contenus dans la librairie. Plus les reads contenus dans une librairie sont longs et plus le pourcentage d'ETs prédit dans le génome correspondant est important.

La composition et la charge d'ETs au sein du genre *Meloidogyne* ne peuvent être reliées précisément à un trait biologique ou à un héritage phylogénétique distinct.

Compte tenu des biais précédemment identifiés, j'ai effectué une analyse comparative du contenu en ETs au sein des génomes des nématodes du genre *Meloidogyne* a partir de données homogènes en longueur entre librairies (longueur des reads comprise entre 95 et 101 pb pour l'ensemble des librairies, voir méthode) et en me plaçant à une couverture cible commune (0.25X, valeur choisie arbitrairement) (voir Figure 4.4.4 & Tableau annexe 4.6.2).

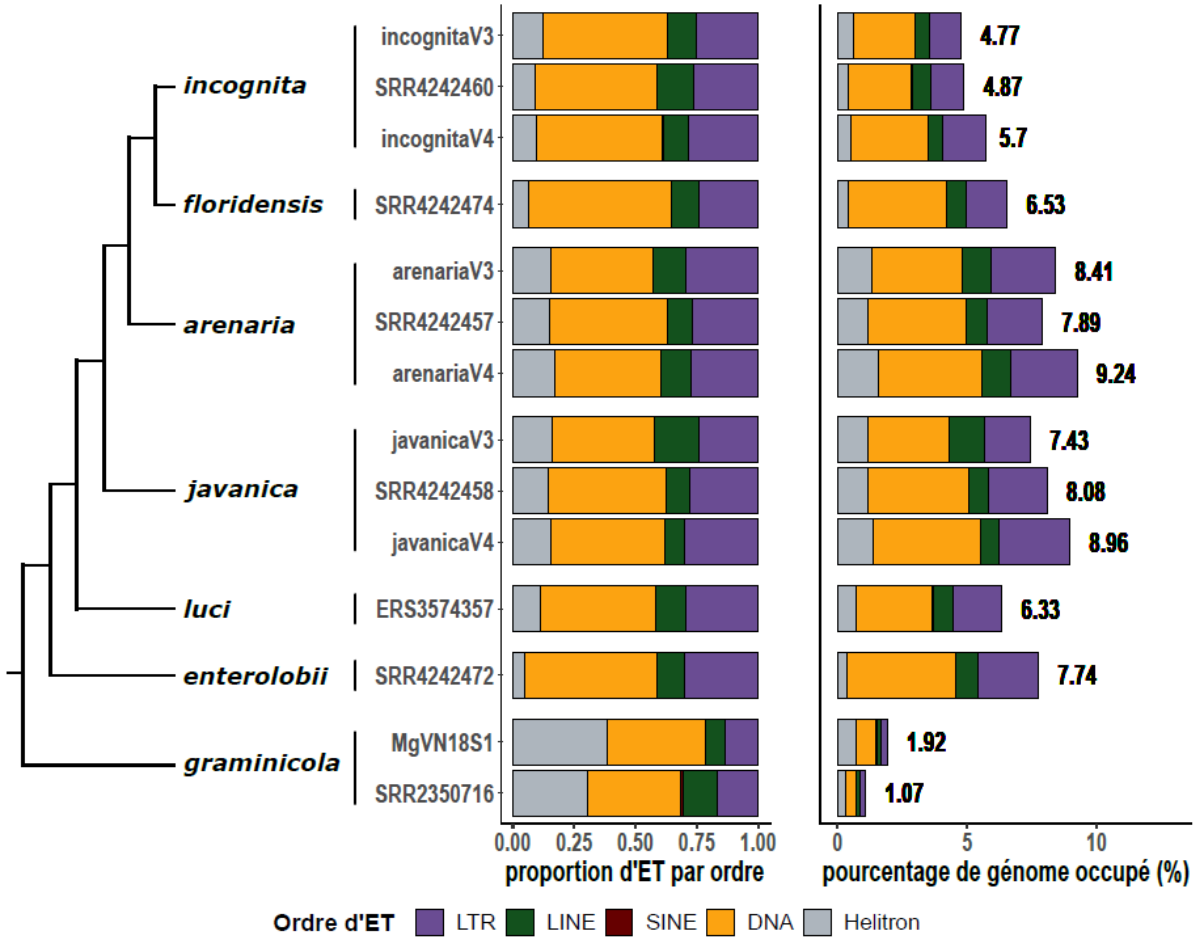


Figure 4.4.4 : composition et charge en ETs au sein des génomes d'espèces du genre *Meloidogyne*.

L'arbre représente la phylogénie des espèces de Meloidogyne étudiées, telle que décrite dans (Álvarez-Ortega et al., 2019). Le barplot central représente la proportion relative de chaque ordre d'ET. Le barplot de droite représente la charge d'ET par espèce et par ordre. Les valeurs données renseignent le pourcentage total du génome occupé par les ETs par librarie de reads utilisée.

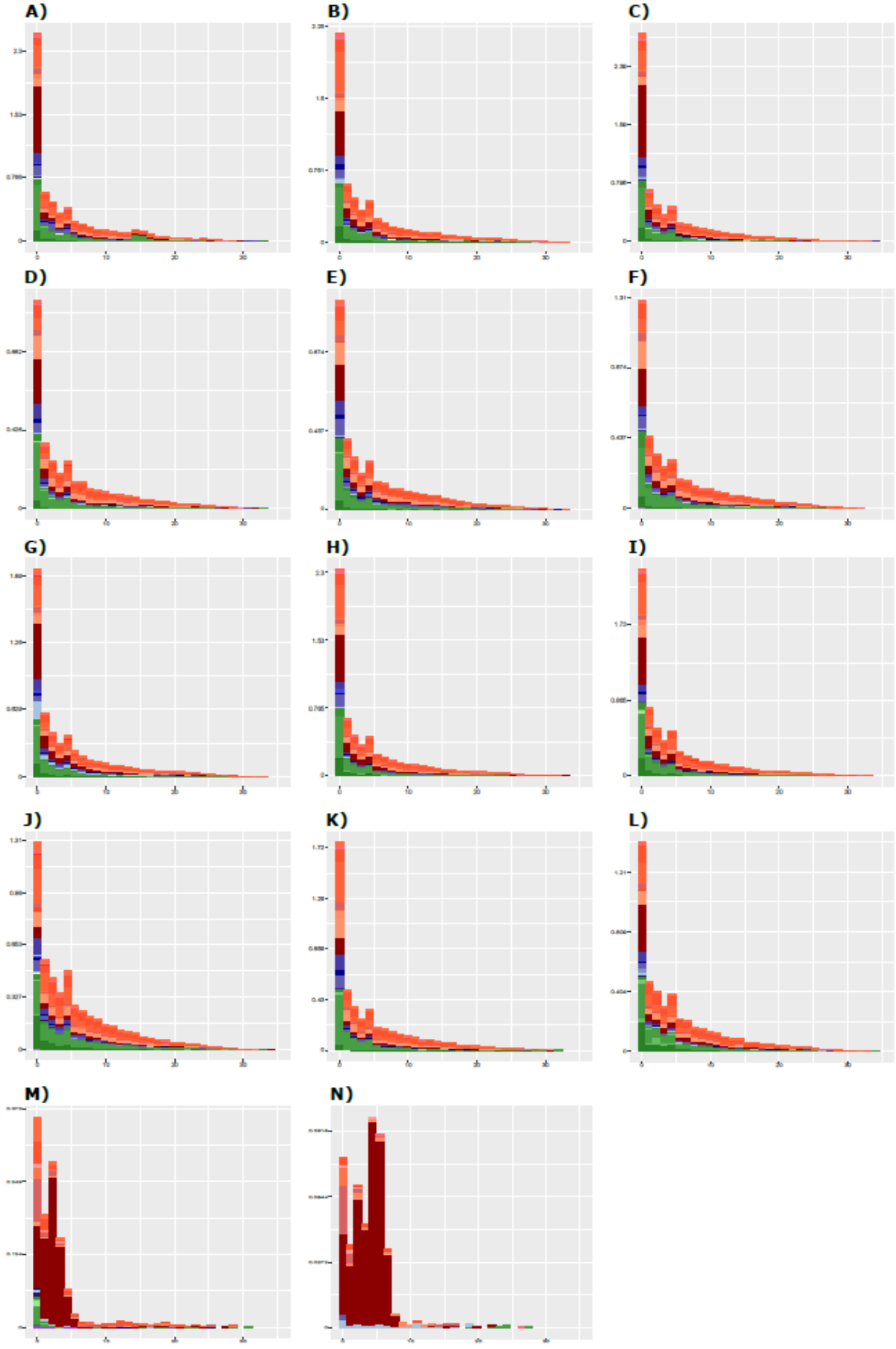
La charge en ETs varie entre espèces de ~1-2% du génome pour *M. graminicola* jusqu'à ~8-9% pour *M. arenaria* et *M. javanica*. Lorsque plusieurs réplicats sont disponibles pour une même espèce (*M. incognita*, *M. arenaria*, *M. javanica* et *M. graminicola*), on note une certaine variabilité entre ces réplicats. La seule espèce véritablement décrite comme diploïde et sexuée (*M. graminicola*) montre une charge en ETs nettement plus faible que toutes les autres espèces. Cependant, toutes les autres espèces sont phylogénétiquement très proches et forment un groupe monophylétique. En l'absence d'une autre espèce diploïde sexuée phylogénétiquement distincte, il est impossible à ce stade de conclure à une influence du mode de reproduction sur la charge en ETs. De même, d'après mes analyses de ploïdie toutes les espèces autres que *M. graminicola* seraient polyploïdes. Il serait donc tentant de conclure à une influence du niveau de ploïdie mais là encore, l'absence de réplicats phylogénétiquement distincts ne nous permet pas de conclure. Enfin, en observant l'arbre en Figure 4.4.4, il ne semble pas que les valeurs de charge en ETs suivent un patron particulier lié à un héritage phylogénétique. En revanche, il

existe une forte corrélation positive entre la charge en ETs et la taille des génomes des espèces de *Meloidogyne* étudiées (test de corrélation de Pearson ; R=0.948 & p-valeur=2.466e-07) (voir Figure annexe 4.6.1).

Bien que la répartition des ETs dans les différents ordres ne soit pas aléatoire (test du chi2; p-valeur 2.2e-16), cette répartition ne varie que légèrement entre les espèces de *Meloidogyne* du clade I, (*i.e.* toutes les espèces de *Meloidogyne* à l'exception de *M. graminicola*; phylogénie des espèces proposée dans (Álvarez-Ortega et al., 2019)). Il semble donc exister un profil de répartition des ETs commun aux *Meloidogyne* du clade I. Néanmoins, il est important de noter qu'il existe une certaine variabilité entre réplicats pour les espèces concernées. Il est donc nécessaire de considérer cette conclusion avec précaution. *M. graminicola* présente un profil de répartition différent avec notamment une plus forte abondance relative en Helitrons. *M. graminicola* et les espèces du clade 1 étant issus de lignées différentes, il est donc envisageable qu'une part de la composition en ETs décrite chez ces espèces soit la résultante de leur histoire évolutive.

La faible divergence entre lectures et les contigs d'ETs suggère une activité récente des ETs chez les Meloidogyne

Pour chaque librairie analysée, dnaPipeTE calcule la distribution de l'âge des ETs en utilisant la divergence entre les lectures et les contigs assemblés comme estimateur d'âge (voir Figure 4.4.5). Les ETs présentant plus de 30% de divergence entre les reads alignés et leur contig de consensus ne sont pas représentés.


A)
B)
C)
D)
E)
F)
G)
H)
I)
J)
K)
L)
M)
N)

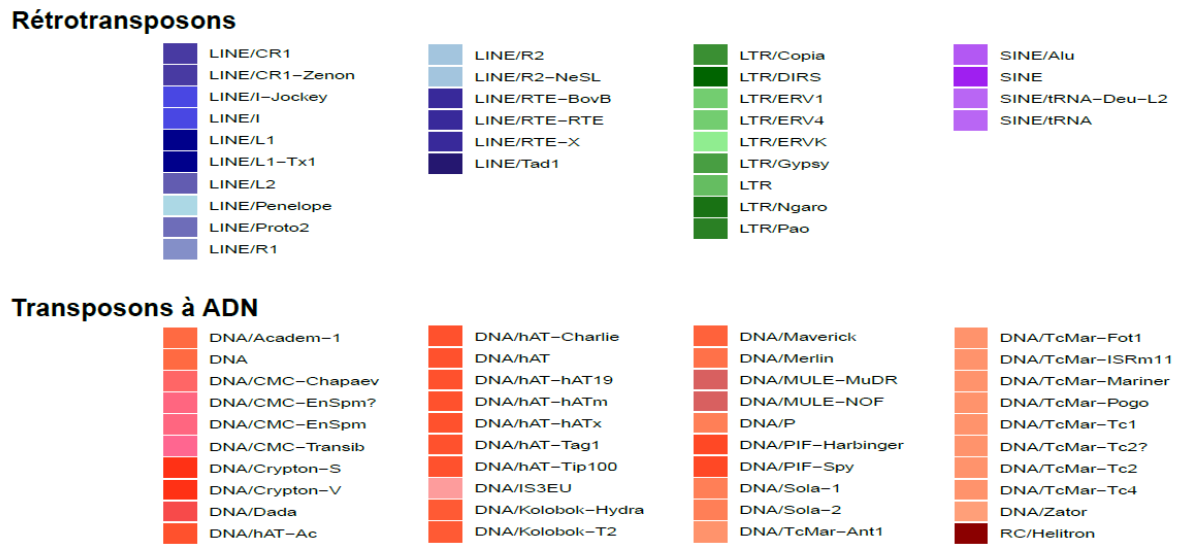


Figure 4.4.5 : Age des ETs au sein des génomes d'espèces du genre Meloidogyne.
Pour chaque librairie de séquençage, l'âge des ETs est estimé à partir de la divergenceentre les données de séquençages échantillonnées et les contigs consensus d'ETs sur lesquels elles s'alignent : ***arenaria*** *[A) arenariaV3, B) SRR4242457, C) arenariaV4];* ***incognita*** *[D) incognitaV3, E) SRR4242460, F) incognitaV4];* ***javanica*** *[G) javanicaV3, H) SRR4242458, I) javanicaV4] ;* ***enterolobii*** *[J) SRR4242472];* ***floridensis*** *[K) SRR4242474];* ***luci*** *[L) ERS3574357]; et* ***graminicola*** *[M) MgVN18S1, N) SRR2350716]*

L'ensemble des librairies analysées présentent une majorité d'ETs avec des lectures présentant un faible pourcentage de divergence avec les contigs de référence, ce qui peut être considérée comme un proxy de leur activité récente (Bast et al. 2015 ; Lerat et al. 2019). Bien que cette observation soit aussi valable pour elles, les deux librairies relatives à *M. graminicola* présentent néanmoins des profils de distributions différents des autres librairies étudiées (voir Figure 4.4.5 M & N). Pour ces deux librairies, la quasi-totalité des ETs détectés présentent moins de 10% de divergence, signe de l'absence ou presque d'ETs plus anciens.

Enfin, on notera que les transposons à ADN semblent être les plus actifs dans ces génomes, une majorité des copies récentes étant issues de superfamilles correspondant à ce type d'ET.

5 - Discussion

Dans cette analyse, nous avons pu identifier plusieurs facteurs influençant les résultats produits par le programme dnaPipeTE et qui pourraient donc constituer des biais .

Nous avons pu mettre en évidence que les résultats de dnaPipeTE sont sensibles à la couverture de sous-échantillonnage utilisée ainsi qu'aux variations de la longueur des reads que ce soit entre librairies différentes mais aussi au sein d'une même librairie (résultats non présentés ici). Le fait que la charge en ETs varie en fonction de ces deux paramètres rend impossible toute méta analyse sans uniformisation préalable de ces deux paramètres. De plus, cela ne permet pas de conclure quant à la charge réelle en ETs au sein de ces organismes. Seule une analyse de la charge en ETs relative au sein de ces espèces est possible, cette charge devant être interprétée avec précaution.

Nous avons tous de même pu remarquer qu'en fixant la couverture cible à 0.25x et en homogénéisant la taille des reads à 100 pb dans l'ensemble des librairies, les estimations de charges en ETs réalisées avec dnaPipeTE sur des librairies non assemblées sont comparables aux valeurs estimées avec REPET à partir des génomes assemblés, et ce pour trois espèces. La charge d'ET estimée chez *M. graminicola* avec ces paramètres (1.92% pour la librairie MgVN18S1) est en effet similaire à celle évaluée via une autre méthodologie (REPET) à partir de l'assemblage du génome réalisé sur les mêmes données brutes (2.64%) (Phan-Thi et al., 2020). La même observation peut être faite pour *M. incognita* (4.77% ici *vs* 4.67% à partir du génome assemblé (Kozlowski et al., 2020)) et pour *M. enterolobii* (7.74% ici *vs.* 8.37% à partir du génome assemblé (Koutsovoulos et al., 2020); cet assemblage étant issu d'autres données de séquençage).

On pourra noter que la similitude obtenue avec une autre méthodologie est aussi valable pour l'estimation de l'âge des ETs puisque le profil d'activité décrit dans cette analyse avec dnaPipeTE pour *M. graminicola* avec ces valeurs de paramètres est très similaire à celui présenté dans le chapitre IV (voir Figure annexe 2). La même observation peut être faite pour les espèces *M. enterolobii* et *M. incognita*. Les résultats obtenus dans la présente analyse (charge d'ET, et profils d'activité/âge des ETs) avec ce paramétrage sont donc cohérents avec ceux obtenus selon d'autres méthodologies. Ce paramétrage de l'outil dnaPipeTE semble donc réduire les biais méthodologiques identifiés.

Il serait intéressant dans le futur de confirmer si les biais identifiés à partir des données chez *Meloidogyne* se confirment sur d'autres espèces. Si c'est le cas, il serait important de communiquer à propos de ces biais auprès des auteurs et des utilisateurs de dnaPipeTE. Par ailleurs, il serait également intéressant d'estimer si le lien entre charge en ETs estimée et couverture de sous-échantillonnage atteint un plateau à partir d'une certaine valeur ou continue à varier.

En limitant l'impact des biais méthodologiques identifiés, il a été possible de comparer la charge et la composition en ETs entre les espèces de *Meloidogyne*.

Il en ressort que la charge en ETs varie de manière parfois importante entre ces espèces. On notera néanmoins que *M. graminicola*, la seule espèce étudiée dans la présente analyse à être issue d'un autre groupe d'espèce (clade III de la phylogénie issue de (Álvarez-Ortega et al., 2019), toutes les autres espèces appartenant au clade I) et aussi la seule espèce diploïde sexuée avérée, présente une charge en ETs bien inférieure à celle des autres espèces, et ce pour les deux "réplicats" étudiés.

Il semble par ailleurs exister un paysage d'ETs (i.e. composition) commun aux espèces de *Meloidogyne* du clade I (i.e. ensemble des espèces sauf *M. graminicola*). La diversité et la proportion relative des ETs au sein de ces espèces pourraient donc avoir été héritées d'un ancêtre commun dans chaque clade puis avoir varié en charge au gré d'une activité propre à chaque espèce.

Il n'est cependant pas possible de conclure à un lien entre mode de reproduction ou niveau ploïdie et la charge ou la composition en ET, *M. graminicola* étant la seule espèce véritablement décrite comme capable de se reproduire sexuellement et étant diploïde. Toutes les autres espèces sont polyploïdes et, excepté *M. floridensis*, sont parthénogénétiques mitotiques. De plus, comme évoqué plus haut, elles forment un seul groupe monophylétique (clade I). Seul l'ajout dans le futur d'autres espèces à reproduction sexuée vs asexuée ; polyploïdes vs diploïdes et méiotiques vs. mitotiques dans les différents clades de *Meloidogyne* permettra à terme de conclure. Il serait aussi intéressant de réaliser le même type d'analyse chez d'autres groupes de nématodes phytoparasites aussi diversifiés en termes de traits biologiques afin d'évaluer si un patron spécifique à ce mode de vie peut être mis en avant.

Dans cette analyse, nous avons aussi pu montrer que les ETs, et en particulier les transposons à ADN, ont vraisemblablement été actifs au sein des génomes des espèces de *Meloidogyne*. La possible activité récente des ETs observée est cohérente avec une hausse d'activité des ETs lors d'évènements d'hybridation (choc génomique initialement proposé par Barbara McClintock). En effet, tous les nématodes du clade I étudiés ici (*M. arenaria, javanica, incognita, floridensis* et *enterolobii*) sont probablement des espèces hybrides (Blanc-Mathieu et al., 2017; Koutsovoulos et al., 2019, Szitenberg et al. 2017) et leurs profils d'activité/d'âge des ETs sont très similaires. *M. graminicola* est la seule espèce à présenter un profil d'activité/âge des ETs différent. Chez *M. graminicola*, l'extrême majorité des copies d'ETs sont hautement identiques à leur consensus. Un tel profil pourrait être dû à une explosion récente du nombre d'ET, cependant la charge en ET est plus faible chez *M. graminicola* que chez toutes les autres espèces. Donc, il pourrait s'agir au contraire d'une très forte répression qui ferait qu'il n'existe que quelques copies actives qui sont très rapidement contrôlées et éliminées du génome.

M. graminicola étant une espèce parthénogénétique facultative capable de faire de la recombinaison méiotique, il est possible que les ETs soient éliminés plus facilement du génome.

6 - Annexes

Tableau annexe 4.6.1 : statistiques des librairies de reads après pré-traitement des données ("Quality trimming" et longueur minimale de 95 pb).

Les librairies ERR1212566 et enterolobiiV3 présentent des signes de contamination et ont été retirées de l'analyse.

espèce	source [local, DB]	num. accès / version	nb. de reads	longueur des reads	%GC	pic de longeur	contient des Ns [Y, N]	signe de contamination (a partir != estimation GC) [Y,M,N]	debut pic qualité (Phred Score)	contient des adaptateurs [Y, N]
arenaria	DB	SRR4242457	37524005	95-125	29	125	N	N	34	N
arenaria	DB	SRR4242457	37506237	95-125	29	125	N	N	34	N
arenaria	local	arenariaV3	688809639	95-101	28	101	N	N	32	N
arenaria	local	arenariaV3	65712338	95-101	28	101	N	N	32	N
arenaria	local	arenariaV4	44487122	95-251	30	95-251	N	N	33	N
arenaria	local	arenariaV4	41852786	95-251	30	95-251	N	N	33	N
javanica	DB	SRR4242458	112984513	95-125	29	125	N	N	34	N
javanica	DB	SRR4242458	115940014	95-125	29	125	N	N	34	N
javanica	local	javanicaV3	54969315	95-101	28	101	N	N	32	N
javanica	local	javanicaV3	50476288	95-101	28	101	N	N	32	N
javanica	local	javanicaV4	58267856	95-251	31	95-251	N	N	34	N
javanica	local	javanicaV4	56506916	95-251	31	95-251	N	N	34	N
incognita	DB	SRR4242460	22788462	95-125	28	125	N	N	33	N
incognita	DB	SRR4242460	21623275	95-125	27	125	N	N	33	N
incognita	local	incognitaV3	64840795	95-101	28	101	N	N	32	N
incognita	local	incognitaV3	62109784	95-101	28	101	N	N	32	N
incognita	local	incognitaV4	32375230	95-251	30	95-251	N	N	33	N
incognita	local	incognitaV4	29859655	95-251	30	95-251	N	N	33	N
enterolobii	DB	SRR4242472	118137682	95-125	29	125	N	N	33	N
enterolobii	DB	SRR4242472	97683725	95-125	29	125	N	N	33	N
enterolobii	local	enterolobiiV3	147643003	95-101	33	101	N	Y	32	N
enterolobii	local	enterolobiiV3	139423886	95-101	32	101	N	Y	32	N
fonidensis	DB	SRR4242474	84788534	95-150	30	95-150	N	N	34	N
fonidensis	DB	SRR4242474	616222738	95-150	30	95-150	N	N	34	N
graminicola	DB	SRR2350716	26705835	95-100	27	100	N	N	32	N
graminicola	DB	SRR2350716	27325736	95-100	27	100	N	N	32	N
graminicola	local/DB	MgVN18S1	7116578	95-100	24	100	N	N	32	N
graminicola	local/DB	MgVN18S1	7206394	95-100	24	100	N	N	32	N
chitwoodi	DB	ERR1212566	18558197	95-301	23	95-301	N	Y	33	N
chitwoodi	DB	ERR1212566	16053804	95-301	22	95-301	N	Y	33	N
luci	DB	ERS3574357	174107045	95-150	29	147	N	N	34	N
luci	DB	ERS3574357	135756880	95-150	28	147	N	N	34	N

Tableau annexe 4.6.2 : composition et charge d'ETs dans les génomes des espèces de *Meloidogyne* (données homogénéisées en longueur : 95-101 pb)

Les valeurs expriment le pourcentage de génome occupé par chaque type de répétition et ce pour chaque librairie analysée. Ces valeurs ont été obtenues avec dnaPipeTE à partir d'un sous échantillonnage à 0.25X de données homogénéisées en longueur (95-101 pb).

	M. arenaria			***M. incognita***			***M. floridensis***
Ordre	Arenaria V3	SRR4242457	Arenaria V4	Incognita V3	SRR4242460	Incognita V4	SRR4242474
LTR	2.230	2.761	4.598	1.157	1.373	2.790	2.342
LINE	1.069	1.145	1.266	0.605	0.842	1.031	0.995
SINE	0.002	0.012	0.037	0.010	0.028	0.033	0.025
DNA	3.973	3.688	6.234	2.586	3.251	5.628	5.053
Helitron	1.229	1.211	1.856	0.694	0.725	1.019	0.893
MITE	0	0	0	0	0	0	0
na	11.311	15.897	13.334	10.279	11.566	11.931	14.818
others	1.391	1.303	1.884	1.510	1.419	1.705	1.018
rRNA	0.658	0.854	0.822	0.454	0.402	0.603	1.162
Satellite	0.165	0.247	0.219	0.519	0.423	0.463	0.064
Simple_repeat	3.896	6.551	11.206	2.705	4.454	8.823	6.810
Tandem_repeats	0	0	0	0	0	0	0
Low_Complexity	1.604	2.731	4.833	1.024	1.567	3.642	2.594

	M. javanica			*M. graminicola*		*M. luci*	*M. enterolobii*
Ordre	Javanica V3	SRR4242458	Javanica V4	MgVN18S1	SRR2350716	ERS3574357	SRR4242472
LTR	2.337	3.052	4.434	0.256	0.178	2.349	2.833
LINE	0.663	0.939	1.348	0.116	0.168	0.840	1.103
SINE	0.026	0.008	0.034	0.000	0.006	0.007	0.005
DNA	3.608	4.184	6.403	1.745	0.388	4.149	6.624
Helitron	1.259	1.256	1.618	0.679	0.313	0.729	0.466
MITE	0	0	0	0	0	0	0
na	11.827	15.128	12.904	5.041	2.860	13.086	14.339
others	1.002	0.983	1.168	0.039	0.440	1.975	1.022
rRNA	0.701	0.879	0.912	0.271	0.293	0.952	0.627
Satellite	0.231	0.246	0.301	0.005	0.017	0.265	0.294
Simple_repeat	4.175	6.293	10.232	2.127	1.661	6.965	6.637
Tandem_repeats	0	0	0	0	0	0	0
Low_Complexity	1.344	2.359	4.066	0.782	0.253	2.532	2.483

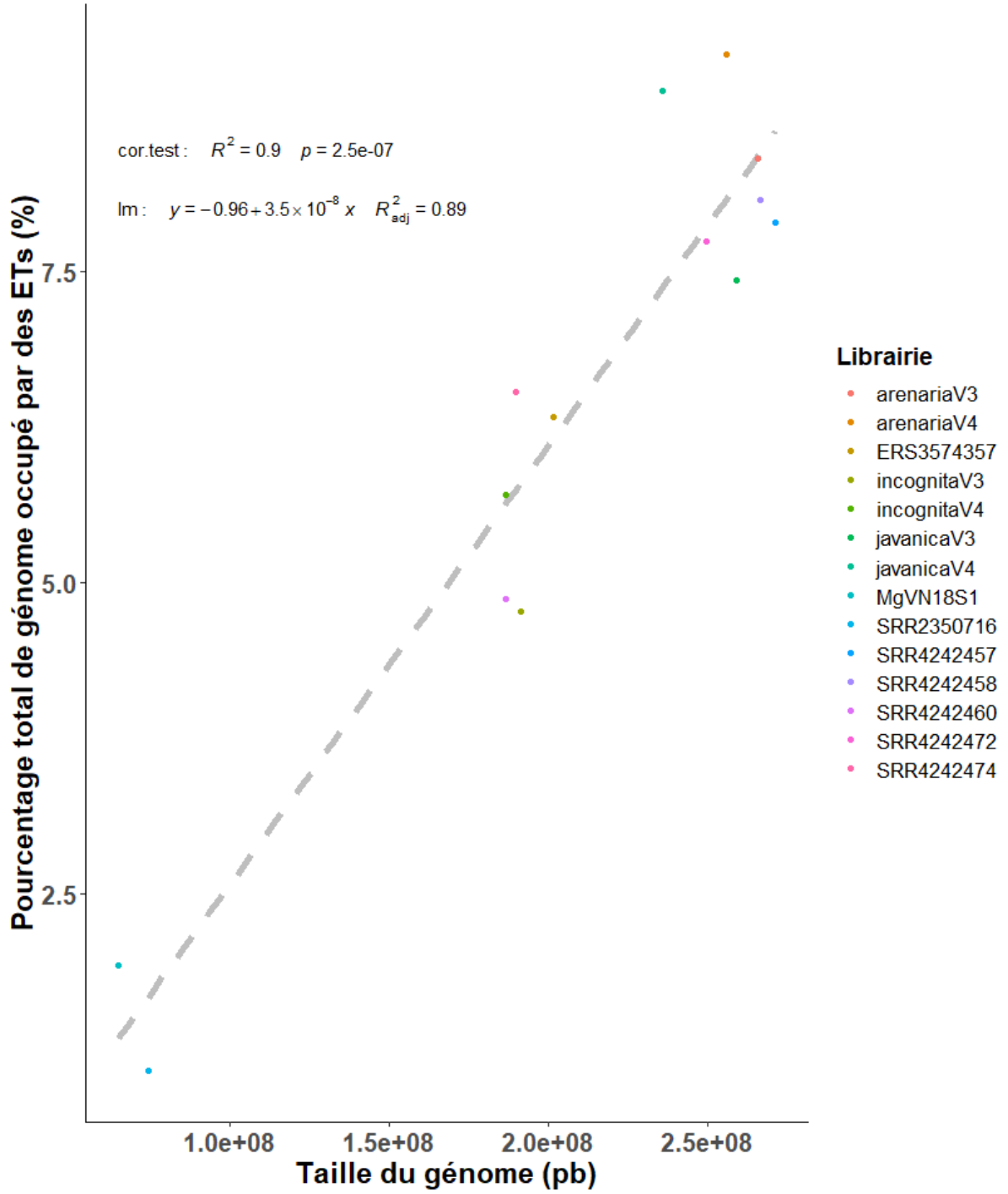


Figure annexe 4.6.1 : relation entre la charge d'ETs et la taille du génome par librairie.

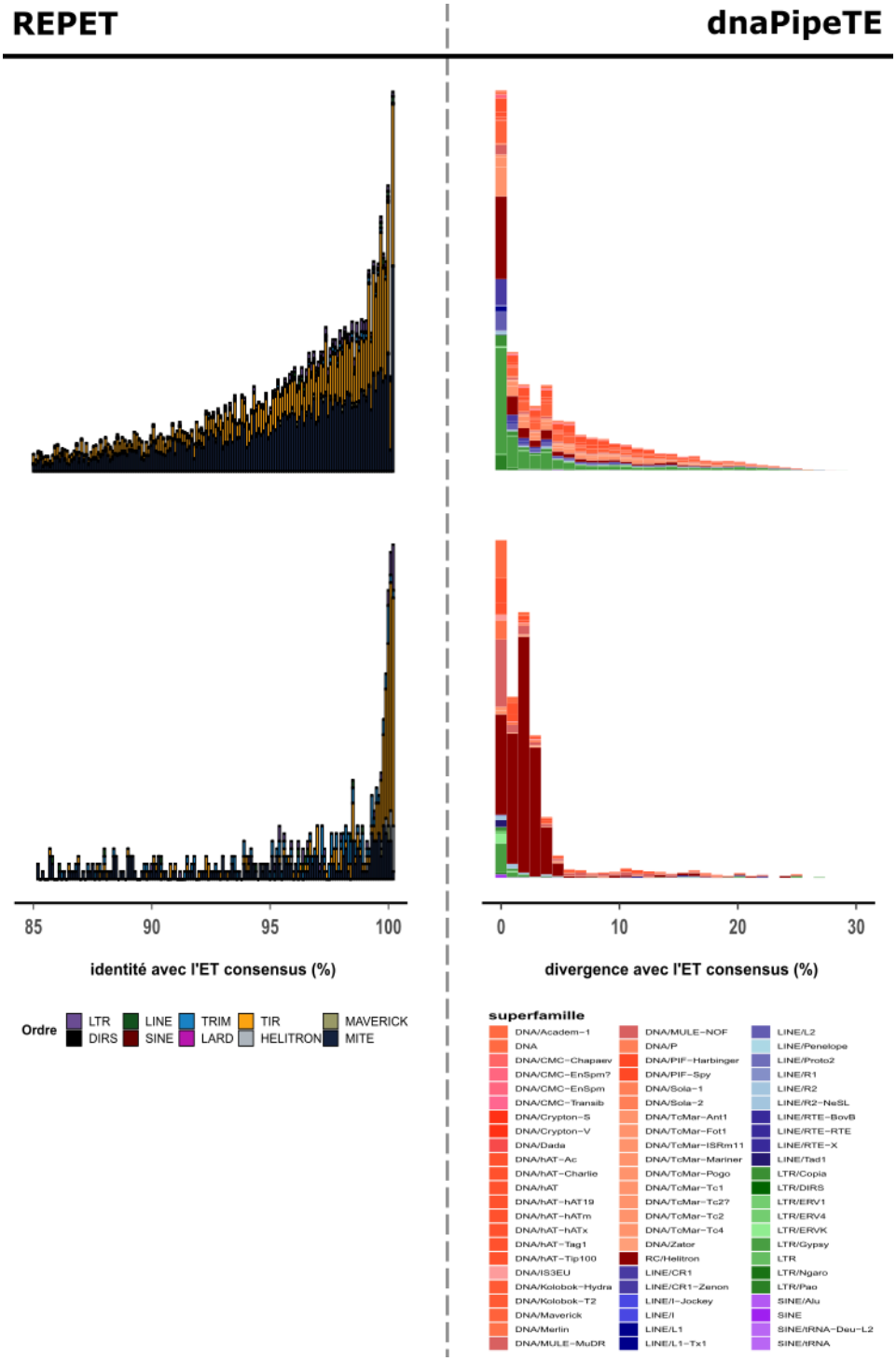


Figure annexe 4.6.2 : âge des ETs, comparaison des résultats obtenus via 2 méthodes.
Le panel de gauche représente le taux d'identité par copie avec leur consensus pour M.incognita (genome assemblé à partir de la librairie incognitaV3) en haut et pour M. graminicola (génome assemblé à partir de la librairie MgVN12S1) en bas. Ces résultats ont été obtenus à partir d'une analyse réalisée avec REPET.
Le panel de droite représente le taux de divergence par copie avec leur consensus pour M.incognita (librairie incognitaV3) en haut et pour M. graminicola (librairie MgVN12S1) en bas. Ces résultats ont été obtenus à partir d'une analyse réalisée avec dnaPipeTE.

VI – Activité des ETs au sein d'une espèce de *Meloidogyne* : le cas de *M.incognita*

1 - Avant-propos

L'article présenté ci-dessous est disponible dans la boîte de dépôt BioRxiv à l'adresse suivante : https://doi.org/10.1101/2020.04.30.069948 . Cet article a été évalué et recommandé par Peer Community In (PCI) Evolutionary Biology. Le résumé de cette recommandation est visible à l'adresse suivante : 10.24072/pci.evolbiol.100106 . Les suppléments sont disponibles en annexes (chapitre X-B).

Contribution(s) des auteurs :

Djampa KOZLOWSKI – conception générale de l'étude, conception des validations expérimentales, analyses bio-informatiques, rédaction de l'article.

Rahim HASSANALY-GOULAMHOUSSEN - conception des validations expérimentales, validations expérimentales, rédaction partielle de l'article.

Martine DA-ROCHA & Georgios KOUTSOVOULOS – analyses bio-informatiques.

Marc BAILLY-BECHET & Etienne GJ DANCHIN - conception générale de l'étude, rédaction de l'article, encadrement.

2 - Contexte

Au travers des résultats précédents, nous avons pu constater que les ETs ont probablement été récemment actifs au sein des génomes des espèces de *Meloidogyne*. Le fait que l'on trouve une majorité de jeunes copies indique que des ETs se sont récemment multipliés et déplacés dans ces génomes. De ce fait, les ET ont donc probablement participé à la plasticité génomique de ces espèces et il est envisageable que cette activité ait eu des répercussions fonctionnelles. Dans cette optique, j'ai concentré mes efforts sur *M. incognita*, l'espèce à reproduction asexuée la plus préjudiciable pour l'agriculture (Trudgill and Blok, 2001).

Après avoir annoté en détail le contenu en ETs de cette espèce, j'ai réalisé une analyse populationnelle entre 12 isolats géographiques présentant des variations de gammes d'hôtes (plantes pouvant être

infectées) afin de i) mieux caractériser l'activité des ETs au sein de cette espèce, et ii) définir si l'activité des ETs pouvait être mise en relation avec des variations de compatibilité d'hôtes entre ces isolats.

Les résultats obtenus sont présentés en détail dans l'article ci-dessous.

3 - Article

Peer Community In Evolutionary Biology

RESEARCH ARTICLE

Open Access
Open Data
Open Code
Open Peer-Review

Transposable Elements are an evolutionary force shaping genomic plasticity in the parthenogenetic root-knot nematode *Meloidogyne incognita*

Djampa KL Kozlowski[1], Rahim Hassanaly-Goulamhoussen[1], Martine Da Rocha[1], Georgios D Koutsovoulos[1], Marc Bailly-Bechet[1]*, Etienne GJ Danchin[1]*.

[1] Université Côte d'Azur, INRAE, CNRS, ISA – Sophia Antipolis, France
* equal contribution

This article has been peer-reviewed and recommended by
Peer Community in Evolutionary Biology
https://doi.org/10.24072/pci.evolbiol.100106

Cite as: Kozlowski DK, Hassanaly-Goulamhoussen R, Da Rocha M, Koutsovoulos GD, Bailly-Bechet M, Danchin EG (2020) Transposable Elements are an evolutionary force shaping genomic plasticity in the parthenogenetic root-knot nematode *Meloidogyne incognita*. bioRxiv, 2020.04.30.069948, ver. 4 peer-reviewed and recommended by PCI Evolutionary Biology. https://doi.org/10.1101/2020.04.30.069948

Posted: 03 Aug 2020

Recommender: Inés Alvarez

Reviewers: Daniel Vitales and two anonymous reviewers

Correspondence:
etienne.danchin@inrae.fr
djampa.kozlowski@outlook.com

ABSTRACT

Despite reproducing without sexual recombination, the root-knot nematode *Meloidogyne incognita* is adaptive and versatile. Indeed, this species displays a global distribution, is able to parasitize a large range of plants and can overcome plant resistance in a few generations. The mechanisms underlying this adaptability without sex remain poorly known and only low variation at the single nucleotide polymorphism level have been observed so far across different geographical isolates with distinct ranges of compatible hosts. Hence, other mechanisms than the accumulation of point mutations are probably involved in the genomic dynamics and plasticity necessary for adaptability. Transposable elements (TEs), by their repetitive nature and mobility, can passively and actively impact the genome dynamics. This is particularly expected in polyploid hybrid genomes such as the one of *M. incognita*. Here, we have annotated the TE content of *M. incognita*, analyzed the statistical properties of this TE content, and used population genomics approach to estimate the mobility of these TEs across 12 geographical isolates, presenting phenotypic variations. The TE content is more abundant in DNA transposons and the distribution of TE copies identity to their consensuses sequence suggests they have been at least recently active. We have identified loci in the genome where the frequencies of presence of a TE showed variations across the different isolates. Compared to the *M. incognita* reference genome, we detected the insertion of some TEs either within genic regions or in the upstream regulatory regions. These predicted TEs insertions might thus have a functional impact. We validated by PCR the insertion of some of these TEs, confirming TE movements probably play a role in the genome plasticity with possible functional impacts.

Keywords: *transposons, genomic plasticity, evolution, agricultural pest, parthenogenesis, hybridization*

Introduction

Agricultural pests cause substantial yield loss to the worldwide life-sustaining production (Savary et al. 2019) and threaten the survival of different communities in developing countries. With a constantly growing human population, it becomes more and more crucial to reduce the loss caused by these pests while limiting the impact on the environment. In this context, understanding how pests evolve and adapt both to the control methods deployed against them and to a changing environment is essential. Among Metazoa, nematodes and insects are the most destructive agricultural pests. Nematodes alone are responsible for crop yield losses of ca. 11% which represents up to 100 billion € economic loss annually (Agrios 2005; McCarter 2009). The most problematic nematodes to worldwide agriculture belong to the genus Meloidogyne (Jones et al. 2013) and are commonly named root-knot nematodes (RKN) owing to the gall symptoms their infection leaves on the roots. Curiously, the RKN species showing the wider geographical distribution and infecting the broadest diversity of plants reproduce asexually via mitotic parthenogenesis (Trudgill and Blok 2001; Castagnone-Sereno and Danchin 2014). In the absence of sexual recombination, the genomes are supposed to irreversibly accumulate deleterious mutations, the efficiency of selection is reduced due to linkage between conflicting alleles while the combination of beneficial alleles from different individuals is impossible (Muller 1964; Hill and Robertson 1966; Kondrashov 1988; Glémin et al. 2019). For these reasons, asexual reproduction is considered an evolutionary dead end and is actually quite rare in animals (Rice 2002). In this perspective, the parasitic success of the parthenogenetic RKN might represent an evolutionary paradox.

Previous comparative genomics analyses have shown the genomes of the most devastating RKN are polyploid as a result of hybridization events (Blanc-Mathieu et al. 2017; Szitenberg et al. 2017). In the parthenogenetic RKN *M. incognita*, the gene copies resulting from allopolyploidy not only diverge at the nucleotide level but also in their expression patterns, suggesting this peculiar genome structure could support a diversity of functions and might be involved in the parasitic success despite the absence of sexual reproduction (Blanc-Mathieu et al. 2017). This hypothesis seems consistent with the 'general purpose genotype' concept, which proposes successful parthenogens have a generalist genotype with good fitness in a variety of environments (Vrijenhoek and Parker 2009). An alternative non mutually exclusive hypothesis is the 'frozen niche variation' concept which proposes parthenogens are successful in stable environments because they have a frozen genotype adapted to this specific environment (Vrijenhoek and Parker 2009). Interestingly, the frequency of parthenogenetic invertebrates is higher in agricultural pests, probably because the anthropized environments in which they live are more stable and uniform (Hoffmann et al. 2008).

However, although a general purpose genotype brought by hybridization might contribute to the wide host range and geographical distribution of these parthenogenetic RKNs, this alone, cannot explain how these species evolve and adapt to new hosts or environments without sex. For instance, initially, avirulent populations of some of these RKN, controlled by a resistance gene in a tomato, are able to overcome the plant resistance in a few generations, leading to virulent sub-populations, in controlled laboratory experiments (Castagnone-Sereno et al. 1994; Castagnone-Sereno 2006). Emergence of virulent populations, not controlled anymore by resistance genes have also been reported in the field (Barbary et al. 2015).

The mechanisms underlying the adaptability of parthenogenetic RKN without sex remain elusive. Recent population genomics analyses showed that only a few single nucleotide variations (SNV) could be identified by comparing different Brazilian *M. incognita* isolates showing distinct ranges of host compatibility (Koutsovoulos

et al. 2020). Addition of further isolates from different geographical locations across the world did not substantially expand the number of variable positions in the genome. Furthermore, the few identified SNV showed no significant correlation with either the geographical location, the host range or the currently infected crop species. However, these SNV could be used as markers to confirm the absence of sexual meiotic recombination in *M. incognita*. Thus, the low nucleotide variability that was observed between isolates is probably not the main driver of the genomic plasticity underlying the adaptability of *M. incognita*.

Consistent with these views, convergent gene copy number variations were observed following resistance breaking down by two originally avirulent populations of *M. incognita* from distinct geographic origins (Castagnone-Sereno et al. 2019). The mechanisms supporting these gene copy numbers and other genomic variations possibly involved in the adaptive evolution of *M. incognita* remain to be described.

Transposable elements (TEs), by their repetitive and mobile nature, can both passively and actively impact genome plasticity. Being repetitive, they can be involved in illegitimate genomic rearrangements leading to loss of genomic portions or expansion of gene copy numbers. Being mobile, they can insert in coding or regulatory regions and have a functional impact on the gene expression or gene structure / function itself. For instance, TE neo-insertions have been shown to affect gene expression in a species-specific manner in amniotes (Zeng et al. 2018) and, in rodents, TE insertions account for ca. 20% of gene expression profile divergence between mice and rats (Pereira et al. 2009). At shorter evolutionary scales, differential presence / absence of TE across *Arabidopsis* populations revealed rare variants associated with extremes of gene expression (Stuart et al. 2016). TE insertions in coding regions can disrupt a gene and this disruption might eventually have an adaptive effect. For example, a TE insertion has caused disruption of a Phytochrome A gene in some soybean strains, which caused photoperiod insensitivity and was in turn associated with adaptation to high latitudes in Japan (Kanazawa et al. 2009). Moreover, in *Drosophila*, insertion of a TE in the *CHKov1* gene caused four new alternative transcripts and this modification is associated with resistance to insecticide and viral infection (Aminetzach et al. 2005; Magwire et al. 2011). In parallel, although TE movements can provide beneficial genomic novelty or plasticity, their uncontrolled activity can also be highly detrimental and put the organism at risk. For instance, some human diseases such as hemophilia (Kazazian et al. 1988) or cancers (Miki et al. 1992) are caused by TE insertions in coding or regulatory regions.

Concerning agricultural pests themselves, TEs are a major player of adaptive genome evolution by both passively and actively impacting the genome structure and sequence in some fungal phytopathogens (Faino et al. 2016). Whether TEs also play an important role in the genome plasticity and possibly adaptive evolution of parasitic animals, engaged in a continuous arms race with their hosts, remains poorly known. According to the Red Queen hypothesis, host-parasites arms race is a major justification for the prevalence of otherwise costly sexual reproduction (Lively 2010) and, in the absence of sex, other mechanisms should provide the necessary plasticity to sustain this arms race.

From an evolutionary point of view, the parthenogenetic root-knot nematode *M. incognita* represents an interesting model to study the activity of TEs and their impact on the genome, including in coding or regulatory regions. Indeed, being a plant parasite, *M. incognita* is engaged in an arms race with the plant defence systems and point mutations alone are not expected to be a major mechanism supporting adaptation in this species (Koutsovoulos et al. 2020).

In a broader perspective, little is known yet about the TE dynamics in nematode genomes and their possible impact on adaptive evolution, including in the model *C. elegans*, despite being the first sequenced animal

genome since 1998 (The C. elegans Genome Sequencing Consortium 1998). Transposition activity of Tc1 TIR element was shown to be positively linked to the overall mutation rate in *C. elegans* mutator strains, one of which is characterized by high transposition in the germline, hence constituting a considerable evolutionary force (Bégin and Schoen 2007). However, these results may be hindered by the fact that, in wild-type *C. elegans*, although Tc1 excision frequency is substantial in somatic cells, it is negligible in the germ-cells (Emmons and Yesner 1984).

Besides Tc1, a more comprehensive analysis using population genomics approach in *C. elegans* represents the most advanced study of the TE dynamics in this species to date (Laricchia et al. 2017). By analyzing hundreds of wild populations of *C. elegans*, the authors have shown a substantial level of activity for multiple families of TEs in these genomes compared to the N2 reference strain. The study points at a population-wide variability of this activity, and, surprisingly, towards little evident phenotypic effect of this activity, even when TEs were found inserted into coding sequences. Concerning the possible functional impact of TE activity in nematodes, an investigation of TE expression in *C. elegans* germline in a single cell framework has shown significant differences between the expression pattern of LTR, non-LTR elements and DNA TE, associated with differentiated vs. undifferentiated cell types (Ansaloni et al. 2019). These complex cell-type specific differential expression patterns suggest TE activity plays an important role in the *C. elegans* embryonic development, although the exact role remains elusive. Overall, while it is now clearly established that TE are active in *C. elegans* and probably contribute to the genome plasticity, their possible functional implication or role in nematode adaptive evolution has not been shown so far.

In this study, we have tested whether the TE activity could represent a mechanism supporting genome plasticity in *M. incognita*, a prerequisite for adaptive evolution. We have re-annotated the 185Mb triploid genome of *M. incognita* (Blanc-Mathieu et al. 2017) for TEs, using stringent filters to identify canonical TEs, possibly active in the genome. We analyzed the statistical properties of the TE content and the distribution of TE sequence identity levels to their consensuses was used as a reporter of the recentness of their activity. We have then tested whether the frequencies of presence/absence of these TEs across the genome varied between different isolates. To test for variations in frequencies, we have used population genomics data from eleven *M. incognita* isolates collected on different crops and locations and differing in their ranges of compatible hosts (Koutsovoulos et al. 2020). From the set of TE loci that presented the most contrasted patterns of presence/absence across the isolates, we investigated whether some could represent neo-insertions. To estimate the possible functional impact of TE insertions, we checked whether some were inserted within coding or possible regulatory regions. Finally, we validated by PCR assays some of these neo-insertions in coding or regulatory regions, predicted by population genomics data. Overall, our study represents the first estimation of TE activity as a mechanism possibly involved in the genome plasticity and the associated functional impact in the most devastating nematode to worldwide agriculture. Besides *C. elegans*, little was known about the role of TE in the genome dynamics of *Nematoda*, one of the most species-rich animal phylum. Because this study focuses on an allopolyploid and parthenogenetic animal species, it also opens new evolutionary perspectives on the fate and potential adaptive impact of TEs in these singular organisms.

Results

The *M. incognita* TE landscape is diversified but mostly composed of DNA transposons.

We used the REPET pipeline (Quesneville et al. 2005; Flutre et al. 2011) to predict and annotate the *M. incognita* repeatome (see methods). Here, we define the repeatome as all the repeated sequences in the genome, excluding Simple Sequence Repeats (SSR or microsatellites). The repeatome spans 26.38 % of the *M. incognita* genome length (sup.Table S1). As we wanted to assess whether TEs actively contributed to genomic plasticity, we applied a series of stringent filters on the whole repeatome to retain only repetitive elements presenting canonical signatures of TEs (see methods and (Kozlowski 2020a)). We identified 480 different TE-consensus sequences that allowed annotation of 9,633 canonical TE, spanning 4.67% of the genome (Table 1). Both retro (Class I) and DNA (Class II) transposons (Wicker et al. 2007) compose the *M. incognita* TE landscape with 5/7 and 4/5 of the known TE orders represented respectively, showing a great diversity of elements (Fig 1). Canonical retro-transposons and DNA-transposons respectively cover 0.90 and 3.77 % of the genome. Terminal Inverted Repeats (TIR) and Miniature Inverted repeat Transposable Elements (MITEs) DNA-transposons alone represent almost two-thirds of the *M. incognita* canonical TE content (64.49 %). Hence, the *M. incognita* TE landscape is diversified but mostly composed of DNA-transposons.

As a technical validation of our repeatome annotation protocol (see methods; sup. Fig S7), we performed the same analysis in *C. elegans*, using the PRJNA13758 assembly (The C. elegans Genome Sequencing Consortium 1998). We compared our results (Kozlowski 2020b) to the reference report of the TE landscape in this model nematode (Bessereau 2006) (sup. Table S2). We estimated that the *C. elegans* repeatome spans 11.81% of its genome, which is close to the 12 % described in (Bessereau 2006). The same resource also reported that MITEs and LTR respectively compose ~2% and 0.4% of the *C. elegans* genomes while we predicted 1.8% and 0.2%. Predictions obtained using our protocol are thus in the range of previous predictions for *C. elegans;* which suggest our repeatome prediction and annotation protocol is accurate.

The wormbook resource (Bessereau 2006) mentioned that most of *C. elegans* TE sequences "are fossil remnants that are no longer mobile", and that active TEs are DNA transposons. This suggests a stringent filtering process is necessary to isolate TEs that are the most likely to be active (e.g. the 'canonical' ones). Using the same post-processing protocol as for *M. incognita*, we estimated that canonical TEs span 3.60% of the *C. elegans* genome, with DNA-transposon alone representing 76.6% of these annotations (sup. Fig S1 & sup Table S3).

Table 1: Per-order summary of *M. incognita* canonical TE annotations.

Autonomous TE orders (*) regroup elements known to present transposition machinery and thus able to transpose by themselves. On the opposite, non-autonomous orders (**) regroup elements lacking transposition machinery and therefore relying on autonomous elements to transpose.

	order autonomous (*) / non-autonomous (**)	**nb. of features**	**total length (bp)**	**genome percentage (%)**	**median length (bp)**	**median identity with consensus (%)**
Retro - transposon	SINE (**)	9	4,522	0.002	528.0	99.7
	LARD (**)	45	6,342	0.035	1433.0	97.05
	TRIM (**)	174	104,018	0.057	525.0	97.7
	LINE (*)	145	313,224	0.171	1971.0	96.6
	LTR (*)	373	1,164,836	0.635	2415.0	97.0
DNA - transposon	Helitron (*)	18	86,666	0.047	5080.0	94.4
	Maverick (*)	189	1,307,068	0.712	6224.0	95.3
	MITE (**)	5085	2,755,381	1.501	525.0	96.2
	TIR (*)	3595	2,777,270	1.513	737.0	97.3
	Total	**9,633**	**8,576,405**	**4.673**		

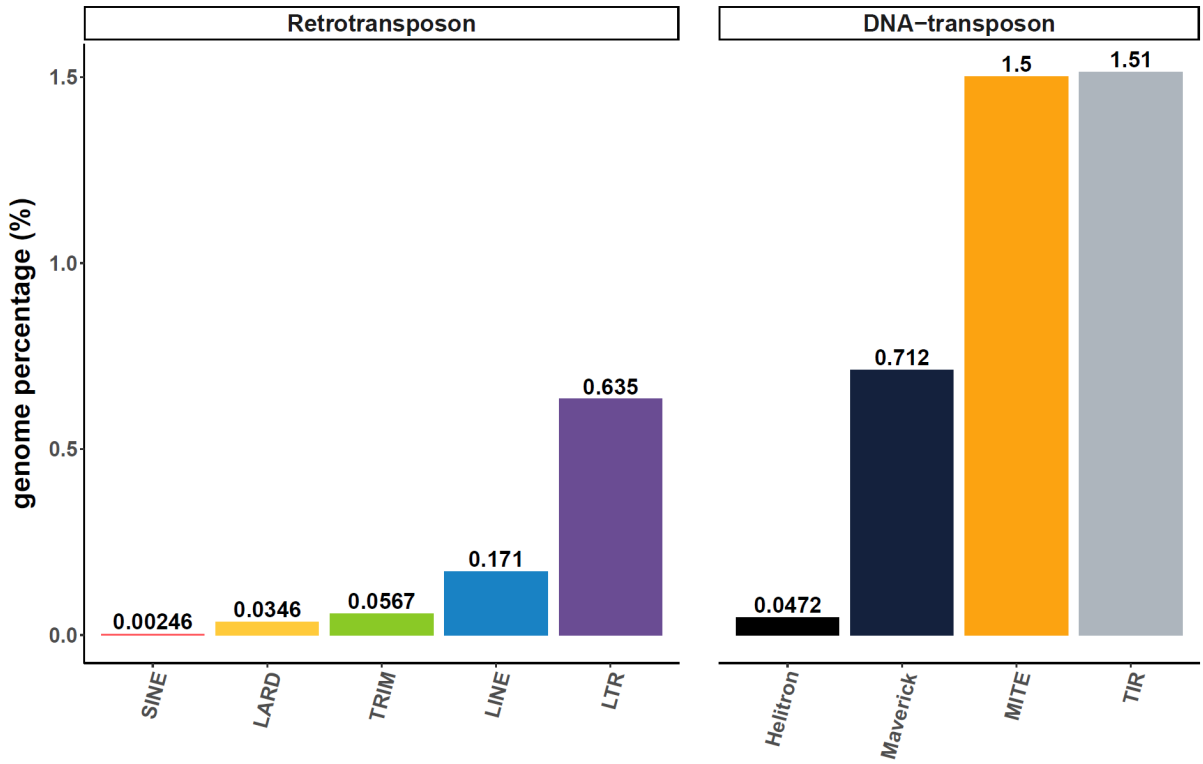


Fig 1: Canonical TE annotations distribution in *M. incognita* genome

Genome percentage is based on a *M. incognita* genome size of 183,531,997 bp (Blanc-Mathieu et al. 2017).

Canonical TE annotations are highly identical to their consensus sequences and some present evidence for transposition machinery.

Canonical TE annotations have a median nucleotide identity of 97% with their respective consensus sequences, but the distribution of identity values varies between TE orders (Fig 2, sup. Table S4). Most of the TEs within an order share a high identity level with their consensuses, the lowest values being observed for Helitron and Maverick elements. Yet, more than half of those elements share above 94% identity with their consensuses, (sup. Fig S3). Although it might be hypothesized the lower identities would be due to bigger length, we showed no evident correlation between the % identity copies share with their consensus and the proportion of consensus length covered (sup. Fig S3). Even considering our inclusion threshold at minimum 85% identity (see methods), the overall distribution of average % identities tends to be asymmetrical, and skewed towards higher values (Fig 2).

Among DNA-transposons, identity profiles of MITEs and TIRs to their consensuses were the most shifted to high values; one fourth of the TIRs annotations sharing above 99% identity with their consensus (Fig 2; sup. Tables 2 and 4).

Among retrotransposon, SINEs (present in very low numbers) and TRIMs show similar profiles with a quite narrow peak at more than 97% identity. Overall, these results indicate that notwithstanding small differences between orders, the canonical TEs show a high similarity with their consensuses.

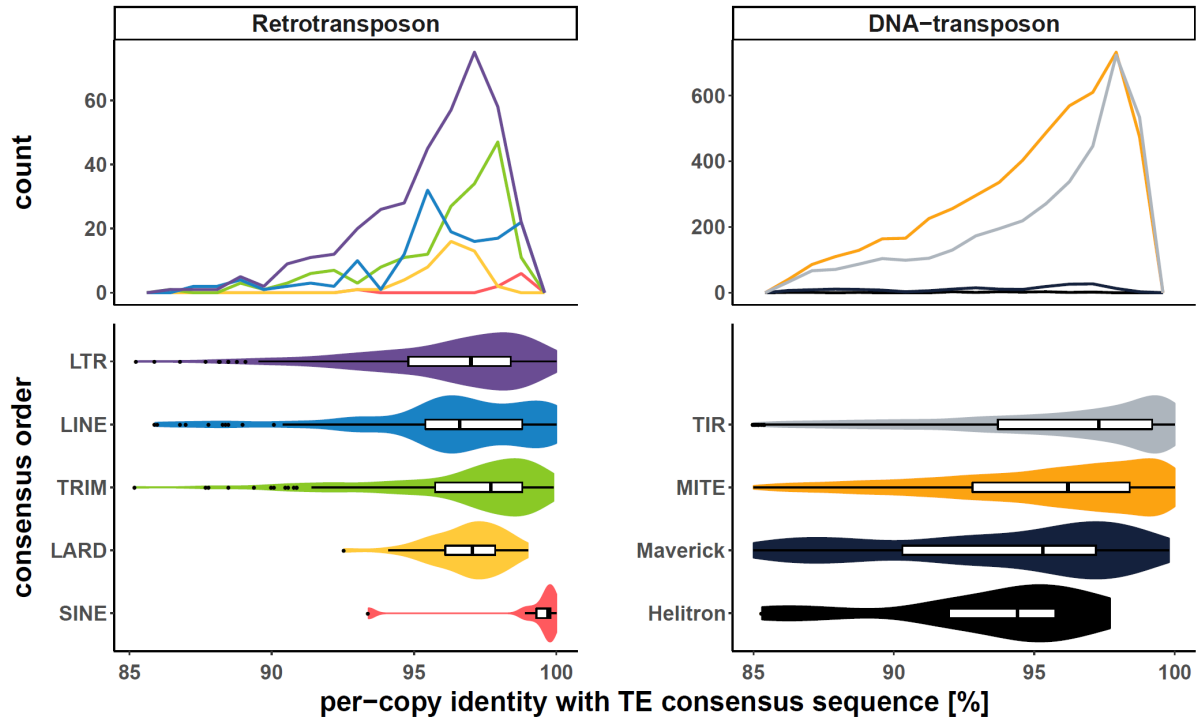


Fig 2: **per-copy identity rate with consensus**

Top frequency plots show the distribution of TE copies count per order in function of the identity % they share with their consensus sequence. To facilitate inter-orders comparison, bottom violin plots display the same information as a density curve, but also encompass boxplots. Each colour is specific to a TE order.

High identity of TE annotations to their consensus can be considered a proxy of their recent activity (Bast et al. 2015; Lerat et al. 2019). To further investigate whether some TEs might be (or have been recently) active, we searched for the presence of genes involved in the transposition machinery within *M. incognita* canonical TEs (see methods). Among the canonical TE annotations, 6.21% (598/9,633) contain at least one predicted protein-coding gene, with a total of 893 genes involved. Of these 893 genes, 344 code for proteins with at least one conserved domain known to be related to transposition machinery. We found that 31.98% (110/344) of the transposition machinery genes had substantial expression support from RNA-seq data. In total, 106 canonical TE-annotations contain at least one substantially expressed transposition machinery gene (Kozlowski, Da Rocha, et al. 2020). These 106 TE annotations correspond to 39 different TE-consensuses, and as expected, only consensuses from the autonomous TE orders, e.g. LTRs, LINEs, TIRs, Helitron, and Maverick present TE-copies with substantially expressed genes coding for transposition machinery (sup. Table S5). Conversely, the non-autonomous TEs do not contain any transposition machinery gene at all. This suggests that some of the detected TEs have functional transposition machinery, which in turn could be hijacked by the non-autonomous elements.

Overall, the presence of a substantial proportion of TE annotations highly similar to their consensuses combined with the presence of genes coding for the transposition machinery and supported by expression data suggest some TE might be active in the genome of *M. incognita*.

Thousands of loci show variations in TE presence frequencies across *M. incognita* isolates.

We used the PopoolationTE2 (Kofler et al. 2016) pipeline on the *M. incognita* reference genome (Blanc-Mathieu et al. 2017) and the canonical TE annotation to detect variations in TE frequencies across the genome between 12 geographical isolates (see methods; (Kozlowski 2020b); sup. Fig S7). One isolate comes from Morelos in Mexico, which is the isolate that was used to produce the *M. incognita* reference genome. The 11 other isolates come from different locations across Brazil, and present four different ranges of compatible hosts (referred to as R1, R2, R3, R4, see sup. Fig S4) and currently infected crop species (Koutsovoulos et al. 2020). Pool-seq paired-end Illumina data has been generated for all these isolates. For each locus, each isolate has an associated frequency value representing the proportion of individuals in the pool having the TE detected at this location.

We identified 3,514 loci where the frequency variation between at least two isolates was higher than our estimated PopoolationTE2 error rate (0.00972 *i.e.* less than 1%, see methods).

Overall, the distribution of within-isolate frequencies is bimodal (Fig 3-A), and this pattern is common to all the isolates, including the reference Morelos isolate (Fig 3-B). On average, 21.1% of the loci have within-isolate frequencies < 25%, 60.7% have frequencies > 75%, and only 18.2% show intermediate frequencies Hence, most of the within-isolate TE frequencies pack around extreme values e.g. <25% or >75%.

Nevertheless, these statistics provide no information about the frequency variability between isolates for a given locus. To address this question, for each locus, we computed the absolute maximum frequency difference between isolates (Fig 3-C). We found that the maximum frequency variation across the isolates is smaller than 20% in 75% of the loci (2,634/3,514). Hence, most of the loci show little to moderate isolate-wide variations in frequencies. Combined to the previous result, this implies that for most loci, the TEs are present either at a high or a low frequency among all isolates. However, some TE loci show more contrasted variations and will be the focus of further studies in our pipeline.

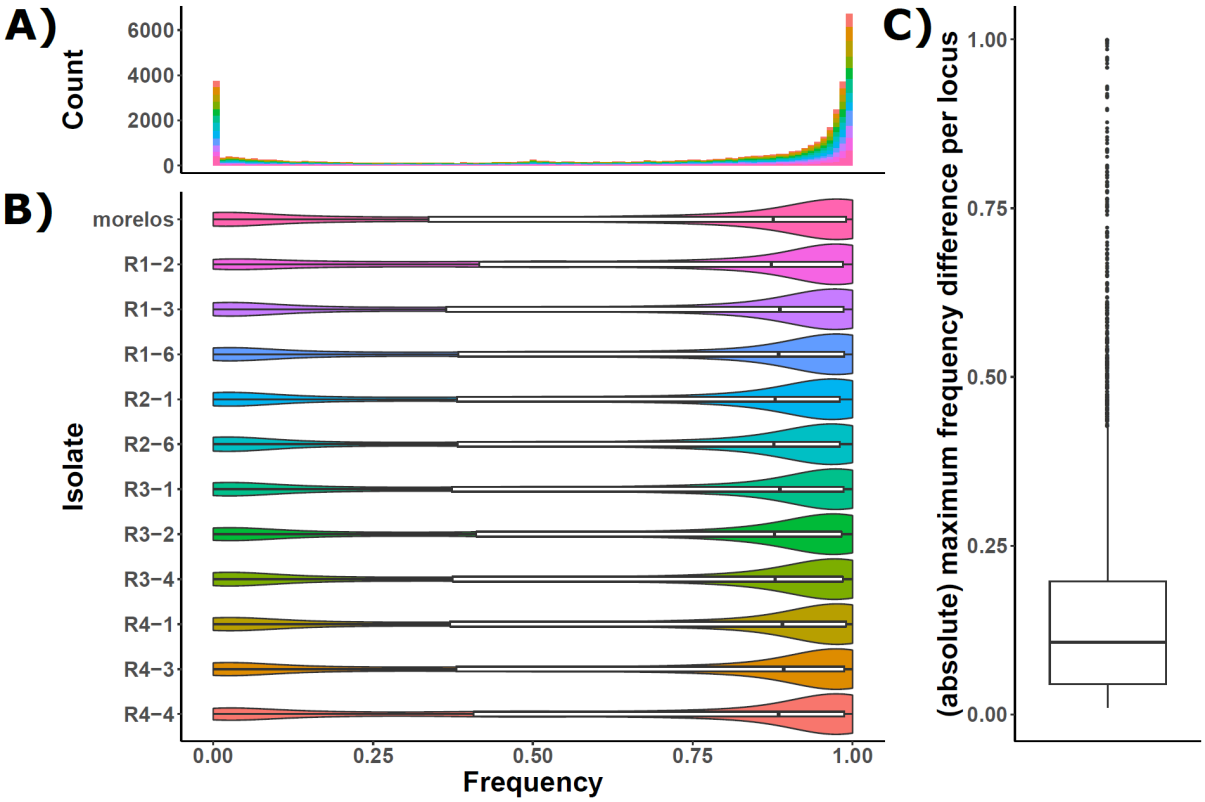


Fig. 3: TE frequency distribution.

The histogram (A) and violin plot (B) represent the TE frequency distribution per isolate. The colour chart is identical between the two figures. Both representations reveal that in all the isolates, only a few TE are found with intermediate frequencies. Right boxplot (C) represents the frequency absolute maximum difference per locus. For a given locus, it illustrates the frequency variability between isolates. The higher is the value; the more important is the frequency difference between at least two isolates. A value of 1 implies that the TE is absent in at least one isolate while it is present in 100% of the individuals of at least another isolate.

Variations of TE frequencies across isolates recapitulate their divergence at the sequence level

We performed a Neighbour-joining phylogenetic analysis of *M. incognita* isolates based on a distance matrix constructed from TE frequencies (3,514 loci; see methods). We also performed a Maximum Likelihood (ML) analysis based on SNV in coding regions as previously identified in (Koutsovoulos et al. 2020) adding the reference isolate Morelos.

As shown in Fig.4, the TE-based and SNV-based tree topologies are highly similar. In particular, the two trees allowed defining four highly supported clades, with bootstrap support values ≥ 98. The four clades were identical, including branching orders for clades 2 and 4 (the two other clades containing each only two isolates). R1-6 and R2-1 positions slightly differed between the SNV-based (A) and TE-based (B) trees. However, in both trees, R1-6 is more closely related to clusters 1 and 2 than the rest of the isolates, and similar observations can be drawn for R2-1 with clusters 3 and 4.

Altogether, the similarity between the SNV-based and TE frequency-based trees indicates that most of the phylogenetic signal coming from variations in TE-frequencies between isolates recapitulates the SNV-based genomic divergence between isolates.

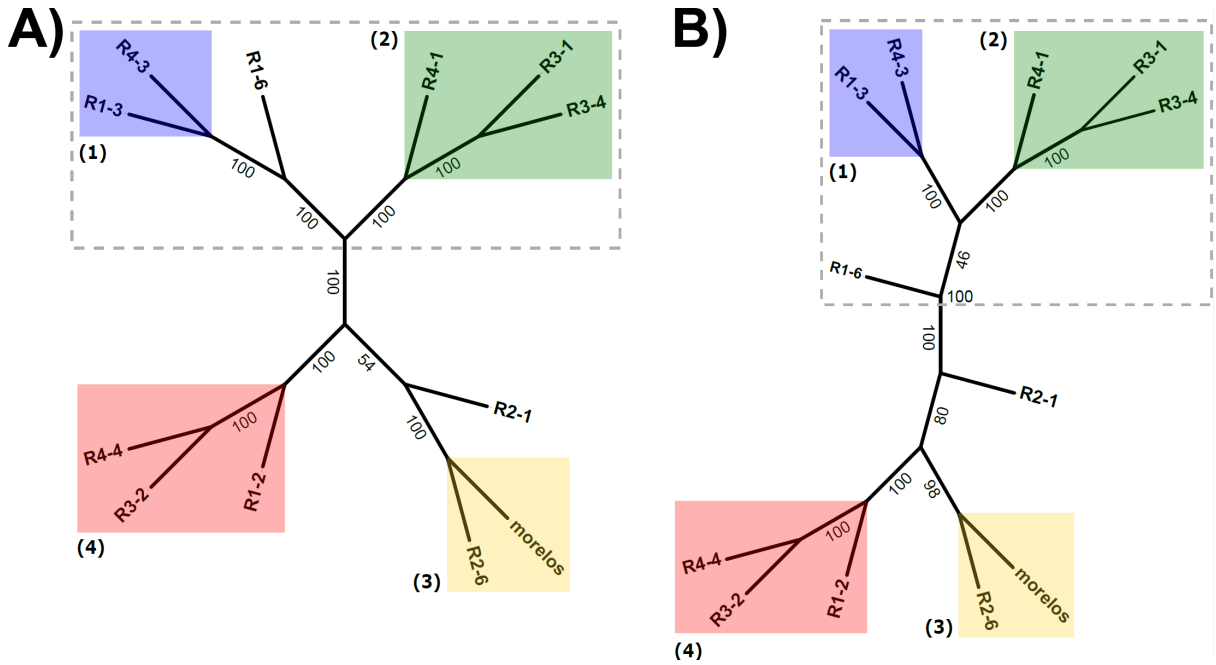


Fig 4: Phylogenetic tree for *M. incognita* isolates.

A- Phylogenetic tree based on SNV present in coding sequences. Maximum Likelihood (ML) tree reconstruction. Branch length not displayed (see sup. Fig S5 for a version with branch length displayed). B- Phylogenetic tree based on TE-frequencies euclidean distances between isolates. Neighbor-Joining (NJ) tree reconstruction. Branch length not displayed (see sup. Fig S5 for a version with branch length displayed). In both trees, bootstrap support values are indicated on the branches. Isolates enclosed in the dashed area form a super-cluster composed of the clusters (1) and (2), and the isolate R1-6.

Most of the TE frequency variations across the isolates concern TE present in the reference genome although additional TE loci were identified.

As explained below (see also methods sup. Figs S7 & S8), we categorized all the loci with TE frequency variations between the isolates by (i) comparing their position to the TE annotation in the reference genome, (ii) analysing TE frequency in the reference isolate Morelos, (iii) comparing TE-frequencies detected for each isolate to the reference isolate Morelos. This allowed defining, on the one hand, non-polymorphic and hence stable reference annotation, and on the other hand, 3 categories of polymorphic (variable) loci (Fig 5).

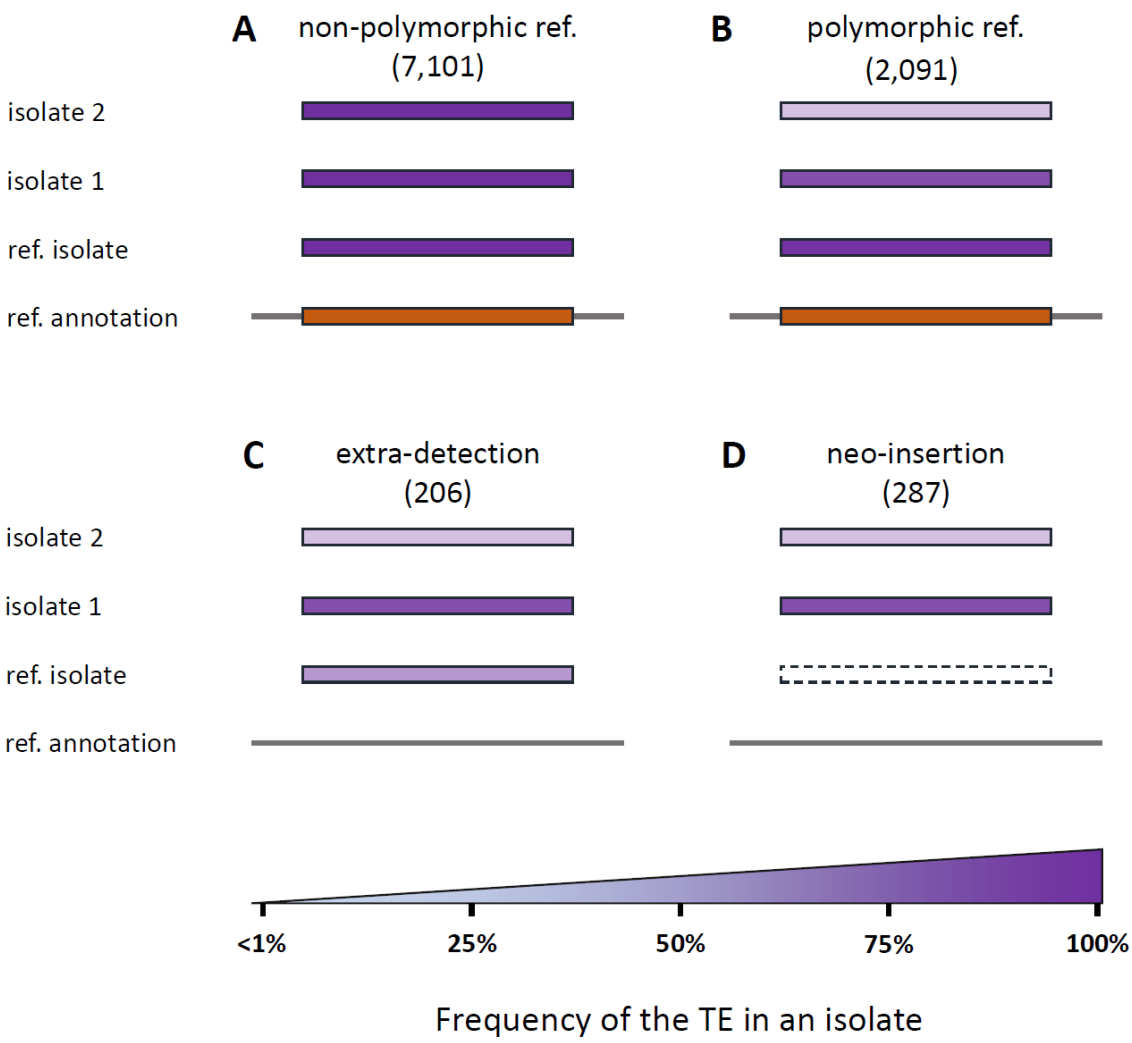


Fig 5: Categories of polymorphic TE loci

Orange boxes illustrate the presence of a TE at this locus in the reference genome annotation. Purple boxes illustrate the percentage of individuals in the isolates for which the TE is present at this locus (i.e. frequency). Frequency values are reported as colour gradients. A - non-polymorphic ref. TE locus: a TE is predicted in the reference annotation (orange box) AND no frequency variation exceeding 1% between isolates (Morelos included) is detected. B - polymorphic ref. locus: a TE is predicted in the reference annotation, is detected in the reference isolate Morelos with a frequency > 75%, and the presence frequency varies (>1%) in at least one isolate. C - extra-detection: no TE is predicted at this locus in the reference annotation but one is detected at a frequency >25% in the reference isolate Morelos, and optionally in other isolates. D - neo-insertion: no TE is predicted at this locus in the reference genome annotation and none is detected in the reference isolate (dashed box, frequency < 1%), but a TE is detected in at least another isolate with a frequency >= 25%.

Overall, 73.5% (2,584/3,514) of the loci with TE frequency variations could be assigned to one of the 3 categories of TE-polymorphisms (B, C, D in Fig 5) and the decomposition per TE order is given in Fig 6 and sup. Table S6.

The vast majority of the polymorphic loci (80.92 %; 2,091/2,584) corresponds to an already existing TE-annotation in the reference genome and the corresponding TE is fixed (frequency > 75 %) at least in the reference isolate Morelos but varies in at least another isolate. These polymorphic loci cover ~21.6% (2,091/9,702) of the canonical TE annotations, in total. These loci will be referred to as 'polymorphic reference loci' from now on (Fig 5B) and they encompass both DNA- and Retro-transposons.

Then, we considered as 'neo-insertion' TEs present at a frequency >25% in at least one isolate at a locus where no TE was annotated in the reference genome and the frequency of TE presence was higher than the estimated error rate (~1%) in the reference Morelos isolate (Fig 5D). In total, 11.11 % (287/2,584) of the detected TE polymorphisms correspond to such neo-insertions. It should be noted here that we consider neo-insertions as regard to the reference Morelos isolate only and some of these so-called neo-insertions might represent TE loss in Morelos. Comparison with the phylogenetic pattern of presence / absence will allow distinguishing further the most parsimonious of these two possibilities (see next sections).

Finally, we classified as 'extra-detection' (Fig 5C) (7.97%; 206/2,584) the loci where no TE was initially annotated by REPET in the reference genome, but a TE was detected at a frequency >25% at least in the ref isolate Morelos by PopoolationTE2. It should be noted that 58.73% (121/206) of these loci correspond to draft annotations that have been discarded during the filtering process to only select the canonical annotations. These draft annotations might represent truncated or diverged versions of TE that exist in a more canonical version in another locus in the genome. Half of the remaining 'extra-detections' (42/85) are detected with low to moderate frequency (<42.6%) in the reference isolate Morelos. We hypothesise that because they represent the minority form, these regions were not taken into account during the assembly of the genome. This would explain why these TEs could not be detected in the genome assembly by REPET (assembly-based approach) but were identified with a read mapping approach on the genome plus repeatome by PopoolationTE2. The remaining 'extra-detections' might correspond to REPET false negatives, PopoolationTE false positives, or a combination of the two. Nonetheless, we can notice these cases only represent 1.63% (42/2,584) of the detected polymorphic TEs.

TIR and MITE elements are overrepresented among TE-polymorphisms.

By themselves, MITE and TIR elements encompass 94.58% (2,444/2,584) of the categorized TE-polymorphisms (Fig 6).

We showed that the polymorphism distribution varies significantly between the four categories presented in Fig. 5 (Chi-square test, p-value < 2.2e-16), indicating that some TE orders are characterised by specific polymorphisms types.

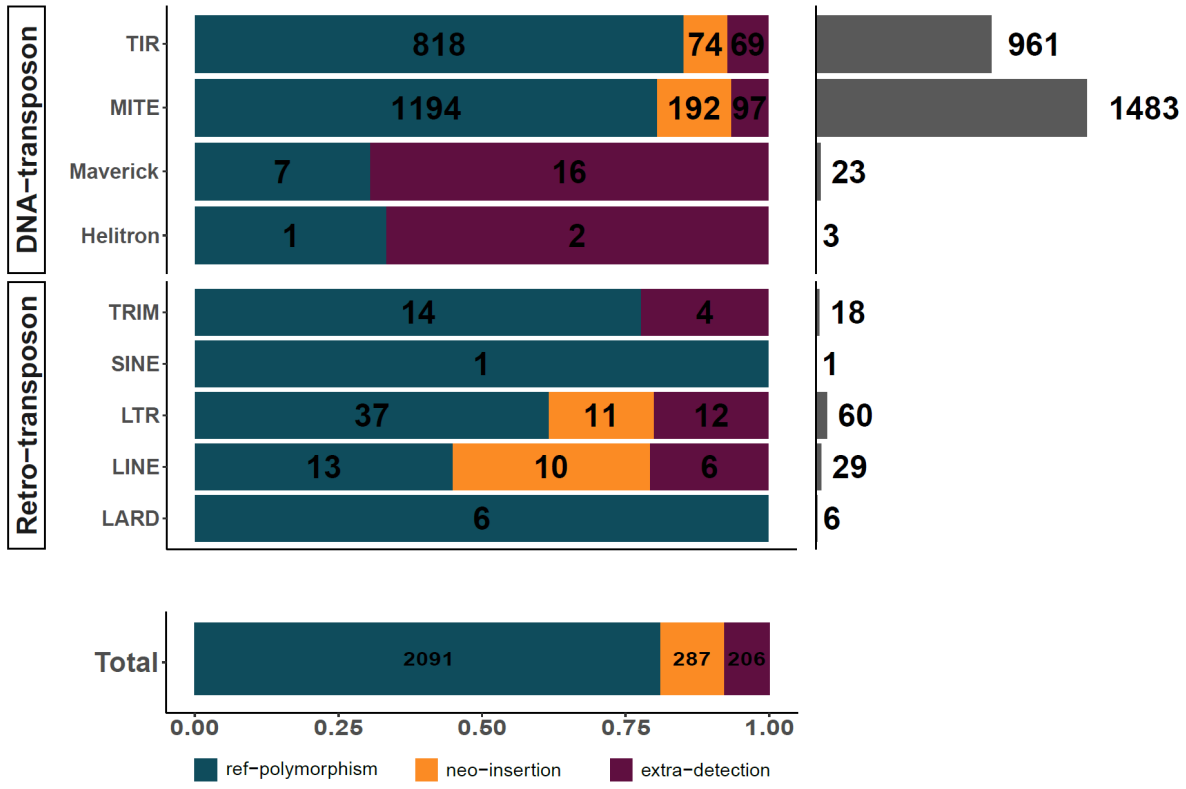


Fig 6: TE polymorphisms count per orders and types.

The top left barplot shows TE polymorphisms distribution per type and per order. The bottom-left barplot summarizes TE polymorphisms distribution per type. In both barplots, the values in black represent the count per polymorphism type. The top-right barplot illustrates the total number of polymorphisms per order.

The analysis of the chi-square residuals (sup. Fig S6) shows MITEs and TIRs are the only orders presenting a relative lack of non-polymorphic TEs. Hence, in addition to being the most abundant in the genome, these two TE orders are significantly enriched among polymorphic loci. MITEs are over-represented in both TE polymorphisms types (polymorphic ref. loci and neo-insertions, Fig5 B and D), suggesting a variety of activities within this order. On the other hand, TIRs are found in excess in ref-polymorphisms but lack in neo-insertions. This lack of neo-insertions in TIRs may indicate a recent lower activity in this order, or a more efficient negative selection.

Finally, we observed a strong excess of Maverick among the extra-detection as almost 70% of Maverick polymorphisms (16/23) (Fig 6) fell into this category. Consistent with the observation that, globally, >50% of the extra detections were actually draft annotations eliminated afterwards during filtering steps; ¾ (12/16) of the Maverick elements were also actually present in the draft annotations but later eliminated during filtering steps.

Overall, in proportion, MITEs and TIRs elements are significantly over-represented in TE-polymorphisms. This observation suggests TEs from MITE and TIR orders, in addition to being the most numerous canonical TEs, might have been more active in the genome of *M. incognita* than elements from other TE-orders.

Some polymorphic loci with contrasted frequency variations between isolates most probably represent true neo-insertions.

We investigated the variability of TE presence frequency per locus between the 12 isolates for all the categorized polymorphic loci in the genome.

In ~ 3/4 (1,911/2,584) of the categorized polymorphic TE loci, the TE presence frequency is homogeneous between isolates (see methods; sup. Fig S8). Said differently, it means that although we observe variations in frequencies between isolates above the estimated error rate (<1%), these variations remain at low amplitude (maximum frequency variation between isolates ≤25% for a given locus). The vast majority (97.95%; 1,872/1,911) concerns loci where the TE is present at a high frequency in all isolates (> 75%). These loci might be considered as fixed in all the isolates. In the remaining 2.04% (39/1,911), the TE frequency is either between 25 and 50% or between 50 and 75% in all isolates. As expected given our methodology, all the high-frequency loci correspond to ref-polymorphisms while all the intermediate frequency loci belong to extra-detections.

In the 673 remaining polymorphic TE loci, TE frequency is heterogeneous, meaning the frequency difference between at least two isolates is > 25% (median difference = 31.35%). Among the most extreme cases of frequency variation per locus, we identified 33 loci in which the TE is found with high frequencies (> 75%) for some isolate(s) while it is absent or rare (frequency <25 %) in the other(s). These loci will be from now on referred to as HCPTEs standing for "Highly Contrasted Polymorphic TE" loci. Because they are highly contrasted, these loci might represent differential fixation/loss across isolates and will be the focus of the following analyses.

HCPTEs encompass 19 MITE elements, 12 TIRs and 2 LINEs (sup. Table S7). We can also notice that some consensuses are more involved in HCPTEs as two TE consensuses alone are responsible for 72.72% (18/33) of these polymorphisms (one MITE consensus involved in 10 HCPTEs, one TIR consensus involved in 8 HCPTEs).

Interestingly, all the HCPTEs loci correspond to neo-insertions regarding the reference genome, meaning that no TE was annotated in the reference genome at this location and the TE presence frequency is < 1% in the Morelos reference isolate. As described in Fig. 7, most of these fixed neo-insertions (20/33) are specific to an isolate and most probably represent lineage-specific neo insertions rather than multiple independent losses.

However, we also found neo-insertions shared by two (10/33), three (2/33) or even six isolates (1/33). Interestingly, all the shared neo-insertions were between isolates present in a same cluster in the phylogenetic trees (TE-based and SNV-based in Fig. 4), suggesting they might have been fixed in a common ancestor and then inherited. For example, two neo-insertions are shared by isolates R4-4, R1-2 and R3-2 which belong to the same cluster 1 and one neo-insertion is shared by isolates R4-3 and R1-3 which belong to the same cluster 2. Even the neo-insertion shared by 6 isolates follows this pattern as all the concerned isolates belong to the same super-cluster composed of the cluster 2 and 3 plus isolate R1-6 (dashed line in Fig 4).

Hence, the phylogenetic distribution reinforces the idea that these cases are more likely to represent branch-specific neo-insertions than multiple independent losses, including in the reference isolate Morelos.

Isolates R1-2, R3-2, and R4-4 show the highest number of neo-insertions. However, their profiles are quite different. In R1-2, 10/12 HCPTEs are isolate-specific while most of the HCPTEs involving R3-2 and R4-4 are neo-insertions shared with closely related isolates. This is also consistent with the topology and branch

lengths of the SNV-based and TE-based phylogenies (sup. Fig S5), which shows that R1-2 is the most divergent isolate with the longest branch length, while R3-2 is quite close to R4-4 and has a relatively short branch.

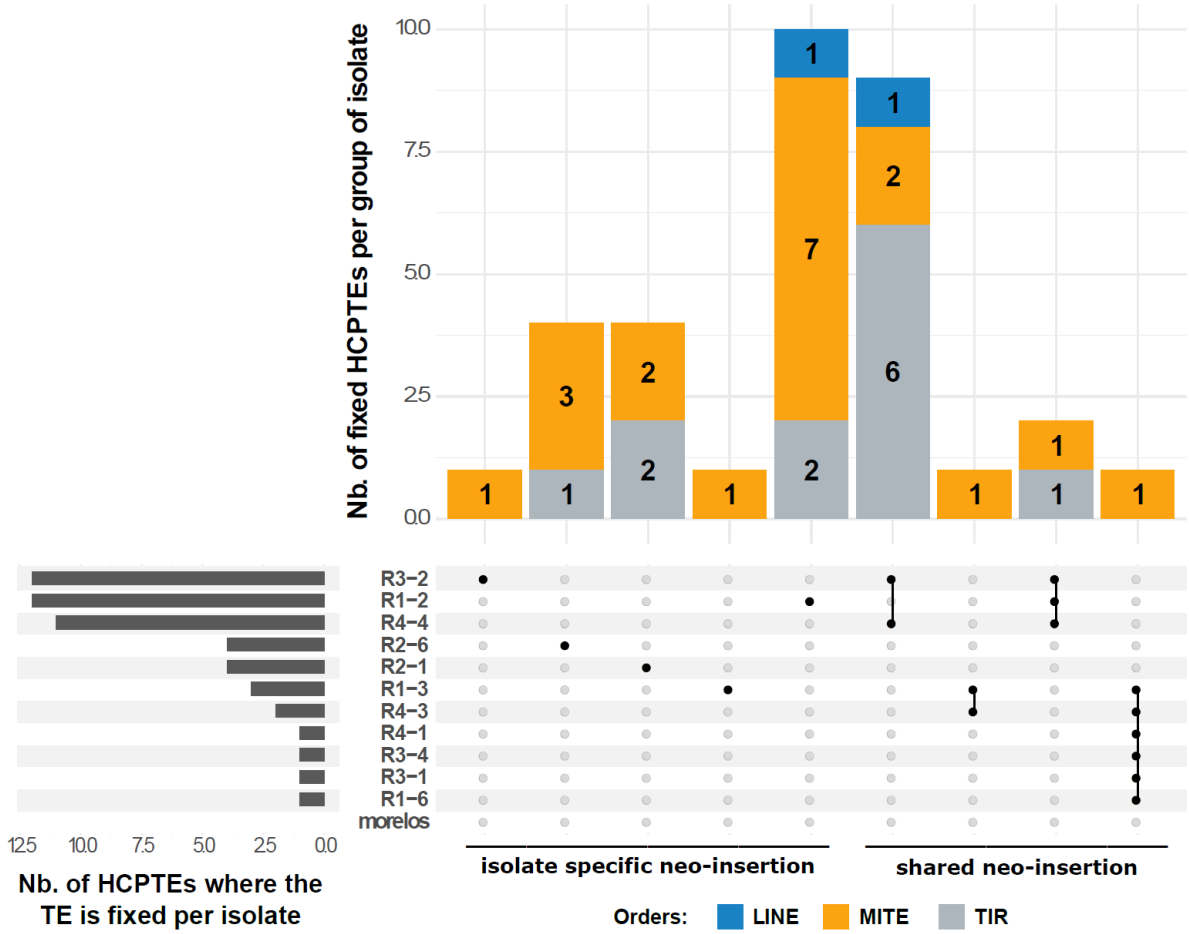


Fig 7: HCPTEs Neo-insertions specificity among the isolates.

The central plot shows how many and which isolate(s) share common HCPTEs neo-insertion(s), every line representing an isolate. Columns with several dots linked by a line indicate shared HCPTEs neo-insertion(s) between isolates. Each dot represents which isolate is involved. Columns with a single dot design isolate-specific HCPTEs neo-insertion(s). The top bar plot indicates how many HCPTEs neo-insertions the corresponding group of isolate shares. The left side barplot specifies how many HCPTEs neo-insertion(s) occurred in a given isolate.

Functional impact of TE neo-insertion and validation of in silico predictions

Interestingly, two-thirds (22/33) of the fixed HCPTEs are inserted inside a gene or in a possible regulatory region (1 kb region upstream of a gene). These fixed neo-insertions might have a functional impact in *M. incognita*. Overall, 27 different genes (26 coding for proteins and one tRNA gene) are possibly impacted by the 22 neo-insertions, some genes being in the opposite direction at a neo-insertion point (overlapping this insertion point or being at max 1kb downstream). More than 80% of these genes (22/27) show a substantial expression level during at least one life stage of the nematode life cycle (in the Morelos isolate), suggesting the impacted genes are functional in the *M. incognita* genome (see methods). Some of the impacted genes (40.74%, 11/27) are specific to the *Meloidogyne* genus (they have no predicted orthologs in other nematodes, according to WormBase Parasite). Ten of these Meloidogyne-specific genes are widely conserved in multiple Meloidogyne species, reinforcing their possible importance in the genus, and one is so far only present in *M. incognita*. Interestingly, further similarity search using BLASTp against the NCBI's nr library returned no significant hits, suggesting these proteins are so far Meloidgyne-specific and do not originate from horizontal gene transfers of non-nematoda origin. Among the remaining genes, one is present in multiple *Meloidogyne* species and otherwise only found in other Plant Parasitic Nematodes species (PPN) (*Ditylenchus destructor, Globodera rostochiensis*) (sup. Table S8). Conservation of these genes across multiple PPN but exclusion from the rest of the nematodes or other species suggest these genes might be involved in important functions relative to these organisms' lifestyle, including plant parasitism itself.

To experimentally validate *in-silico* predictions of TE neo-insertions with potential functional impact, we performed PCR experiments on 5 of the 22 HCPTEs loci falling in coding or possible regulatory regions (see methods for selection criteria). To perform these PCR validations, we used the DNA remaining from previous extractions performed on the *M. incognita* isolates for population genomics analysis (Koutsovoulos et al. 2020). Basically, the principle was to validate whether the highly contrasted frequencies (>75% / <25%) obtained by PopoolationTE2 actually corresponded to absence/presence of a TE at the locus under consideration (see methods). One isolate (R3-1) presented no amplification in any of the tested loci nor in the positive control. After testing the DNA concentration in the sample, we concluded that the DNA quantity was too low in this isolate and decided to discard it from the analysis.

For four of the five tested HCPTEs loci, we could validate by PCR the *in-silico* predicted differential presence/absence of a sequence at this position, across the different isolates (Fig 8; (Kozlowski, Hassanaly-Goulamhoussen, et al. 2020)).

In one of the five tested loci, named locus 1, we could i) validate by PCR the presence of a sequence at this position for the isolates presenting a PopoolationTE2 frequency >75% and absence for those having a frequency <25%; ii) also validate by sequencing that the sequence itself corresponded to the TE under consideration (a MITE). This case is further explained in detail below and in Fig. 8.

According to PopoolationTE2 frequencies, in the concerned locus, 1 MITE is inserted and fixed in 3 isolates (R1-2, R3-2, R4-4) as the estimated frequencies are higher than 75% in these isolates. We assumed the TE is absent from the rest of the isolates as all of them display frequencies <5%. To validate this differential presence across the isolates, we designed specific primers from each side of the estimated insertion point so that the amplicon should measure 973 bp with the TE insertion and 180 bp without.

The PCR results are consistent with the frequency predictions as only R1-2, R3-2, and R4-4 display a ~1 kb amplicon while all the other isolates show a ~0.2 kb amplicon (Fig 8). Hence, as expected, only the 3 isolates with a predicted TE frequency >75% at this locus exhibit a longer region, compatible with the MITE insertion.

To validate the amplified regions corresponded to the expected MITE, we sequenced the amplicons for the 3 predicted insertions and aligned the sequences to the TE consensus and the genomic region surrounding the estimated insertion point (Kozlowski, Hassanaly-Goulamhoussen, et al. 2020). Amplicon sequences of R-1_2, R-3_2, and R-4_4 all covered a significant part of the TE consensus sequence length (> 78%) with high % identity (> 87%) and only a few gaps (<5%). These results confirm that the inserted sequence corresponds to the predicted TE consensus. Moreover, all the 3 amplicons aligned on the genomic region downstream of the insertion point with high % identity (>= 99%), which helped us further determine the real position of the insertion point. The real insertion point is 26 bp upstream of the one predicted by PopoolationTE2 and falls in the forward primer sequence. This explains why the amplicon sequences do not align on the region upstream the insertion point.

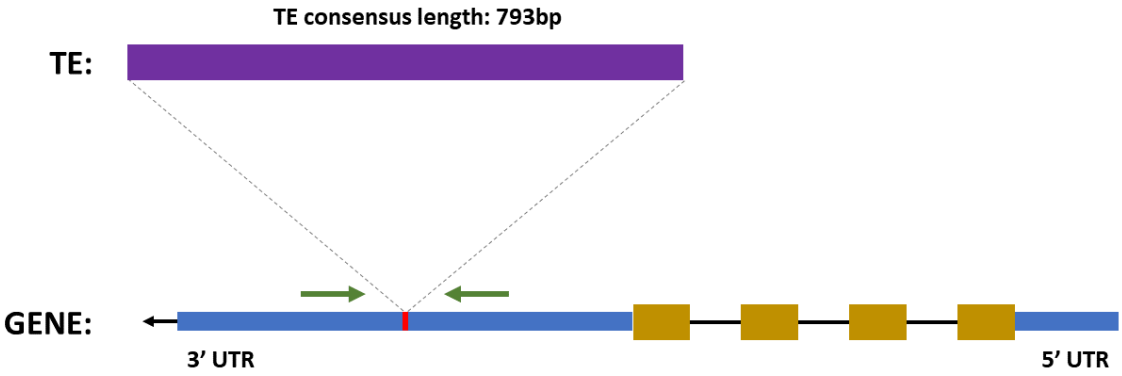
We also noticed that the inserted TE sequences slightly diverged between the isolates while the genomic region surrounding the insertion point remains identical. Interestingly, the level of divergence in the TE sequence does not follow the phylogeny as R-4_4 is closer to R-1_2 than to R-3_2 (sup. Table S9).

Finally, in the Morelos, R-2_1, and R-2_6 isolates, the sequencing of the amplicon validated the absence of insertions. Indeed, the sequences aligned on the genomic region surrounding the insertion point with high % identity (99, 97, 87 % respectively) but not with the MITE consensus.

Hence, we fully validated experimentally the presence/absence profile across isolates predicted *in silico* at this locus.

In the *M. incognita* genome, this neo-insertion is predicted to occur in the 3' UTR region of a gene (Minc3s00026g01668). This gene has no obvious predicted function, as no conserved protein domain is detected and no homology to another protein with an annotated function could be found. However, orthologs were found in the genomes of several other *Meloidogyne* species (*M. arenaria*, *M. javanica*, *M. floridensis*, *M. enterolobii*, and *M. graminicola*), ruling out the possibility that this gene results from a prediction error from gene calling software. The broad conservation of this gene in the *Meloidogyne* genus suggests this gene might be important for *Meloidogyne* biology and survival.

In the Morelos isolate, for which no TE was inserted at this position, this gene is supported by transcriptomic RNA-seq data during the whole life cycle of the nematode (Kozlowski, Da Rocha, et al. 2020), suggesting this gene is probably functionally important in *M. incognita* and other root-knot nematodes. Consequently, the insertion of the TE in R-1_2, R-3_2, and R-4_4 genome at this locus could have functional impacts.



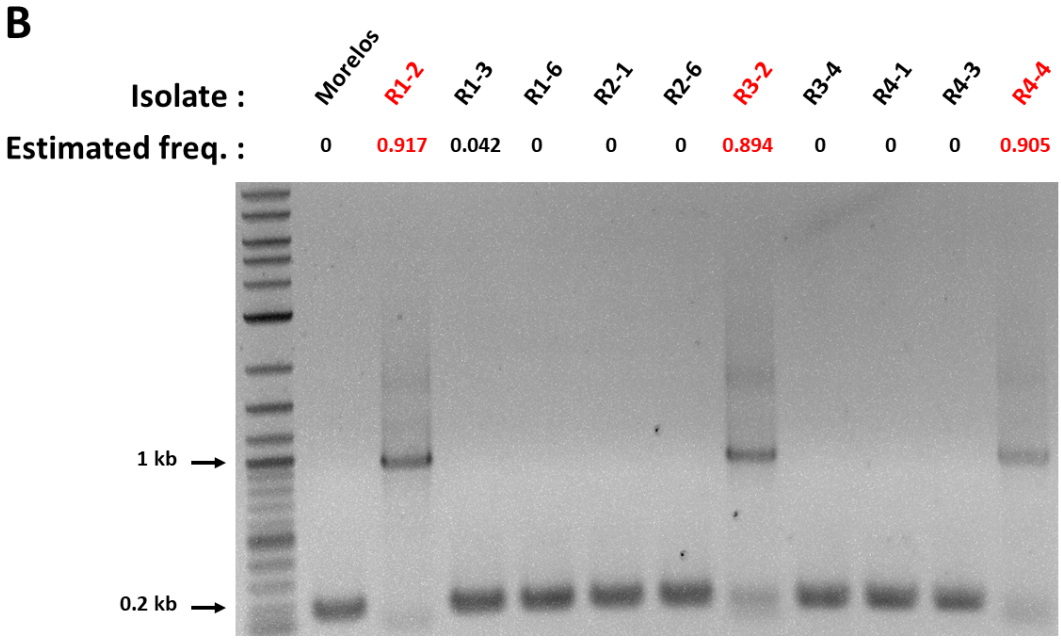


Fig 8: Experimental validation of a predicted neo-insertion.

A- Diagram of the TE neo-insertion. The neo-insertion of the MITE element occurs in the 3'UTR region of the gene (Minc3s00026g01668). Blue boxes illustrate the 3' and 5' UTR regions of the gene while the yellow boxes picture the exons. Green arrows represent the primers used to amplify the region. Gene subparts and TE representations are not at scale. Predicted size of the amplicon: 973 bp with the TE insertion, 180 bp without. B- PCR validation of the TE neo-insertion. Estimated freq. values correspond to the proportion of individuals per isolate predicted to have the TE at this position (PopoolationTE2). Isolates in red were

predicted to have the TE inserted at this locus. Only these isolates show an amplicon with a size suggesting an insertion (sequences are available in (Kozlowski, Hassanaly-Goulamhoussen, et al. 2020)).

Discussion

TE landscape in nematode genomes and possible recent activity in *M. incognita*

In this analysis, we have annotated TEs in the genome of *M. incognita* and used variations in TE frequencies between geographical isolates across loci as a reporter of their activity. The *M. incognita* TE landscape is more abundant in DNA than retro-transposons and using the same methodology, we confirmed a similar trend in the genome of *C. elegans*. Interestingly, even if the methodology used was different, a similar observation was made at the whole nematoda level (Szitenberg et al., 2016), suggesting a higher abundance of DNA transposons might be a general feature of nematode genomes.

We have shown 75% of the polymorphic TE loci in *M. incognita* display moderate frequency variations between isolates (<25%); a majority being found with high frequencies (> 75%) in all the isolates simultaneously. Hence, a substantial part of the TE can be considered as stable and fixed among the isolates.

Nevertheless, the remaining quarter of polymorphic TE loci present frequency variations across the isolates exceeding 25%. This observation concerns both the TE already present in the reference genome, but also the neo-insertions. We even detected loci where the TE frequencies were so contrasted between the isolates (HCPTEs) that we could predict the TE presence/absence pattern among the isolates. Such frequency variations between isolates, and the fact that part of the HCPTEs are isolate-specific neo-insertions, constitute strong evidence for TE activity in the *M. incognita* genome.

In *C. elegans*, multiple TE families have also shown a substantial level of activity across different populations (Laricchia et al. 2017). However, this analysis was based on binary presence / absence data of TE at loci across populations and thus provided no information about the amplitude of TE frequencies variability within isolates. In our analysis we provided this extra layer of information and this also allowed estimating the amplitude of TE frequency variations between *M. incognita* isolates.

It should be noted here that the total TE activity in the *M. incognita* genome is probably underestimated, in part because of our strategy to eliminate false positives as much as possible by applying a series of stringent filters, and in another part because of the intrinsic limitations of the tools, such as the incapacity of PopoolationTE2 to detect nested TEs (Kofler et al. 2016).

We then evaluated how recent this activity could be, using % identity of the TE copies with their respective consensuses as a proxy for their age as previously proposed in other studies (Bast et al. 2015; Lerat et al. 2019). We showed that a substantial proportion of the canonical TE annotations were highly similar to their consensus, indicating most of these TE copies were recent in the genome. The probable recent hybrid origin of *M. incognita* (Blanc-Mathieu et al. 2017) is consistent with a recent TE burst in the genome. Indeed, as further explained in the last section of the discussion, it is well established that hybridization events can lead to a relaxation of the TE silencing mechanisms and consequently to a TE expansion (Belyayev 2014; Guerreiro 2014; Rodriguez and Arkhipova 2018).

However, as suggested in (Bourgeois and Boissinot 2019), the extent of this phenomenon might differ depending on the TE order. In *M. incognita,* MITEs and TIRs alone account for ~2/3 of the canonical TE

annotations, but their fate in the genome seems to have followed different paths. Indeed, as illustrated in Figure 2, MITEs show a wide range of identity rate with their consensus, which suggests they might have progressively invaded the genome being uncontrolled or poorly controlled as suggested for the rice genome (Lu et al. 2017). On the opposite, almost all the TIR copies share high percentage identity with their consensuses which could be reminiscent of a rapid and recent burst. Nevertheless, this burst could have quickly been under control as, according to chi-square residuals (sup. Fig S6), TIR neo-insertions are significantly less numerous than expected owing to their abundance in the genome. Interestingly, in *C. elegans*, the Tc1 / Mariner TIR DNA element was shown to be the most active while, so far, no evidence for active retro-transposition was shown in this species (Bessereau 2006; Laricchia et al. 2017).

Because no molecular clock is available for *M. incognita*, it is impossible to evaluate more precisely when TE bursts would have happened and how fast each TE from each order would have spread in the genome. Such bursts can be very recent, including in animal genomes as exemplified by the P-element which invaded the genome of some Drosophila populations in just 40 years (Anxolabéhère et al. 1988). While an absolute dating of TE activities in *M. incognita* is currently not possible, a relative timing of the events regarding population diversification can still be deduced from the distribution of TE loci frequencies across isolates. Indeed, we have shown (Figure 7) that some neo-insertion were shared between isolates and that in each case, the concerned isolates belonged to a same monophyletic cluster (Figure 4). The most parsimonious scenario is that these neo-insertions occurred in *M. incognita*, after the separation of the different main clusters but before the diversification of the phylogenetically-related isolates, within a cluster, in a common ancestor. Other TE neo-insertions, in contrast, were so far isolate-specific, suggesting some TE movements were even more recent and that TE mobility might be a continuous phenomenon. No information is available about the ancientness of cultivated lands in Brazil on which the different isolates have been sampled. However, because there is no significant correlation between the isolates geographical distribution and the phylogenetic clusters, whether it is TE-based (this study) or SNV-based (Koutsovoulos et al. 2020), we can hypothesize these isolates have been recently spread by human agricultural activity in the last centuries.

Overall, the presence of isolate-specific TE neo-insertions, the distribution of percent identities of some TE copies to their consensuses shifted towards high value, as well as transcriptional support for some of the genes involved in the transposition machinery, suggest TE have recently been active in *M. incognita* and are possibly still active.

Functional impact of TEs activity in *M. incognita* and other nematodes

M. incognita is a parthenogenetic mitotic nematode of major agronomic importance. How this pest adapts to its environment in the absence of sexual recombination remains unresolved. In this study, we investigated whether TE movements could constitute a mechanism of genome plasticity compatible with adaptive evolution.

In *M. javanica*, a closely related root-knot nematode, comparison between an avirulent line unable to infect tomato plants carrying a nematode resistance gene and another virulent line that overcame this resistance, led to the identification of a gene present in the avirulent nematodes but absent from the virulent ones. Interestingly, the gene under consideration is present in a TIR-like DNA transposon and its absence in

the virulent line suggests this is due to excision of the transposon and thus that TE activity plays a role in *M. javanica* adaptive evolution (Gross and Williamson 2011).

In *M. incognita*, convergent gene losses at the whole genome level between two virulent populations compared to their avirulent populations of origin were recently reported (Castagnone-Sereno et al. 2019). Gene copy number variation CNV are indeed known to be involved in genomic plasticity and in adaptive evolution (Katju and Bergthorsson 2013), and TE can actively (e.g. by gene hitchhiking) or passively (e.g. through illegitimate recombination) participate in these variations. This CNV analysis in *M. incognita* was done on an older version of the genome (Abad et al. 2008), that was partially incomplete, and the possible contribution of TEs in these CNV could not be assessed. Although the current version of the genome (Blanc-Mathieu et al. 2017) is more complete and consistent with the estimated genome size, it is still fragmentary with thousands of scaffolds and a relatively low N50 length (38.6 kb). This fragmentation prevents a thorough identification of TE-rich and TE-poor regions and possible co-localization with CNV loci at the whole genome scale. Availability of long read-based more contiguous genome assembly in the future will certainly allow reinvestigating CNV and the possible involvement of TEs in association to an adaptive process such as resistance breaking down.

As previously evoked, in *M. incognita*, we found that the genome-wide pattern of variations of TE frequencies across the loci between the different populations recapitulated almost exactly the phylogeny of the isolates built on SNV in coding regions (Fig 4). Hence, most of the divergence in terms of TE pattern follows the divergence at the nucleotide level and thus the phylogeny of the isolates. Almost the same conclusion was drawn by comparing SNV and TE variation data across different *C. elegans* populations (Laricchia et al. 2017). In *M. incognita*, the phylogeny of isolates does not significantly correlate with the monitored biological traits, namely geographical distribution, range of compatible host plants and nature of the crop currently infected (Koutsovoulos et al. 2020). Interestingly, no correlation was also observed between variations in TE frequencies and geographical distribution for European *Drosophila* populations (Lerat et al. 2019). The lack of evident correlation between the phylogenetic signal regardless whether it is TE-based or SNV-based and the biological traits under consideration suggests most of the variations follow the drift between isolates and are not necessarily adaptive, which is not surprising. A similar conclusion was also drawn recently by analyzing 625 fungal genomes and observing that most TE movements were presumably neutral and adaptive ones being marginal (Muszewska et al. 2019).

On another note, as explained in the first section of the discussion, TE activity is possibly very recent in *M. incognita* and this might contribute to the current lack of evidence for association between TE activity, including invasion or decay across populations and adaptive traits.

Yet, we detected, and confirmed by PCR the neo-insertions of TE in some functionally important loci, inside genes or possible regulatory regions. We found that more than 90% of the TEs involved were TIRs or MITEs, which echoes their enrichment among the most active TEs in *M. incognita*. In the Mulberry genome, MITEs inserted near genes were shown to regulate gene expression via small RNAs while those inserted within genes were associated with alternative splice variants (Xin et al. 2019). Similarly, in the wheat genome, MITEs of the mariner superfamily played an instrumental role in generating the diversity of micro-RNAs involved in important adaptive traits such as resistance to pathogens (Poretti et al. 2020). The exact functional impact of TE insertions in *M. incognita* would need to be evaluated in the future. Generating transcriptomics data for the different isolates would enable studying associated differences in gene

expression patterns or transcript diversity. As a complementary approach, proteomic studies would allow direct search for differences at the encoded protein level.

Regardless of the future experimental validation of the functional impact, one important question concerns the current preliminary evidence for a possible role in the nematode adaptive evolution. Because some of the impacted genes are specific to plant-parasitic species and yet conserved in several of these phyto-parasites, a role in plant parasitism is possible. Interestingly, TE movements can be involved in the emergence of species or genus-specific 'orphan' genes (Ruiz-Orera et al. 2015; Wu and Knudson 2018; Jin et al. 2019). However, in the absence of known protein domains or functional characterization of these genes, the exact biochemical activity or biological processes in which they might be involved remains elusive.

Ploidy, (a)sexuality and hybridization: a complex interplay on TE load and composition

M. incognita is an asexual (mitotic parthenogenetic), polyploid, and hybrid species. These three features are expected to impact TE load in the genome with various intensities and possibly conflicting effects.

Contradictory theories exist concerning the activity/proliferation of TEs as a function of the reproductive mode. The higher efficacy of selection under sexual reproduction can be viewed as an efficient system to purge TEs and control their proliferation. Supporting these views, in parasitoid wasps, TE load was shown to be higher in asexual lineages induced by the endosymbiotic Wolbachia bacteria than in sexual lineages (Kraaijeveld et al. 2012). However, whether this higher load is a consequence of the shift in reproductive mode or of Wolbachia infection remains to be clarified.

In an opposite theory, sexual reproduction can also be considered as a way for TEs to spread across individuals within the population whereas in clonal reproduction the transposons are trapped exclusively in the offspring of the holding individual. Under this view, asexual reproduction is predicted to reduce TE load as TE are unable to spread in other individuals, and are thus removed by genetic drift and/or purifying selection in the long term (Wright and Finnegan 2001). Consistent with this theory, comparison of sexual and asexual *Saccharomyces cerevisiae* populations showed that the TE load decreases rapidly under asexual reproduction (Bast et al. 2019).

Hence, whether the TE-load is expected to be higher or lower in clonal species compared to sexual relatives remains unclear and other conflicting factors such as TE excision rate and the effective size of the population probably blur the signal (Glémin et al. 2019). The breeding system has been shown to constitute an important factor,of TE distribution in Caenorhabditis genomes (Dolgin et al. 2008): TEs in self-fertilizing populations seem to be selectively neutral and segregate at higher frequency than in outcrossing populations, where they are submitted to purifying selection. Interestingly, at a broader scale, a comparative analysis of different lineages of sexual and asexual arthropods revealed no evidence for differences in TE load according to the reproductive modes (Bast et al. 2015). Similar conclusions were drawn at the whole nematoda phylum scale (Szitenberg et al. 2016), although only one apomictic asexually-reproducing species (i.e. *M. incognita*) was present in the comparative analysis.

Polyploidy, in contrast, is commonly accepted as a major event initially favouring the multiplication and activity of TEs. This is clearly described with numerous examples in plants (Vicient and Casacuberta 2017) and some examples are also emerging in animals (Rodriguez and Arkhipova 2018). When hybridization and polyploidy are combined, this can lead to TE bursts in the genome. As originally proposed by Barbara McClintock, allopolyploidization produces a "genomic shock", a genome instability associated with the

relaxation of the TE silencing mechanisms and the reactivation of ancient TEs (McClintock 1984; Mhiri et al. 2019).

Hybridization, polyploidy and asexual reproduction are combined in *M. incognita* with relative effects on the TE load extremely challenging, if not impossible, to disentangle. Initial comparisons of the TE loads in three allopolyploid clonal *Meloidogyne* against a diploid facultative sexual relative suggested a higher TE load in the clonal species (Blanc-Mathieu et al. 2017). However, to differentiate the relative contribution of each of these three features to the *M. incognita* TE load, it would be necessary to conduct comparative analysis with a same method on diploid asexuals, on polyploid sexuals as well as on diploid asexuals in the genus *Meloidogyne,* and ideally with and without hybrid origin. So far, genomic sequences are only available for other polyploid clonal species, which are all suspected to have a hybrid origin (Blanc-Mathieu et al. 2017; Szitenberg et al. 2017; Koutsovoulos et al. 2019; Susič et al. 2020), and, apart from that, only two diploid facultative sexual species (Opperman et al. 2008; Somvanshi et al. 2018). Hence, further sampling of *Meloidogyne* species with diverse ploidy levels and reproductive modes will be necessary to disentangle the relative contribution of ploidy level, hybridization and reproductive mode on the TE abundance and composition.

Concluding remarks

In this study we used population genomics technique and statistical analyses of the results to assess whether TE might contribute to the genome dynamics of *M. incognita* and possibly to its adaptive evolution. Overall, we provided a body of evidence suggesting TE have been at least recently active and might still be active. With thousands of loci showing variations in TE presence frequencies across geographical isolates, there is a clear impact on the *M. incognita* genome plasticity. Some TE being neo-inserted in coding or regulatory regions might have a functional impact. Although no clear connection with a role in adaptive evolution could be made so far, based on the few impacted coding loci we experimentally checked in this study, this is not to be excluded given the current lack of large-scale functional information for this species. This pioneering study constitutes a valuable resource and opens new perspectives for future targeted investigation of the potential effect of TE dynamics on the evolution, fitness and adaptability of *M. incognita* as well as in the whole nematoda phylum.

Materials and Methods

Material

The genome of *M. incognita*

We used the genome assembly published in (Blanc-Mathieu et al. 2017) as a reference for TE prediction and annotation (ENA assembly accession GCA_900182535, bioproject PRJEB8714) as well as for read-mapping of the different geographical isolates (Koutsovoulos et al. 2020), used for prediction of TE presence frequencies.

Briefly, the triploid *M. incognita* genome is 185Mb long with ~12,000 scaffolds and a N50 length of ~38 kb. Although the genome is triploid, because of the high nucleotide divergence between the genome copies (8% on average), most of these genome copies have been correctly separated during genome assembly, which can be considered effectively haploid (Blanc-Mathieu et al. 2017; Koutsovoulos et al. 2020). This reference genome originally came from a *M. incognita* population from the Morelos region of Mexico and was reared on tomato plants from the offspring of one single female in our laboratory.

The genome of *C. elegans*

We used the *C. elegans* genome (The C. elegans Genome Sequencing Consortium 1998) assembly (PRJNA13758) to perform its repeatome prediction and annotation and compare our results to the literature as a methodological validation.

Genome reads for 12 *M. incognita* geographical isolates

To predict the presence frequencies at TE loci across different *M. incognita* isolates, we used whole-genome sequencing data from pools of individuals from 12 different geographical regions (sup. Fig S4 & sup. Table S10). One pool corresponds to the Morelos isolates used to produce the *M. incognita* reference genome itself, as described above. The 11 other pools correspond to different geographical isolates across Brazil as described in (Koutsovoulos et al. 2020).

All the samples were reared from the offspring of one single female and multiplied on tomato plants. Then, approximately 1 million individuals were pooled and sequenced by Illumina paired-end reads (2*150bp). Libraries sizes vary between 74 and 76 million reads (Koutsovoulos et al. 2020).

We used cutadapt-1.15 (Martin 2011) to trim adapters, discard small reads, and trim low-quality bases in reads boundaries (–max-n=5 -q 20,20 -m 51 -j 32 -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCA -A AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT). Then, for each library, we performed a fastqc v-0.11.8 (Andrew S., 2010: http://www.bioinformatics.babraham.ac.uk/projects/fastqc) analysis to evaluate the quality of the reads. FastQC results analyses showed that no additional filtering or cleaning step was needed and no further read was discarded.

Methods

We performed the statistical analysis and the graphical representation using R' v-3.6.3 and the following libraries: ggplot2, cowplot, reshape2, ggpubr, phangorn, tidyverse, and ComplexUpset. All codes and analysis workflows are publicly available in the INRAE Dataverse (Kozlowski 2020a; Kozlowski 2020c; Kozlowski, Da Rocha, et al. 2020). For experimental validations, see (Kozlowski, Hassanaly-Goulamhoussen, et al. 2020). A diagram recapitulating the main steps of the analysis has been provided in supplementary (sup. Fig S7); as well as a decision tree summarising the polymorphism characterisation (sup. Fig S8).

M. incognita and *C. elegans* repeatome predictions and annotations.

We predicted and annotated the *M. incognita* and *C. elegans* repeatomes following the same protocol as thoroughly explained in (Koutsovoulos et al. 2019). We define the repeatome as all the repeated sequences in the genome, excluding Simple Sequence Repeats (SSR) and microsatellites. Then, following the above-mentioned protocol, we further analysed each repeatome to isolate annotations with canonical signatures of Transposable Elements (TEs).

Below, we briefly explain each step and describe protocol adjustments.

Genome pre-processing.

Unknown nucleotides 'Ns' encompass 1.81% of *the M. incognita* reference genome and need to be trimmed before repeatome predictions. We created a modified version of the genome by splitting it at N stretches of length 11 or more and then trimming all N, using dbchunk.py from the REPET package (Quesneville et al. 2005; Flutre et al. 2011). As this increases genome fragmentation and may, in turn, lead to false positives in TE detection, we only kept chunks of length above the L90 chunk length threshold, which is 4,891 bp. This modified version of the genome was only used to perform the *de novo* prediction of the TE consensus library (below). The TE annotation was performed on the whole reference genome.

The *C. elegans* reference genome was entirely resolved (no N), at the chromosome-scale. Hence, we used the whole assembly as is to perform the *de novo* prediction analysis.

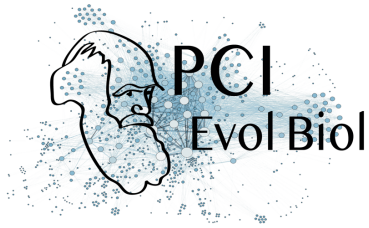
De novo prediction: constituting draft TE-consensus libraries.

For each species, we used the TEdenovo pipeline from the REPET package to generate a draft TE-consensus library..

Briefly, TEdenovo pipeline i) realises a self-alignment of the input genome to detect repetitions, ii) clusters the repetitions, iii) performs multiple alignments from the clustered repetitions to create consensus sequences, and eventually, iv) classify the consensus sequence following the Wicker's classification (Wicker et al. 2007) using structural and homology based information. One of the most critical steps of this process concerns the clustering of the repetitions as it requires prior knowledge about assembly ploidy and phasing quality.

We ran the analysis considering the modified *M. incognita* reference assembly previously described as triploid and set the 'minNbSeqPerGroup' parameter to 7 (*i.e* 2n+1). As the *C. elegans* assembly was haploid, we set the same parameter to 3.

All the remaining parameters values set in these analyses can be found in the TEdenovo configuration files (Kozlowski 2020a).



Automated curation of the TE-consensus libraries.

To limit the redundancy in the previously created TE consensus libraries and the false positives, we performed an automated curation step. Briefly, for each species, i) we performed a minimal annotation (steps 1, 2, 3, 7 of TEannot) of their genome with their respective draft TE-consensus libraries, and ii) only retained consensus sequences with at least one Full-Length Copy (FLC) annotated in the genome. All parameters values are described in the configuration files available in (Kozlowski 2020a).

Repeatome annotation

For each species, we performed a full annotation (steps 1, 2, 3, 4, 5, 7, and 8) of their genome with their respective cleaned TE-consensus libraries using TEannot from the REPET package. The obtained repeatome annotations (excluding SSR and microsatellites) were exported for further analyses. All parameters values are described in the configuration files available in (Kozlowski 2020a).

Repeatome post-processing: identifying annotations with canonical signatures of TEs.

Using in house scripts (Kozlowski 2020a), we analysed REPET outputs to retain annotations with canonical signatures of Transposable Elements (TEs) from the rest of the repeatomes. The same parameters were set for *M. incognita* and *C. elegans*. Briefly, for each species, we only conserved TE annotations i) classified as retro-transposons or DNA-transposons, ii) longer than 250 bp, iii) sharing more than 85% identity with their consensus sequence, iv) covering more than 33% of their consensus sequence length, v) first aligning with their consensus sequence in a BLAST analysis against the TE-consensus library, and vi) not overlapping with other annotations. TE annotations respecting all the described criterion were referred to as canonical TE annotations.

Putative transposition machinery identification (*M. incognita* only)

We analysed the *M. incognita* predicted proteome and transcriptome (Blanc-Mathieu et al. 2017) and crossed the obtained information with the canonical TE-annotation to identify TE containing genes putatively involved in the transposition machinery and evaluate TE-related gene expression levels in comparison to the rest of the genes in the genome.

Finding genes coding for proteins with TE-related HMM profiles

We performed an exhaustive HMMprofile search analysis on the whole *M. incognita* predicted proteome and then looked for proteins with TE-related domains. First, we concatenated two HMMprofile libraries into one: Pfram32 (Finn et al. 2016) library and Gypsy DB 2.0 (Llorens et al. 2011), a curated library of HMMprofiles linked to viruses, mobile genetic elements, and genomic repeats. Then, using this concatenated HMM profile library, we performed an exhaustive but stringent HMM profile search on the *M. incognita* proteome using hmmscan (-E 0.00001 --domE 0.001 --noali).

Eventually, using in house script (Kozlowski, Da Rocha, et al. 2020), we selected the best non-overlapping HMM profiles for each protein and then tagged corresponding genes with TE-related HMM profiles thanks to a knowledge-based function from the REPET tool 'profileDB4Repet.py'. We kept as genes with TE-related profiles all the genes with at least one TE-related HMM-profile identified.

Genes expression level

To determine the *M. incognita* protein-coding genes expression patterns, we used data from a previously published life-stage specific RNA-seq analysis of *M. incognita* transcriptome during tomato plant infection (Blanc-Mathieu et al. 2017). This analysis encompassed four different life stages: (i) eggs, (ii) pre-parasitic

second stage juveniles (J2), (iii) a mix of late parasitic J2, third stage (J3) and fourth stage (J4) juveniles and (iv) adult females, all sequenced in triplicates.

The cleaned RNA-seq reads were retrieved from the previous analysis and re-mapped to the *M. incognita* annotated genome assembly (Blanc-Mathieu et al. 2017) using a more recent version of STAR (2.6.1) (Dobin et al. 2013) and the more stringent end-to-end option (*i.e.* no soft clipping) in 2-passes. Expected read counts were calculated on the predicted genes from the *M. incognita* GFF annotation as FPKM values using RSEM (Li and Dewey 2011) to take into account the multi-mapped reads via expectation maximization. To reduce amplitude of variations, raw FPKM values were transformed to Log10(FPKM+1) and the median value over the 3 replicates was kept as a representative value in each life stage. The expression data are available in (Danchin and Da Rocha 2020).

Then, for each life stage independently, i) we ranked the gene expression values, and ii) defined gene expression level corresponding to the gene position in the ranking. We considered as substantially expressed all the genes that presented an expression level >= 1st quartile in at least one life stage.

TE annotations with potential transposition machinery

To identify TE-annotations including predicted genes involved in transposition machinery (inclusion >= 95% of the gene length), we performed the intersection of the canonical TE annotation and the genes annotation BED files (Kozlowski, Da Rocha, et al. 2020) using the intersect tool (-wo -s -F 0.95) from the bedtools v-2.27.1 suite (Quinlan and Hall 2010).

We then cross-referenced the obtained file with the list of the substantially expressed genes and the list of the TE-related genes previously elaborated to identify the TEs containing potential transposition machinery genes and their expression levels.

Evaluation of TE presence frequencies across the different *M. incognita* isolates

We used the popoolationTE2 v-1.10.04 pipeline (Kofler et al. 2016) to compute isolate-related support frequencies of both annotated, and *de novo* TE-loci across the 12 *M. incognita* geographical isolates previously described. To that end, we performed a 'joint' analysis as recommended by the popoolationTE2 manual. Briefly, popoolationTE2 uses both quantitative and qualitative information extracted from paired-end (PE) reads mapping on the TE-annotated reference genome and a set of reference TE sequences to detect signatures of TE polymorphisms and estimate their frequencies in every analysed isolate. Frequency values correspond to the proportion of individuals in an isolate for which a copy of the TE is present at a given locus.

Preparatory work: creating the TE-hierarchy and the TE-merged-reference files.

We used the canonical TE-annotation set created above (Kozlowski 2020a) and the *M. incognita* reference genome to produce the TE-merged reference file and the TE-hierarchy file necessary to perform the popoolationTE analysis (Kozlowski 2020c).

We used getfasta and maskfasta commands (default parameters) from the bedtools suite to respectively extract and mask the sequences corresponding to canonical TE-annotations in the reference genome. Then we concatenated both resulting sequences in a 'TE-merged reference' multi fasta file. The 'TE-hierarchy' file was created from the TE-annotation file from which it retrieves and stores the TE sequence name, the family, and the TE-order for every entry.

Reads mapping

For each *M. incognita* isolate library, we mapped forward and reverse reads separately on the "TE-merged-references" genome-TE file using the local alignment algorithm bwa bwasw v-0.7.17-r1188 (Li and Durbin 2009) with the default parameters. The obtained sam alignment files were then converted to bam files using samtools view v-1.2 (Li et al. 2009).

Restoring paired-end information and generating the ppileup file.

We restored paired-end information from the previous separate mapping using the sep2pe (--sort) tool from popoolationTE2-v1.10.03. Then, we created the ppileup file using the 'ppileup' tool from popolationTE2 with a map quality threshold of 15 (--map-qual 15).

For every base of the genome, this file summarises the number of PE reads inserts spanning the position (physical coverage) but also the structural status inferred from paired-end read covering this site.

Estimating target coverage and subsampling the ppileup to a uniform coverage

As noticed by R. Kofler, heterogeneity in physical coverage between populations may lead to discrepancies in TE frequency estimation. Hence, we flattened the physical coverage across the *M. incognita* isolates by a subsampling and a rescaling approach.

We first estimated the optimal target coverage to balance information loss and homogeneity using the 'stats-coverage' tool from PopoolationTE2 (default parameter) and set this value to 15X. We then used the 'subsamplePpileup' tool (--target-coverage 15) to discard positions with a physical coverage below 15X and rescale the coverage of the remaining position to that value.

Identify signatures of TE polymorphisms

We identified signatures of TE polymorphisms from the previously subsampled file using the 'identifySignature' tool following the joint algorithm (--mode joint; --min-count 2; --signature-window minimumSampleMedian; --min-valley minimumSampleMedian).

Then, for each identified site, we estimated TE frequencies in each isolate using the 'frequency' tool (default parameters). Eventually, we paired up the signatures of TE polymorphisms using 'pairupSignatures' tool (--min-distance -200; --max-distance -- 300 as recommended by R. Kofler), yielding a final list of potential TE-polymorphisms positions in the reference genome with their associated frequencies for each one of the isolates.

Evaluation of PopoolationTE2 systematic error rate in the TE-frequency estimation.

To estimate PopoolationTE2 systematic error rate in the TE-frequency estimation, we ran the same analysis (from the PE information restoration step) but comparing each isolate against itself (12 distinct analyses).

We then analysed each output individually, measuring the frequency difference between the two 'replicates' in all the detected loci with FR signatures (see below for more explanations).

We tested the homogeneity of the frequency-difference across the 12 analyses with an ANOVA and concluded that the mean values of the frequencies differences between the analysis were not significantly heterogeneous (p. value = 0.102 > 0.05). Hence, we concatenated the 12 analysis frequency-difference and set the systematic error rate in the TE-frequency estimation to 2 times the standard deviation of the frequency differences, a value of 0.97 % .

TE polymorphism analysis

Isolating TE loci with frequency variation across M. incognita isolates.

We parsed PopoolationTE2 analysis output to identify TE loci with enough evidence to characterise them as polymorphic in frequency across the isolates.

PopoolationTE2 output informs for each detected locus i) its position on the reference genome, ii) its frequency value for every sample of the analysis (e.g each isolate), and iii) qualitative information about the reads mapping signatures supporting a TE insertion.

In opposition to separate Forward ('F') or Reverse ('R') signatures, 'FR' signatures mean the locus both boundaries are supported by significant physical coverage. Entries with such type of signature are more accurate in terms of frequency and position estimation. Hence, we only retained candidate loci with 'FR' signatures. Then, for each locus, we computed the maximal frequency variation between all the isolates and discarded the loci with a frequency difference smaller than the PopoolationTE2 systematic error rate in the TE-frequency estimation we computed (0.97 %; see above). We also discarded loci where different TEs were predicted to be inserted. We considered the remaining loci as polymorphic in frequency across the isolates.

Isolates phylogeny

We reconstructed *M. incognita* isolates phylogeny according to their patterns of polymorphism in TE frequencies.

We first computed a euclidean distance matrix from the isolates TE frequencies of all the detected polymorphic loci. We then used the distance matrix to construct the phylogenetic tree using the Neighbor Joining (NJ) method (R' phangorn package v-2.5.5). We computed nodes support values with a bootstrap approach (n=500 replicates) using the boot.phylo function from the ape-v5.4 R package (Paradis and Schliep 2019). The boot.phylo function performs a resampling of the frequency matrix (here the matrix with loci in columns, isolates in row, and values corresponding to the TE presence frequencies).

Also, we created a phylogenetic tree using the SNV from coding regions for all isolates with raxml-ng v-0.9.0 (Kozlov et al. 2019) utilising the model GTR+G+ASC_LEWIS and performing 100 bootstrap replicates. We compared both topologies using Itol v-4.0 viewer (Letunic and Bork 2019).

Polymorphisms characterisation.

We exported the polymorphic TE positions as an annotation file, and we used bedtools intersect (-wao) to perform their intersection with the reference canonical TE annotation. We then cross-referenced the results with the filtered popoolationTE2 output and defined a decision tree to characterise the TE-polymorphism detected by popoolationTE2 as 'reference-TE polymorphism' (ref-polymorphism), 'extra-detection', or 'neo-insertion' (sup Fig S8).

We considered a reference TE-annotation as polymorphic (e.g. ref-polymorphism locus) if:

i) The position of the polymorphism predicted by PoPoolationTE2 falls between the boundaries of the reference TE-annotation

ii) Both the reference TE-annotation and the predicted polymorphism belong to the same TE-consensus sequence.

iii) The TE has a predicted frequency > 75% in the reference isolate Morelos.

Canonical TE-annotations that did not intersect with polymorphic loci predicted by PopoolationTE2, or that presented frequency variations <1% across the isolates were considered as non-polymorphic.

We classified as 'neo-insertions' all the polymorphic loci for which no canonical TE was predicted in the reference annotation (polymorphism position is not included in a reference TE-annotation), but which were detected with a frequency > 25% in at least one isolate different from the reference isolate Morelos, in which the TE frequency should be inferior to 1% and thus considered truly absent in the reference genome.

Finally, we classified as 'extra-detection' all the polymorphic loci which did not correspond to a reference annotation but which were detected with a frequency > 25% in the reference isolate Morelos (at least). Polymorphic loci having a frequency between 1% and 25% in Morelos isolate were considered ambiguous and were discarded.

Then, for each TE polymorphism, we investigated the homogeneity of the TE frequency between the isolates.We considered TE frequency was homogeneous between isolates when the maximum frequency variation between isolate was <= to 25%. Above this value, we considered the TE presence frequency was heterogeneous between isolates.

Highly Contrasted Polymorphic TE loci (HCPTEs): isolation, characterisation and experimental validation.

HCPTEs isolation

We considered as highly contrasted all the polymorphic loci for which i) all the isolates had frequency values either < 25% or > 75%, ii) at least one isolate showed a frequency < 25 % while another presented a frequency > 75%. Polymorphic loci fitting with these requirements were exported as an annotation file in the bed format.

HCPTEs possible functional impact

We first identified the genes potentially impacted by the HCPTEs by cross-referencing the HCPTEs annotation file with the gene annotation file, using the bedtools suite. We used the 'closest' program (-D b -fu -io; b being the gene annotation file) to identify the closest (but not intersecting) gene downstream each HCPTE. We only retained the entries with a maximum distance of 1 kb between the HCPTE and gene boundaries. We identified the insertions in the gene using the 'intersect' tool (-wo).

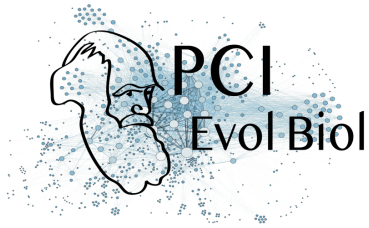
Then, we performed a manual bioinformatic functional analysis for each gene potentially impacted by HCPTEs. Protein sequences were extracted from the *M. incognita* predicted proteome (Blanc-Mathieu et al. 2017) and blasted (blastp; default parameters) against the Non-Redundant protein sequences database (NR) from the NCBI (https://blast.ncbi.nlm.nih.gov/). The same sequences were also used on the InterProScan website (https://www.ebi.ac.uk/interpro/) to perform an extensive search on all the available libraries of conserved protein domains and motifs.

Then, for each gene potentially impacted by HCPTEs, we performed an orthology search on the Wormbase Parasite website (https://parasite.wormbase.org/) using genes accession numbers and the pre-computed ENSEMBL Compara orthology prediction (Herrero et al. 2016).

Finally, we analysed the expression levels of the genes potentially impacted by HCPTEs extracting the information from the RNA-seq analysis of four *M. incognita* life-stages performed previously (see Putative transposition machinery identification section).

Experimental validation of Highly Contrasted Polymorphic TE loci

To experimentally validate in-silico predictions of TE neo-insertions with potential functional impact, we selected 5 candidates among the HCPTEs loci and performed a PCR experiment. To run this experiment, we



used DNA remaining from extractions performed on the M. incognita isolates for a previous population genomics analysis (Koutsovoulos et al. 2020). We selected loci to be validated based on the following criteria:

- The predicted insertion must be in a genic or potential regulatory region (max 1kb upstream of a gene) as the most evident criterion for a potential functional impact.
- The element must be short enough (2.5kb max) to be amplified by PCR and SANGER sequenced using standard techniques and material.
- To validate the predicted impacted gene actually exists, it must be supported by substantial expression data in the reference isolate Morelos.
- To maximize the chances the genes have effects on biological traits characteristic of the root-knot nematodes, the impacted gene must be *Meloidogyne*-specific.

Once all these criteria were applied, we maximized the diversity of TE orders involved and this resulted in the 5 loci presented in the results section.

Primer design and PCR amplification.

We designed primers for the PCR analysis using the Primer3Plus web interface (Untergasser et al. 2007). The set of 10 primers with the corresponding sequence and expected amplicon sizes with, or without TE insertion, is shown in (sup. Table 11 & (Kozlowski, Hassanaly-Goulamhoussen, et al. 2020)). We used primers amplifying the whole actin-encoding gene (Minc3s00960g19311) as positive control.

PCR experiments were performed on *M. incognita* Morelos isolate and 11 Brazilian isolates: R1-2, R1-3, R1-6, R2-1, R2-6, R3-1, R3-2, R3-4, R4-1, R4-3 and R4-4.

R3-1 presented no amplification in any of the tested loci nor the positive control (actin) and was thus discarded from this analysis.

PCR mixture contained 0.5µmol of each primer, 1x MyTaq™ reaction buffer and 1.0 U of MyTaq™ DNA polymerase (Bioline Meridian Bioscience) adjusted to a total volume of 20µL. PCR amplification was performed with a TurboCycler2 (Blue-Ray Biotech Corp.). PCR conditions were as follows: initial denaturation at 95°C for 5 min, followed by 35 cycles of 95°C for 30 s, 56°C for 30 s of annealing, and 72°C for 3 min of extension, the program ending with a final extension at 72°C for 10 min. Aliquots of 5µL were migrated by electrophoresis on a 1% agarose gel (Sigma Chemical Co.) for 70 min at 100 V. The size marker used is 1kb Plus DNA Ladder (New England Biolabs Inc.), containing the following size fragments in bp: 100, 200, 300, 400, 500, 600, 700, 900, 1000, 1200, 1500, 2000, 3000, 4000, 5000, 6000, 8000 and 10000.

Purification and sequencing of PCR amplicons.

Amplicon bands were revealed using ethidium bromide and exposure to ultraviolet radiation. PCR products bands were excised from the agarose gel with a scalpel and purified using MinElute Gel Extraction Kit (Qiagen) before sequencing, following the manufacturer's protocol. PCR products were sequenced by Sanger Sequencing (Eurofins Genomics).

Forward (F) and Reverse (R) sequences were blasted individually (https://blast.ncbi.nlm.nih.gov/ ; Optimised for 'Somewhat similar sequences', default parameters) to the expected TE-consensus sequence and to the genomic region surrounding the predicted insertion point (2 kb region: 1kb upstream the predicted insertion point and 1kb downstream). When no significant hit was found, the sequence was blasted against the *Meloidogyne* reference genomes available (https://meloidogyne.inrae.fr/), the whole TE-consensus library, and the NR database on the NCBI blast website.

Data accessibility

All the raw and filtered data generated in this study as well as details of the experimental procedures, scripts and codes have been deposited and made publicly available in the institutional INRAE Data Portal at this URL: https://data.inrae.fr/dataverse/TE-mobility-in-MiV3 and cited throughout the text where appropriate, with DOIs available in the references.

Supplementary material

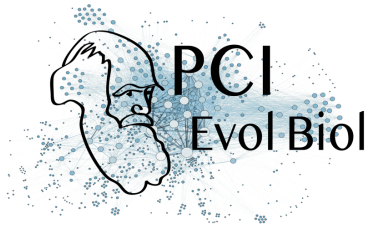
Supplementary material, tables and figures accompany this article and are available online in bioRxiv as a single PDF file.

Acknowledgements

The authors would like to thank Joffrey Mejias for all his advice and all the inspiring discussion. The authors are grateful to Laetitia Perfus-Barbeoch for her advice and support in the experimental validation of TE movements. The authors would like to thank Erika VS Albuquerque for her help and assistance in accessing the DNA extractions from the *M. incognita* Brazilian Isolates. We would also like to thank the BIG bioinformatics platform from the PlantBios infrastructure as well as the URGI team for providing facilities and technical support. This work has been supported by the French government, through the UCA-JEDI "Investments in the Future" project managed by the National Research Agency (ANR) with the reference number ANR-15-IDEX-01. Version 4 of this preprint has been peer-reviewed and recommended by Peer Community In Evolutionary Biology (https://doi.org/10.24072/pci.evolbiol.100106)

Conflict of interest disclosure

The authors of this preprint declare that they have no financial conflict of interest with the content of this article.



References

Abad P, Gouzy J, Aury J-M, Castagnone-Sereno P, Danchin EGJ, Deleury E, Perfus-Barbeoch L, Anthouard V, Artiguenave F, Blok VC, et al. 2008. Genome sequence of the metazoan plant-parasitic nematode Meloidogyne incognita. *Nat. Biotechnol.* 26:909–915.

Agrios GN. 2005. Plant Pathology, 5th Edition. Burlington, USA: Elsevier Academic Press

Aminetzach YT, Macpherson JM, Petrov DA. 2005. Pesticide Resistance via Transposition-Mediated Adaptive Gene Truncation in Drosophila. *Science* 309:764–767.

Ansaloni F, Scarpato M, Di Schiavi E, Gustincich S, Sanges R. 2019. Exploratory analysis of transposable elements expression in the C. elegans early embryo. *BMC Bioinformatics* 20:484.

Anxolabéhère D, Kidwell MG, Periquet G. 1988. Molecular characteristics of diverse populations are consistent with the hypothesis of a recent invasion of Drosophila melanogaster by mobile P elements. *Mol. Biol. Evol.* 5:252–269.

Barbary A, Djian-Caporalino C, Palloix A, Castagnone-Sereno P. 2015. Host genetic resistance to root-knot nematodes, Meloidogyne spp., in Solanaceae: from genes to the field. *Pest Manag. Sci.* 71:1591–1598.

Bast J, Jaron KS, Schuseil D, Roze D, Schwander T. 2019. Asexual reproduction reduces transposable element load in experimental yeast populations.Coop G, Tautz D, Coop G, Charlesworth B, editors. *eLife* 8:e48548.

Bast J, Schaefer I, Schwander T, Maraun M, Scheu S, Kraaijeveld K. 2015. No Accumulation of Transposable Elements in Asexual Arthropods. *Mol. Biol. Evol.*:msv261.

Bégin M, Schoen DJ. 2007. Transposable Elements, Mutational Correlations, and Population Divergence in Caenorhabditis Elegans. *Evolution* 61:1062–1070.

Belyayev A. 2014. Bursts of transposable elements as an evolutionary driving force. *J. Evol. Biol.* 27:2573–2584.

Bessereau J-L. 2006. Transposons in C. elegans. *WormBook* [Internet]. Available from: http://www.wormbook.org/chapters/www_transposons/transposons.html

Blanc-Mathieu R, Perfus-Barbeoch L, Aury J-M, Rocha MD, Gouzy J, Sallet E, Martin-Jimenez C, Bailly-Bechet M, Castagnone-Sereno P, Flot J-F, et al. 2017. Hybridization and polyploidy enable genomic plasticity without sex in the most devastating plant-parasitic nematodes. *PLOS Genet.* 13:e1006777.

Bourgeois Y, Boissinot S. 2019. On the Population Dynamics of Junk: A Review on the Population Genomics of Transposable Elements. *Genes* 10:419.

Castagnone-Sereno P. 2006. Genetic variability and adaptive evolution in parthenogenetic root-knot nematodes. *Heredity* 96:282–289.

Castagnone-Sereno P, Danchin EGJ. 2014. Parasitic success without sex – the nematode experience. *J. Evol. Biol.* 27:1323–1333.

Castagnone-Sereno P, Mulet K, Danchin EGJ, Koutsovoulos GD, Karaulic M, Rocha MD, Bailly-Bechet M, Pratx L, Perfus-Barbeoch L, Abad P. 2019. Gene copy number variations as signatures of adaptive evolution in the parthenogenetic, plant-parasitic nematode Meloidogyne incognita. *Mol. Ecol.* 28:2559–2572.

Castagnone-Sereno P, Wajnberg E, Bongiovanni M, Leroy F, Dalmasso A. 1994. Genetic variation inMeloidogyne incognita virulence against the tomatoMi resistance gene: evidence from isofemale line selection studies. *Theor. Appl. Genet.* 88:749–753.

Danchin E, Da Rocha M. 2020. M. incognita protein-coding genes expression patterns. *Portail Data INRAE* [Internet]. Available from: https://doi.org/10.15454/YM2DHE

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29:15–21.

Dolgin ES, Charlesworth B, Cutter AD. 2008. Population frequencies of transposable elements in selfing and

outcrossing Caenorhabditis nematodes. *Genet. Res.* 90:317–329.

Emmons SW, Yesner L. 1984. High-frequency excision of transposable element Tc1 in the nematode caenorhabditis elegans is limited to somatic cells. *Cell* 36:599–605.

Faino L, Seidl MF, Shi-Kunne X, Pauper M, Berg GCM van den, Wittenberg AHJ, Thomma BPHJ. 2016. Transposons passively and actively contribute to evolution of the two-speed genome of a fungal pathogen. *Genome Res.* [Internet]. Available from: http://genome.cshlp.org/content/early/2016/07/12/gr.204974.116

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, et al. 2016. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* 44:D279–D285.

Flutre T, Duprat E, Feuillet C, Quesneville H. 2011. Considering Transposable Element Diversification in De Novo Annotation Approaches. *PLoS ONE* 6:e16526.

Glémin S, François CM, Galtier N. 2019. Genome Evolution in Outcrossing vs. Selfing vs. Asexual Species. In: Anisimova M, editor. Evolutionary Genomics: Statistical and Computational Methods. Methods in Molecular Biology. New York, NY: Springer. p. 331–369. Available from: https://doi.org/10.1007/978-1-4939-9074-0_11

Gross SM, Williamson VM. 2011. Tm1: A Mutator/Foldback Transposable Element Family in Root-Knot Nematodes. *PLoS ONE* 6:e24534.

Guerreiro MPG. 2014. Interspecific hybridization as a genomic stressor inducing mobilization of transposable elements in *Drosophila*. *Mob. Genet. Elem.* 4:e34394.

Herrero J, Muffato M, Beal K, Fitzgerald S, Gordon L, Pignatelli M, Vilella AJ, Searle SMJ, Amode R, Brent S, et al. 2016. Ensembl comparative genomics resources. *Database* [Internet] 2016. Available from: https://academic.oup.com/database/article/doi/10.1093/database/bav096/2630091

Hill WG, Robertson A. 1966. The effect of linkage on limits to artificial selection. *Genet. Res.* 8:269–294.

Hoffmann AA, Reynolds KT, Nash MA, Weeks AR. 2008. A high incidence of parthenogenesis in agricultural pests. *Proc. R. Soc. Lond. B Biol. Sci.* 275:2473–2481.

Jin G-H, Zhou Y-L, Yang H, Hu Y-T, Shi Y, Li L, Siddique AN, Liu C-N, Zhu A-D, Zhang C-J, et al. 2019. Genetic innovations: Transposable element recruitment and de novo formation lead to the birth of orphan genes in the rice genome. *J. Syst. Evol.* [Internet] n/a. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/jse.12548

Jones JT, Haegeman A, Danchin EGJ, Gaur HS, Helder J, Jones MGK, Kikuchi T, Manzanilla-López R, Palomares-Rius JE, Wesemael WML, et al. 2013. Top 10 plant-parasitic nematodes in molecular plant pathology. *Mol. Plant Pathol.* 14:946–961.

Kanazawa A, Liu B, Kong F, Arase S, Abe J. 2009. Adaptive Evolution Involving Gene Duplication and Insertion of a Novel Ty1/copia-Like Retrotransposon in Soybean. *J. Mol. Evol.* 69:164–175.

Katju V, Bergthorsson U. 2013. Copy-number changes in evolution: rates, fitness effects and adaptive significance. *Front. Genet.* [Internet] 4. Available from: http://www.frontiersin.org/Evolutionary_and_Population_Genetics/10.3389/fgene.2013.00273/abstract

Kazazian HH, Wong C, Youssoufian H, Scott AF, Phillips DG, Antonarakis SE. 1988. Haemophilia A resulting from de novo insertion of L 1 sequences represents a novel mechanism for mutation in man. *Nature* 332:164–166.

Kofler R, Gómez-Sánchez D, Schlötterer C. 2016. PoPoolationTE2: Comparative Population Genomics of Transposable Elements Using Pool-Seq. *Mol. Biol. Evol.* 33:2759–2764.

Kondrashov AS. 1988. Deleterious mutations and the evolution of sexual reproduction. *Nature* 336:435–440.

Koutsovoulos GD, Marques E, Arguel M-J, Duret L, Machado ACZ, Carneiro RMDG, Kozlowski DK, Bailly-Bechet M, Castagnone-Sereno P, Albuquerque EVS, et al. 2020. Population genomics supports clonal

reproduction and multiple independent gains and losses of parasitic abilities in the most devastating nematode pest. *Evol. Appl.* 13:442–457.

Koutsovoulos GD, Poullet M, Ashry AE, Kozlowski DK, Sallet E, Rocha MD, Martin-Jimenez C, Perfus-Barbeoch L, Frey J-E, Ahrens C, et al. 2019. The polyploid genome of the mitotic parthenogenetic root knot nematode Meloidogyne enterolobii. *bioRxiv*:586818.

Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. 2019. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* 35:4453–4455.

Kozlowski D. 2020a. Transposable Elements prediction and annotation in the M. incognita genome. *Portail Data INRAE* [Internet]. Available from: https://doi.org/10.15454/EPTDOS

Kozlowski D. 2020b. Transposable Elements prediction and annotation in the C. elegans genome. *Portail Data INRAE* [Internet]. Available from: https://doi.org/10.15454/LQCIW0

Kozlowski D. 2020c. TE polymorphisms detection and analysis with PopoolationTE2. *Portail Data INRAE* [Internet]. Available from: https://doi.org/10.15454/EWJCT8

Kozlowski D, Da Rocha M, Danchin E. 2020. TE-related genes: annotation, characterisation, and expression. *Portail Data INRAE* [Internet]. Available from: https://doi.org/10.15454/DLDJVF

Kozlowski D, Hassanaly-Goulamhoussen R, Danchin E. 2020. Experimental validations of TE-impacted coding or regulatory loci. *Portail Data INRAE* [Internet]. Available from: https://doi.org/10.15454/NQAF31

Kraaijeveld K, Zwanenburg B, Hubert B, Vieira C, De Pater S, Van Alphen JJM, Den Dunnen JT, De Knijff P. 2012. Transposon proliferation in an asexual parasitoid. *Mol. Ecol.* 21:3898–3906.

Laricchia KM, Zdraljevic S, Cook DE, Andersen EC. 2017. Natural Variation in the Distribution and Abundance of Transposable Elements Across the Caenorhabditis elegans Species. *Mol. Biol. Evol.* 34:2187–2202.

Lerat E, Goubert C, Guirao-Rico S, Merenciano M, Dufour A-B, Vieira C, González J. 2019. Population-specific dynamics and selection patterns of transposable element insertions in European natural populations. *Mol. Ecol.* 28:1506–1522.

Letunic I, Bork P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* 47:W256–W259.

Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12:323.

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

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

Lively CM. 2010. A Review of Red Queen Models for the Persistence of Obligate Sexual Reproduction. *J. Hered.* 101:S13–S20.

Llorens C, Futami R, Covelli L, Domínguez-Escribá L, Viu JM, Tamarit D, Aguilar-Rodríguez J, Vicente-Ripolles M, Fuster G, Bernet GP, et al. 2011. The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. *Nucleic Acids Res.* 39:D70–D74.

Lu L, Chen J, Robb SMC, Okumoto Y, Stajich JE, Wessler SR. 2017. Tracking the genome-wide outcomes of a transposable element burst over decades of amplification. *Proc. Natl. Acad. Sci.* 114:E10550–E10559.

Magwire MM, Bayer F, Webster CL, Cao C, Jiggins FM. 2011. Successive Increases in the Resistance of Drosophila to Viral Infection through a Transposon Insertion Followed by a Duplication. *PLOS Genet.* 7:e1002337.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17:10–12.

McCarter JP. 2009. Molecular Approaches Toward Resistance to Plant-Parasitic Nematodes. In: Berg RH, Taylor CG, editors. Cell Biology of Plant Nematode Parasitism. Vol. 15. Plant Cell Monographs. Berlin, Heidelberg: Springer Berlin Heidelberg. p. 239–267. Available from:

http://www.springerlink.com/index/10.1007/978-3-540-85215-5_9

McClintock B. 1984. The significance of responses of the genome to challenge. *Science* 226:792–801.

Mhiri C, Parisod C, Daniel J, Petit M, Lim KY, Borne FD de, Kovarik A, Leitch AR, Grandbastien M-A. 2019. Parental transposable element loads influence their dynamics in young Nicotiana hybrids and allotetraploids. *New Phytol.* 221:1619–1633.

Miki Y, Nishisho I, Horii A, Miyoshi Y, Utsunomiya J, Kinzler KW, Vogelstein B, Nakamura Y. 1992. Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. *Cancer Res.* 52:643–645.

Muller HJ. 1964. The Relation of Recombination to Mutational Advance. *Mutat Res* 106:2–9.

Muszewska A, Steczkiewicz K, Stepniewska-Dziubinska M, Ginalski K. 2019. Transposable elements contribute to fungal genes and impact fungal lifestyle. *Sci. Rep.* 9:4307.

Opperman CH, Bird DM, Williamson VM, Rokhsar DS, Burke M, Cohn J, Cromer J, Diener S, Gajan J, Graham S, et al. 2008. Sequence and genetic map of Meloidogyne hapla: A compact nematode genome for plant parasitism. *Proc Natl Acad Sci U A* 105:14802–14807.

Paradis E, Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35:526–528.

Pereira V, Enard D, Eyre-Walker A. 2009. The Effect of Transposable Element Insertions on Gene Expression Evolution in Rodents. *PLoS ONE* [Internet] 4. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2629548/

Poretti M, Praz CR, Meile L, Kälin C, Schaefer LK, Schläfli M, Widrig V, Sanchez-Vallet A, Wicker T, Bourras S. 2020. Domestication of High-Copy Transposons Underlays the Wheat Small RNA Response to an Obligate Pathogen. *Mol. Biol. Evol.* 37:839–848.

Quesneville H, Bergman CM, Andrieu O, Autard D, Nouaud D, Ashburner M, Anxolabehere D. 2005. Combined evidence annotation of transposable elements in genome sequences. *PLoS Comput Biol* 1:166–175.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841–842.

Rice WR. 2002. Experimental tests of the adaptive significance of sexual recombination. *Nat. Rev. Genet.* 3:241–251.

Rodriguez F, Arkhipova IR. 2018. Transposable elements and polyploid evolution in animals. *Curr. Opin. Genet. Dev.* 49:115–123.

Ruiz-Orera J, Hernandez-Rodriguez J, Chiva C, Sabidó E, Kondova I, Bontrop R, Marqués-Bonet T, Albà MM. 2015. Origins of De Novo Genes in Human and Chimpanzee. *PLOS Genet.* 11:e1005721.

Savary S, Willocquet L, Pethybridge SJ, Esker P, McRoberts N, Nelson A. 2019. The global burden of pathogens and pests on major food crops. *Nat. Ecol. Evol.* 3:430–439.

Somvanshi VS, Tathode M, Shukla RN, Rao U. 2018. Nematode Genome Announcement: A Draft Genome for Rice Root-Knot Nematode, Meloidogyne graminicola. *J. Nematol.* 50:111–116.

Stuart T, Eichten SR, Cahn J, Karpievitch YV, Borevitz JO, Lister R. 2016. Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation. *eLife* 5:e20777.

Susič N, Koutsovoulos GD, Riccio C, Danchin EGJ, Blaxter ML, Lunt DH, Strajnar P, Širca S, Urek G, Stare BG. 2020. Genome sequence of the root-knot nematode Meloidogyne luci. *J. Nematol.* 52:1–5.

Szitenberg A, Cha S, Opperman CH, Bird DM, Blaxter ML, Lunt DH. 2016. Genetic drift, not life history or RNAi, determine long term evolution of transposable elements. *Genome Biol. Evol.*:evw208.

Szitenberg A, Salazar-Jaramillo L, Blok VC, Laetsch DR, Joseph S, Williamson VM, Blaxter ML, Lunt DH. 2017. Comparative Genomics of Apomictic Root-Knot Nematodes: Hybridization, Ploidy, and Dynamic Genome Change. *Genome Biol. Evol.* 9:2844–2861.

The C. elegans Genome Sequencing Consortium. 1998. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* 282:2012–2018.

Trudgill DL, Blok VC. 2001. Apomictic, polyphagous root-knot nematodes: exceptionally successful and damaging biotrophic root pathogens. *Annu Rev Phytopathol* 39:53–77.

Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, Leunissen JAM. 2007. Primer3Plus, an enhanced web interface to Primer3. *Nucleic Acids Res.* 35:W71–W74.

Vicient CM, Casacuberta JM. 2017. Impact of transposable elements on polyploid plant genomes. *Ann. Bot.* 120:195–207.

Vrijenhoek RC, Parker ED. 2009. Geographical Parthenogenesis: General Purpose Genotypes and Frozen Niche Variation. In: Schön I, Martens K, Dijk P, editors. Lost Sex. Dordrecht: Springer Netherlands. p. 99–131. Available from: http://www.springerlink.com/index/10.1007/978-90-481-2770-2_6

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. *Nat. Rev. Genet.* 8:973–982.

Wright S, Finnegan D. 2001. Genome evolution: Sex and the transposable element. *Curr. Biol.* 11:R296–R299.

Wu B, Knudson A. 2018. Tracing the De Novo Origin of Protein-Coding Genes in Yeast. *mBio* [Internet] 9. Available from: https://mbio.asm.org/content/9/4/e01024-18

Xin Y, Ma B, Xiang Z, He N. 2019. Amplification of miniature inverted-repeat transposable elements and the associated impact on gene regulation and alternative splicing in mulberry (Morus notabilis). *Mob. DNA* 10:27.

Zeng L, Pederson SM, Kortschak RD, Adelson DL. 2018. Transposable elements and gene expression during the evolution of amniotes. *Mob. DNA* [Internet] 9. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5998507/

VII – Evolution du contenu en ETs au cours du temps au sein d'un isolat de *M. incognita*

1 - Avant-propos

Cette étude (préliminaire) est le fruit d'une collaboration avec Rahim HASSANALY GOULAMHOUSSEN (RHG) (https://orcid.org/0000-0002-0811-7217). J'ai réalisé les prédictions bio-informatiques et la conception générale de l'analyse. RHG et moi-même avons mis en place le protocole de validation expérimentale. RHG a réalisé les validations expérimentales.

2 - Contexte & Design expérimental

Précédemment, nous avons comparé le contenu en ETs de différents isolats géographiques de l'espèce *M. incognita* (Kozlowski et al., 2020). Dans cette étude, nous avons pu montrer que plusieurs milliers de loci d'ET présentent des variations de fréquence (de la présence de l'ET) entre 12 isolats de *M. incognita*, signe d'une activité au moins récente des ETs au sein de cette espèce. Par ailleurs, nous avons aussi pu constater que, bien que l'ensemble des isolats présentent des profils individuels de fréquence de présence d'ETs hautement bimodaux avec un pic à 0 (absence) et l'autre à 1 (fixation dans l'isolat), une part faible mais non négligeable des loci présentaient des fréquences intermédiaires (comprises entre 0.25 et 0.75) ; avec des variations indépendantes pour chaque isolat étudié. Ces résultats suggèrent qu'il existe une variabilité du contenu en ETs intra-isolat en plus de celle observée inter-isolats chez *M. incognita*.

Cette observation est surprenante. En effet, nous étudions ici un organisme à reproduction asexuée et ne faisant pas de méiose et dont la descendance est par conséquent supposée présenter très peu de variations entre individus voir être clonale. Pour chacun des isolats étudiés, la descendance d'une seule femelle a été récupérée (i.e. la masse d'oeufs) et celle-ci a été multipliée indépendamment pendant plusieurs générations. De ce fait, dans le cas « idéal » où chaque génération produirait le même nombre de descendants et où la pression de sélection exercée serait la même, on s'attendrait intuitivement à ce qu'à l'intérieur d'un isolat, le contenu en ETs reste très proche voir identique à celui de la génitrice

originelle au cours du temps et à ce que de potentielles nouvelles insertions restent minoritaires en fréquence.

De cette idée sont nées les deux questions suivantes :

- Est-ce que la composition en ETs (fréquence de présence des ETs) varie au cours du temps au sein d'un même isolat issu d'une espèce supposée clonale, et en particulier en un faible nombre de générations ? Si oui, ces variations suivent-elle une dynamique particulière ?
- Sommes-nous capables de quantifier expérimentalement ces variations de fréquences ?

Afin de répondre à ces questions, j'ai mis en place le design expérimental suivant :

- Réaliser une analyse *in silico* à partir de données de séquençages réalisés à plusieurs points de cinétique sur un même isolat afin de prédire des loci d'ET dont la fréquence varie au cours du temps.

  Pour ce faire nous disposons de 2 librairies de données de séquençage (illumina « Paired End ») produites à 4 ans d'intervalle à partir du même isolat (soit environ 16 à 24 générations d'intervalle, le temps de génération étant de 2-3 mois en environnement contrôlé). Le premier séquençage, réalisé en 2015 et auquel il sera fait référence sous l'appellation « T0 », correspond au matériel ayant permis de produire l'actuel génome de référence de *M. incognita* (Blanc-Mathieu et al., 2017). Le deuxième séquençage, auquel il sera fait référence sous l'appellation T1, a été réalisé en 2019 (données non encore publiées; correspondent à la librairie "incognitaV4" présentée dans le chapitre V).

- Évaluer expérimentalement les fréquences de présence d'ETs candidats à partir de matériel génétique contemporain des différents séquençages (et issu du même isolat).

  Cette étape constitue le point critique de cette analyse : l'estimation expérimentale des fréquences d'ETs au sein de l'isolat à un point de cinétique donné ne peut être réalisée qu'à partir d'un mélange d'ADN issu d'un pool d'individus. En effet, compte tenu de la taille de l'organisme et des techniques actuellement mises en place, il n'est en effet pas envisageable d'extraire suffisamment de matériel génétique par individu pour pouvoir réaliser des expériences individuelles de PCR (validant la présence d'un ET à une position donnée) pour

ensuite évaluer la part de la population pour qui l'ET est présent à cette position. Utiliser une méthode quantitative ou semi-quantitative plutôt que qualitative est donc nécessaire. Dans un premier temps j'avais envisagé une approche par PCR quantitative (qPCR) permettant d'évaluer la quantité d'ADN pour laquelle un ET donné est présent ou absent dans l'échantillon (via l'utilisation d'une séquence en copie unique comme base de normalisation). Nous avons aussi mis en place une expérience de dosage (Qubit) reposant sur l'utilisation des mêmes couples d'amorces à partir d'un produit de PCR en conditions non saturantes.

- Pour les loci dont la fréquence aurait été évaluée expérimentalement, comparer ces valeurs à celles prédites *in silico* (évaluation de la précision des prédictions).

- Réaliser l'évaluation expérimentale des fréquences de ces mêmes loci pour d'autres points de cinétiques (antérieurs) afin d'analyser l'évolution de la présence de ces ETs au sein de l'isolat sur un intervalle de temps plus important.

  Des nématodes issus du même isolat ont en effet été conservés (congelés) à intervalles plus ou moins réguliers depuis 1997. J'avais prévu d'utiliser ce matériel afin de réaliser un point de cinétique tous les 4 ans environ depuis 1997 jusqu'en 2012 en plus des deux points de cinétique précédemment évoqués, soit 7 points de cinétique au total. Une telle analyse permettrait de suivre l'évolution de la fréquence de différents ETs sur 88 à 132 générations au sein de cet isolat de *M. incognita.*

3 - Matériel

Génome et fichiers d'annotation

Pour réaliser cette analyse nous avons utilisé l'assemblage du génome de référence de *M. incognita* (Blanc-Mathieu et al., 2017) (ENA assembly accession GCA_900182535, bioproject 807 PRJEB8714). Le fichier d'annotation des ETs utilisé est issu de (Kozlowski et al., 2020). Le fichier d'annotation des gènes utilisé est tiré de (Blanc-Mathieu et al., 2017).

Données de séquençage

Nous avons utilisé 2 librairies de données de séquençage, appelées T0 et T1, produites à 4 ans d'intervalle à partir du même isolat. La librairie T0 (illumina PE 2*100 ; access. Nb. : ERS1696677), produite en 2015, correspond au matériel ayant permis d'assembler l'actuel génome de référence de *M. incognita* (Blanc-Mathieu et al., 2017). Cette librairie comptabilise 78 374 471 reads. La librairie T1 (illumina PE PCR-free 2*250 ; données non publiées, correspondent à la librairie "incognitaV4" présentée dans le chapitre V) , a été produite en 2019 et comptabilise 39 895 726 reads. Les deux séquençages ont chacun été réalisés à partir d'un pool d'environ 1 million d'individus.

Matériel biologique

Les estimations expérimentales des fréquences d'ET au sein de l'isolat ont été menées en triplicat biologique à partir de matériel biologique (nématodes congelés) contemporain aux deux points de cinétiques étudiés *in silico*. Pour un point de cinétique donné, chaque réplicat biologique correspond au prélèvement de nématodes (protocole décrit dans (Rosso et al., 1999)) à partir d'une lignée différente de l'isolat Morelos puis à leur stockage à -80 °C ; chaque lignée de nématode étant maintenue en parallèle en conditions contrôlées sur tomate (*Solanum esculentum*).

4 - Méthodes

Identification de sites d'ET polymorphes en fréquence au cours du temps.

Nous avons utilisés le pipeline d'analyse popoolationTE2 v-1.10.04 (Kofler et al., 2016) afin de calculer la fréquence de présence d'ET et d'identifier de potentiels sites d'ETs polymorphes en fréquence au cours du temps à partir du génome de référence de *M. incognita*, d'un set d'annotation en ET, ainsi que de données de séquençage à différents points de cinétique.

La méthodologie de prédiction de site d'ET variant en fréquence (popoolationTE2) suivie dans cette analyse est strictement similaire à celle décrite en détails dans (Kozlowski et al., 2020) à la différence près que nous avons ici comparé deux librairies de reads issues d'un même isolat de *M. incognita* plutôt que de comparer des données provenant d'isolats différents.

Nous avons ensuite analysé les résultats produits par popoolationTE2 pour ne retenir que les loci détectés présentant des signatures « FR », i.e les loci pour lesquels la présence de l'ET est confirmé

par une accumulation d'information à gauche « F » et à droite « R » du point d'insertion. Puis nous avons calculé la variation de fréquence au cours du temps en réalisant la différence entre les valeurs de fréquences T1 et T0. Seuls les sites présentant des variations de fréquence > 1% ont été retenus (en valeur absolue ; cette valeur correspond à l'arrondi du taux d'erreur systématique de 0.97% que j'ai estimé pour popoolationTE2, voir chapitre VI). Nous avons comparé les positions des sites polymorphes prédits restants avec celles de l'annotation en ETs de référence en utilisant l'outil « intersect » (-wao) de la suite d'outils bedtools v-2.27.1 (Quinlan and Hall, 2010) afin de caractériser le type du polymorphisme, i.e déterminer si le polymorphisme touche un ET annoté dans le génome de référence ("ET de référence polymorphe") à ce locus ou non ("neo-insertion"), ou encore une "extra detection" (ET détecté dans la population de référence mais non décrit dans l'annotation de référence). Nous avons identifié les gènes potentiellement affectés par les ETs polymorphes en croisant le fichier d'annotation de ces loci avec le fichier d'annotation des gènes via l'utilisation de la suite de logiciels bedtools. Nous avons utilisé l'outil "closest" (-D b -fu -io ; b étant le fichier d'annotation des gènes) pour identifier le gène le plus proche (mais non chevauchant) en aval de chaque site polymorphe. Nous n'avons retenu que les entrées ayant une distance maximale de 10 kb entre le site polymorphe et les extrémités du gène. Nous avons identifié les loci polymorphes insérés dans les gènes en utilisant l'outil "intersect" (-wo). L'ensemble des analyses statistiques réalisées sur les résultats produits par les différents outils ont été faites sous R v-3.6.3.

Enfin, nous avons sélectionné 3 loci polymorphes pour validation expérimentale des fréquences de présence d'ETs à partir de matériel génétique contemporain aux deux points de cinétiques étudiés *in silico*. Ces loci ont été sélectionnés selon les critères suivants. Ils devaient premièrement présenter une variation de fréquence > 10% (valeur absolue) entre les deux points de cinétique pour que les fréquences estimées expérimentalement pour les deux points de cinétique aient plus de chances d'être distinguables. Deuxièmement, différentes amplitudes de variations de fréquences devaient être représentées afin de juger de la sensibilité des méthodes de prédiction et de validation. Enfin, dans les loci choisis, il fallait qu'à la fois les augmentations et les diminutions de fréquence au cours du temps soient représentées.

Estimations expérimentales de fréquences d'ETs à partir de matériel génétique issu d'un pool d'individus.

Afin de réaliser une estimation expérimentale de fréquence de présence/absence d'ETs à partir de matériel génétique issu d'un pool d'individus, nous avons envisagé deux approches quantitatives : la qPCR et le dosage Qubit.

Extraction d’ADN

L'ADN a été extrait et purifié selon le protocole du kit de purification de l'ADN MasterPure™ (Epicentre Biotechnologies). L'ADN a été extrait à partir de nématodes au stade juvénile congelés provenant de différents points de cinétique, avec trois réplicats biologiques par point de cinétique. Afin de s’assurer de la présence de quantité suffisante de matrice dans les extractions, nous avons quantifié l’ADN présent dans chaque échantillon avec le spectrophotomètre Nanodrop 2000 (ThermoFisher Scientific) et le fluoromètre Qubit3 (Thermofisher Scientific).

Design et validation de sondes validant la présence ou l’absence d’ET aux loci étudiés.

Nous avons conçu 6 couples d’amorces (2 pour chaque locus : 1 validant la présence d’un ET donné à cette position, l’autre rapportant l’absence de ce même ET) en utilisant l'interface web Primer3Plus (Untergasser et al. 2007). Pour les couples d’amorces validant la présence d’un ET, une des amorces se situe sur la séquence génomique flanquante du point d’insertion et l’autre se situe dans la séquence de l’ET. Pour les couples d’amorces validant l’absence d’insertion, une des amorces chevauche le point d'insertion rendant impossible l’amplification en cas d’insertion. L'ensemble des amorces sont présentées (séquence et taille attendue de l’amplicon) dans le Tableau annexe 7.7.1.

La validation des sondes a été réalisée par PCR sur un échantillon d’ADN issu d’un prélèvement récent sur le même isolat que celui analysé. Le mélange PCR contenait 0,5µmol de chaque amorce, 1x de tampon de réaction MyTaq™ et 1,0 U de l'ADN polymérase MyTaq™ (Bioline Meridian Bioscience) ajusté à un volume total de 20µL. L'amplification PCR a été réalisée avec un TurboCycler2 (Blue-Ray Biotech Corp.). Les conditions de la PCR étaient les suivantes : dénaturation initiale à 95°C pendant 5 min, suivie de 35 cycles de 95°C pendant 30 s, 56°C pendant 30 s d'annelage, et 72°C pendant 15 s d'extension, le programme se terminant par une extension finale à 72°C pendant 10 min. Des aliquotes de 5µL ont été mis à migrer par électrophorèse sur un gel d'agarose à 1% (Sigma Chemical Co.) pendant 70 min à 100 V. Le marqueur de taille utilisé est le « 1kb Plus DNA Ladder » (New England Biolabs Inc.), contenant les fragments de taille suivants en pb : 100, 200, 300, 400, 500, 600, 700, 900, 1000, 1200, 1500, 2000, 3000, 4000, 5000, 6000, 8000 et 10000.

La révélation des bandes a été réalisée en utilisant du bromure d'éthidium et une exposition aux rayons ultraviolets. Les bandes de produits PCR ont été excisées du gel d'agarose avec un scalpel et purifiées à l'aide du kit d'extraction de gel MinElute (Qiagen) en suivant le protocole du fabricant. Les produits PCR ont ensuite été séquencés par méthode Sanger (Eurofins Genomics).

Amplification par qPCR et calcul d'efficacité de primer

Les expériences de PCR quantitatives ont été effectuées à l'aide du kit qPCR MasterMix Plus pour SYBR® Green I (Eurogentec). L'amplification qPCR a été réalisée avec le système de PCR en temps réel AriaMx (Agilent). Les conditions de qPCR étaient les suivantes : dénaturation initiale à 95°C pendant 5 min, suivie de 40 cycles de 95°C pendant 15 s, 56°C pendant 30 s de recuit, et 72°C pendant 30 s d'extension, et détection de la courbe de fusion à 95°C pendant 30 s, 65°C pendant 30 s et 95°C pendant 30 s.

L'efficacité des amorces a été calculée et est indiquée dans le Tableau annexe 7.7.2.

Dosage d'ADN à partir de produit de PCR en conditions non saturantes.

Nous avons réalisé une analyse de dosage d'ADN visant à estimer la proportion d'ADN ayant/n'ayant pas l'ET à partir d'échantillons de matériel génétique issu de pools d'individus.
Le dosage a été effectué avec le kit de dosage à large gamme d'ADN Qubit (Thermofisher scientific). La quantification a été faite sur des produits d'amplification PCR en conditions non saturantes (20 cycles), et la quantité a été normalisée en utilisant des amplicons de gène d'actine (Minc3s00960g19311).

Les fréquences de présence de l'ET dans chaque échantillon a été évaluée en calculant le rapport suivant : $Qnorm(TE^+)/(Qnorm(TE^+) + Qnorm(TE^-))$; avec Qnorm étant la quantité d'ADN présentant l'ET (TE^+) ou non (TE^-) normalisée par la quantité d'ADN amplifiée pour l'actine.

5 - Résultats

Des ETs varient en fréquence en un faible nombre de génération au sein d'un isolat de *M. incognita*.

Nous avons cherché, via une approche *in silico*, à identifier des copies d'ETs dont la fréquence (de présence) aurait varié au cours du temps au sein d'un isolat de *M. incognita*. Pour ce faire, nous avons réalisé une analyse de génomique des populations pour comparer des données de séquençage produites à 4 ans d'intervalle à partir du même isolat.

Nous avons identifié 231 loci pour lesquels une variation de fréquence > 1% (taux d'erreur interne de PopoolationTE2 ; voir chapitre VI) est prédite entre les séquençages T0 et T1 (voir Figure 7.5.1). Ces 231 loci polymorphes concernent 6 rétrotransposons (2 LINEs et 4 LTRs) et 225 transposons à ADN (156 MITEs et 69 TIRs). L'ensemble des polymorphismes prédits concernent des copies d'ETs décrites dans l'annotation du génome de référence de *M. incognita* (Kozlowski et al., 2020). Ainsi, en une vingtaine de générations, 2.39 % (231/9633) des sites d'ETs de référence ont varié en fréquence au sein de l'isolat. Nous n'avons détecté aucune néo-insertion ou délétion d'ET entre les deux temps de séquençages.

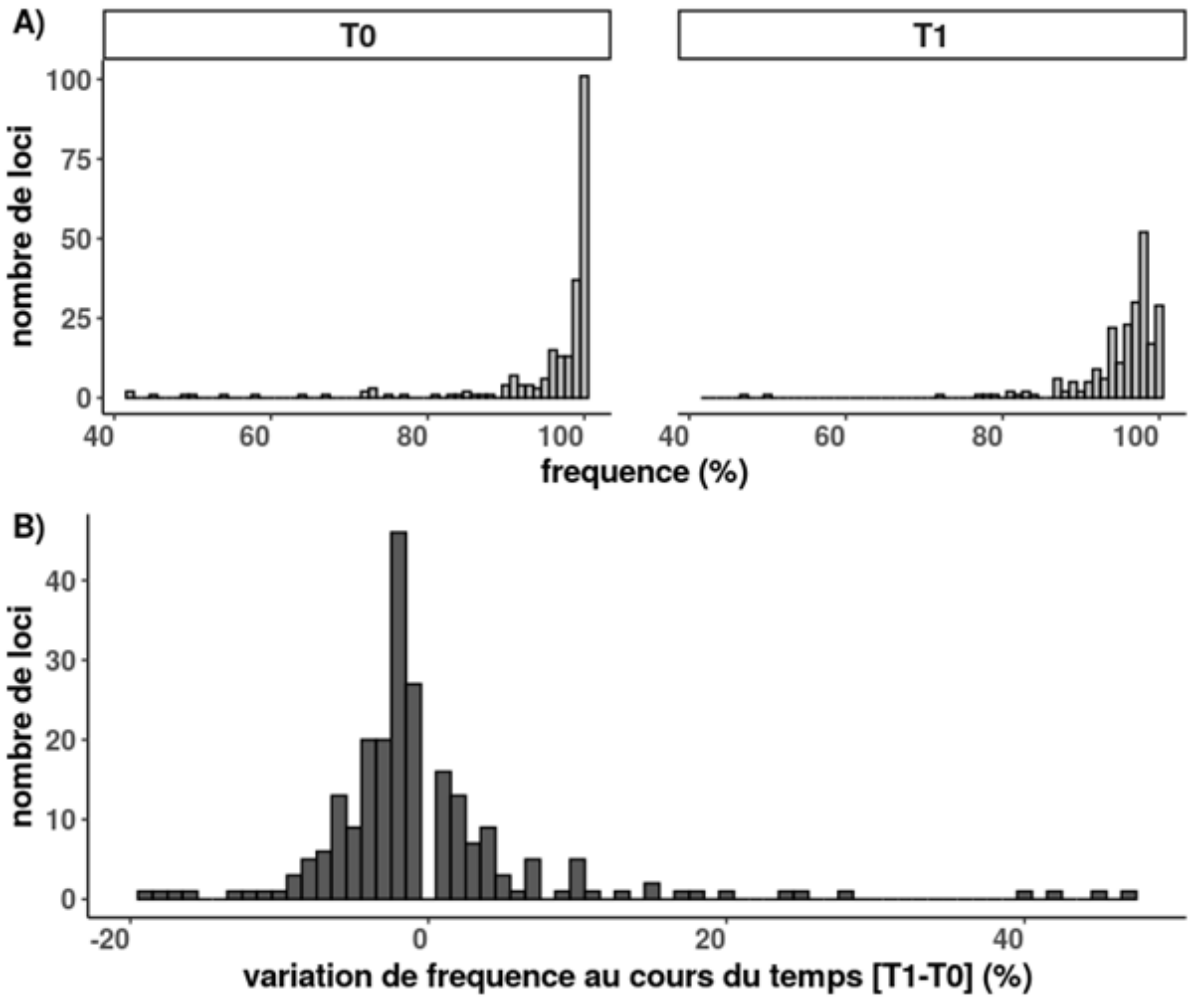


Figure 7.5.1 : fréquences de présence d'ETs et variations de fréquences au cours du temps.

A - *Histogrammes des fréquences prédites par librairies (T0 et T1 e.g. points de cinétique) pour l'ensemble des sites polymorphes détectés (231).*

Aucun seuil bas de fréquence n'a été introduit. Le fait qu'aucune fréquence ne soit inférieure à 40 % est dû au fait que les données utilisées sont issues du même isolat à deux points de cinétique proche. Le matériel T0 correspondant au matériel ayant servi à réaliser le génome de référence (Blanc-Mathieu et al. 2017)*, seuls les ET en fréquence majoritaires ou proches de l'être ont été assemblés, puis potentiellement détectés lors de la prédiction et de l'annotation du génome.*

B - Histogramme des valeurs de variation de fréquence au cours du temps (e.g. fréquences(T1)- fréquences(T0)). Les valeurs positives de variations de fréquences indiquent une augmentation de la fréquence de présence de l'ET au sein de l'isolat au cours du temps. Les valeurs négatives indiquent une diminution de la fréquence de présence de l'ET au cours du temps. Les loci présentant des variations de fréquences au cours du temps inférieures à 1% n'étant pas considérés comme polymorphes, ils ne sont pas représentés sur cette figure.

Pour l'ensemble des histogrammes, l'épaisseur d'une barre représente un intervalle de fréquence de 1%. Les valeurs de fréquences ou de variations de fréquences (axe des abscisses) sont exprimées en pourcentage.

Un faible nombre de loci présentent des fréquences d'ETs inférieures à 75% (15 loci à T0 et 3 à T1). Une extrême majorité des loci présentent donc des fréquences proches de la fixation (Figure 7.5.1-A). Les 3/4 des loci polymorphes détectés montrent des variations de fréquence inférieures à 6% entre les deux temps de séquençages (Figure 7.5.1-B). Néanmoins, certains loci présentent des variations de fréquence plus importantes pouvant aller d'une diminution de 19% à une augmentation de 47% au cours du temps pour les cas les plus extrêmes. Ainsi, à de rares exceptions près, l'amplitude des variations de fréquence entre les deux séquençages est modérée.

Seul un tiers des loci (77/231) présentent une augmentation de fréquence d'ETs au cours du temps (voir Figure 7.5.1-B). L'évolution de la fréquence des ETs analysés n'est donc pas à l'équilibre dans cet isolat. De plus, la comparaison des valeurs d'augmentation et de diminution de fréquence au cours du temps a montré une différence significative entre les deux distributions (test de conformité de Kolmogorov-Smirnov, p-valeur = 0.01584). La répartition des variations de fréquences est donc différente selon le type de fluctuation. Ainsi on peut en conclure que si la majorité des ETs détectés diminuent en fréquence au cours du temps, l'amplitude des variations de fréquence est plus importante en ce qui concerne les augmentations.

Les variations de fréquence observées ne semblent pas contraintes par le type d'ET impliqué ni par l'environnement génique.

Les polymorphismes impliquant des ETs de l'ordre des MITEs sont ceux qui présentent les variations de fréquence les plus importantes. Cependant, compte tenu du faible nombre de LTR et de LINE représentés face aux TIRs et aux MITEs, il est impossible de conclure statistiquement quant à l'homogénéité de répartition des valeurs de variation de fréquence en fonction de l'ordre d'ET. Par ailleurs, les valeurs médianes des variations de fréquence par ordre sont relativement similaires. Globalement, la distribution des variations de fréquences ne diffère pas en fonction de l'ordre d'ET analysé (voir Figure 7.5.2). L'évolution de la présence d'un ET dans cet isolat semble donc indépendante des caractéristiques de l'ET, du moins pour ce niveau de détail de classification.

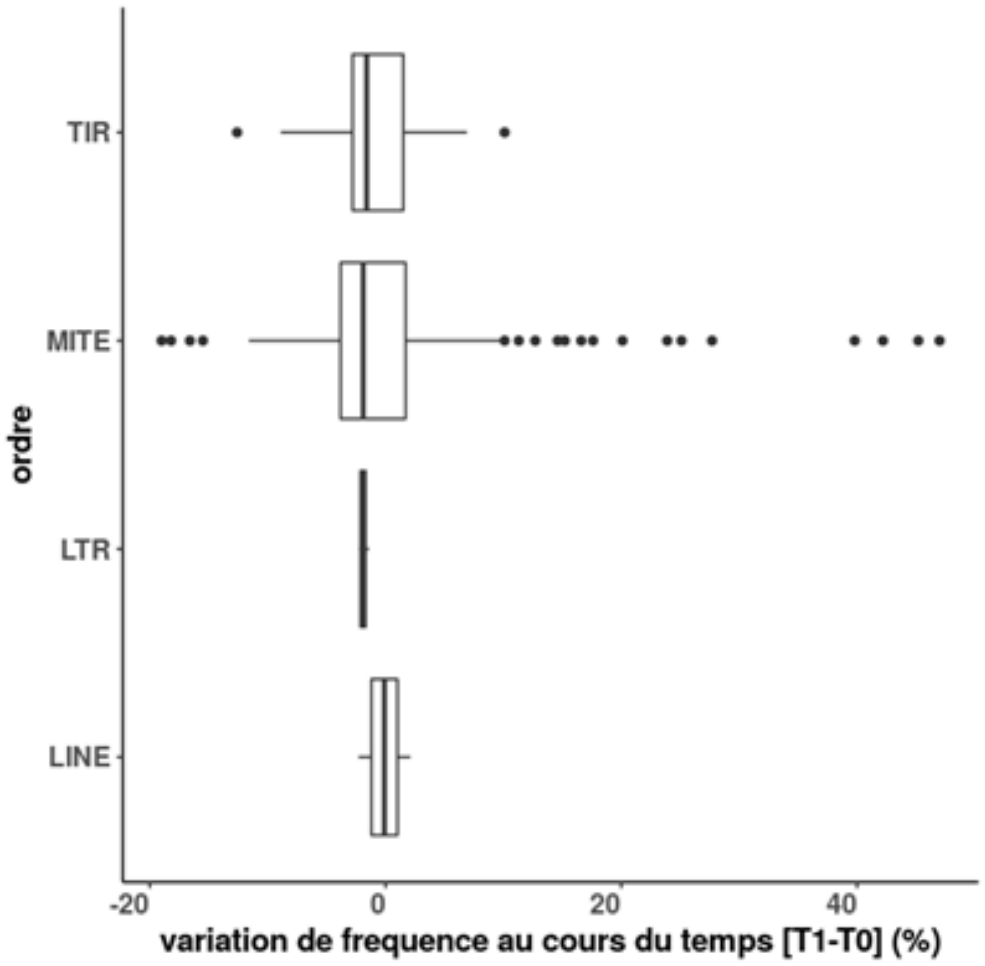


Figure 7.5.2 : distribution par ordre d'ET des variations de fréquence au cours du temps.

Chaque boîte à moustache représente la distribution des valeurs de variations de fréquence au cours du temps pour un ordre d'ET donné. Les traits épais à l'intérieur des boîtes à moustache représentant la valeur médiane.

Nous avons souhaité évaluer l'existence d'une relation entre le type (augmentation ou diminution) et l'amplitude de la variation de fréquence d'un ET à un loci au cours du temps et la présence de gènes à proximité. Pour ce faire, nous avons dans un premier temps isolé les gènes en aval des sites polymorphes (voir Figure 7.5.3). Au total, 174 loci se situent à moins de 10kb en amont de gènes et présentent une variation de fréquence au cours du temps (117 diminutions et 57 augmentations). Un test de corrélation de Pearson a montré que l'amplitude des variations de fréquences au cours du temps n'est pas liée à la proximité par rapport aux gènes, et ce en considérant les augmentations et les diminutions de fréquences conjointement ou indépendamment.

Nous avons dans un second temps isolé les ETs polymorphes insérés dans des gènes. Au total, 62 loci sont concernés (45 diminutions et 17 augmentations de fréquence au cours du temps). Un test exact de Fisher a montré que les proportions d'augmentation ou de diminution de fréquence au cours du temps ne varient pas en fonction du fait d'être proche d'un gène ou inséré dans le corps du gène.

La nature et l'amplitude des variations de fréquences d'ETs au cours du temps semblent donc indépendantes de la proximité d'un ET avec un gène.

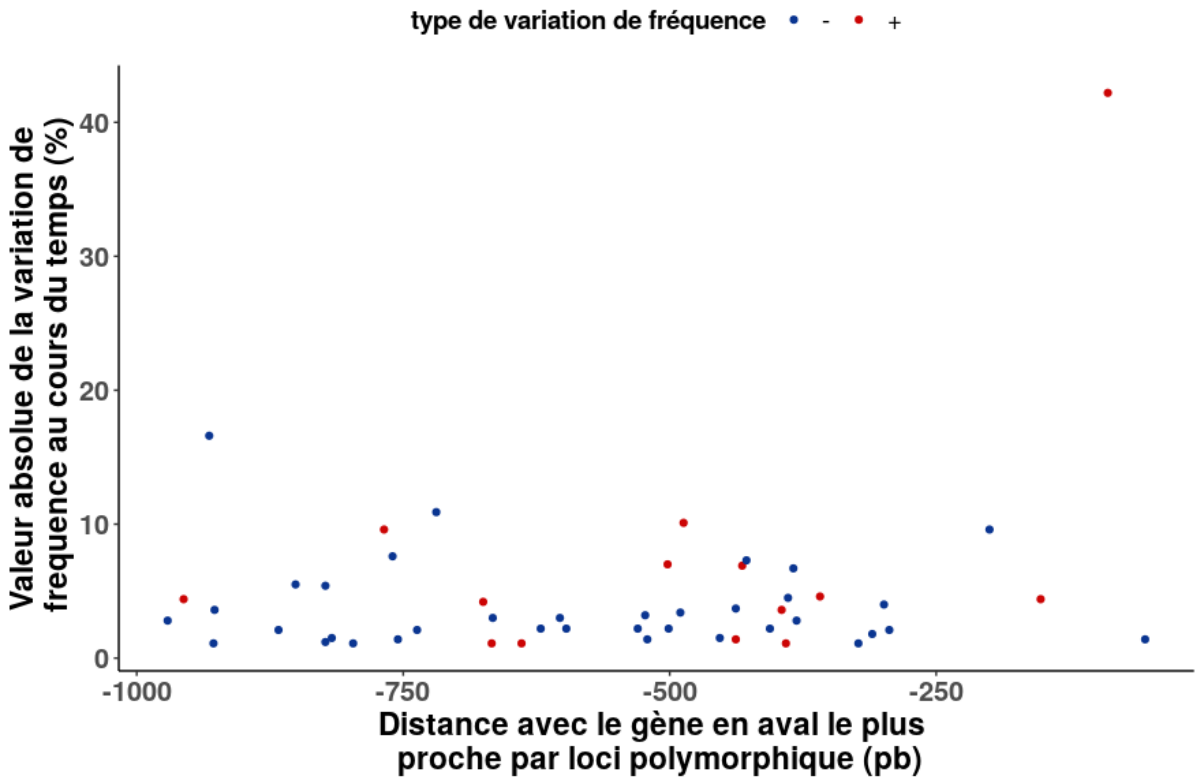


Figure 7.5.3 : Amplitude de la variation de fréquence au cours du temps en fonction de la distance avec le gène en aval le plus proche.

Chaque point représente un site polymorphe. La couleur indique le type de variation de fréquence observé pour ce locus au cours du temps : bleu pour une diminution, rouge pour une augmentation. Les loci polymorphes insérés dans des gènes ne sont pas représentés.

L'expérimentation biologique valide des polymorphismes de présence d'ET au sein de l'isolat Morelos de *M. incognita* mais ne permet pas de conclure quant aux variations de fréquence au cours du temps.

Nous avons sélectionné 3 loci d'ET pour validations expérimentales que nous nommerons à partir de maintenant 01542, 00187 et 00004 (voir méthodes). L'ensemble des ETs sélectionnés appartiennent à l'ordre des MITEs, des transposons à ADN non-autonomes majoritaires dans le génome de *M. incognita* (Kozlowski et al., 2020), mais sont issus de séquences consensus distinctes. Les fréquences de présence de chaque ET estimées par PopoolationTE2 aux deux points de cinétique sont détaillées dans le Tableau 7.5.1. Les loci 01542 et 00004 présentent une augmentation en fréquence de l'ET au cours du temps (respectivement + 25.1 % et +14.6 %) . Le locus 00187 quant à lui, présente une diminution de la fréquence de l'ET au cours du temps (-19.0 %). Pour chaque locus, nous avons conçu deux couples d'amorces en vue de réaliser des PCR: l'un validant l'absence de l'ET à ce locus et l'autre validant la présence de l'ET.

Tableau 7.5.1 : fréquences des ETs polymorphes prédites aux loci sélectionnés pour validation expérimentale.

Les fréquences (%) représentent la proportion d'individus pour lesquels il est prédit que l'ET (colonne de gauche) est présent à un temps donné (T0 ou T1) à ce locus. La colonne T1-T0 représente la variation de fréquence de présence de l'ET dans l'échantillon au cours du temps (%). Le signe de la différence représente le type de variation de fréquence au cours du temps ('+' : augmentation ; '-' : diminution).

	Fréquence de présence de l'ET (%)		
	T0 (2015)	**T1 (2019)**	**T1-T0**
01542	72.8	97.9	+ 25.1
00187	97.7	78.7	- 19.0
00004	85.4	100.0	+ 14.6

Nous avons réalisé une première expérience de PCR afin de tester la spécificité de ces couples d'amorces (voir Figure 7.5.4-A). La migration des produits de PCR sur gel indique que les séquences amplifiées sont bien uniques et donc que les couples d'amorces sont spécifiques. Nous avons ensuite séquencé ces amplicons. Pour chaque locus, la taille ainsi que les séquences des amplicons confirment les résultats attendus i.e la présence ou l'absence de l'ET suivant le couple d'amorces considéré et ce pour les 3 loci analysés. Le fait que les deux formes, avec et sans ETs, coexistent au sein d'un même échantillon pour chacun des loci analysés confirme le fait qu'il existe un polymorphisme de présence de ces ETs entre individus au sein de l'isolat morelos.

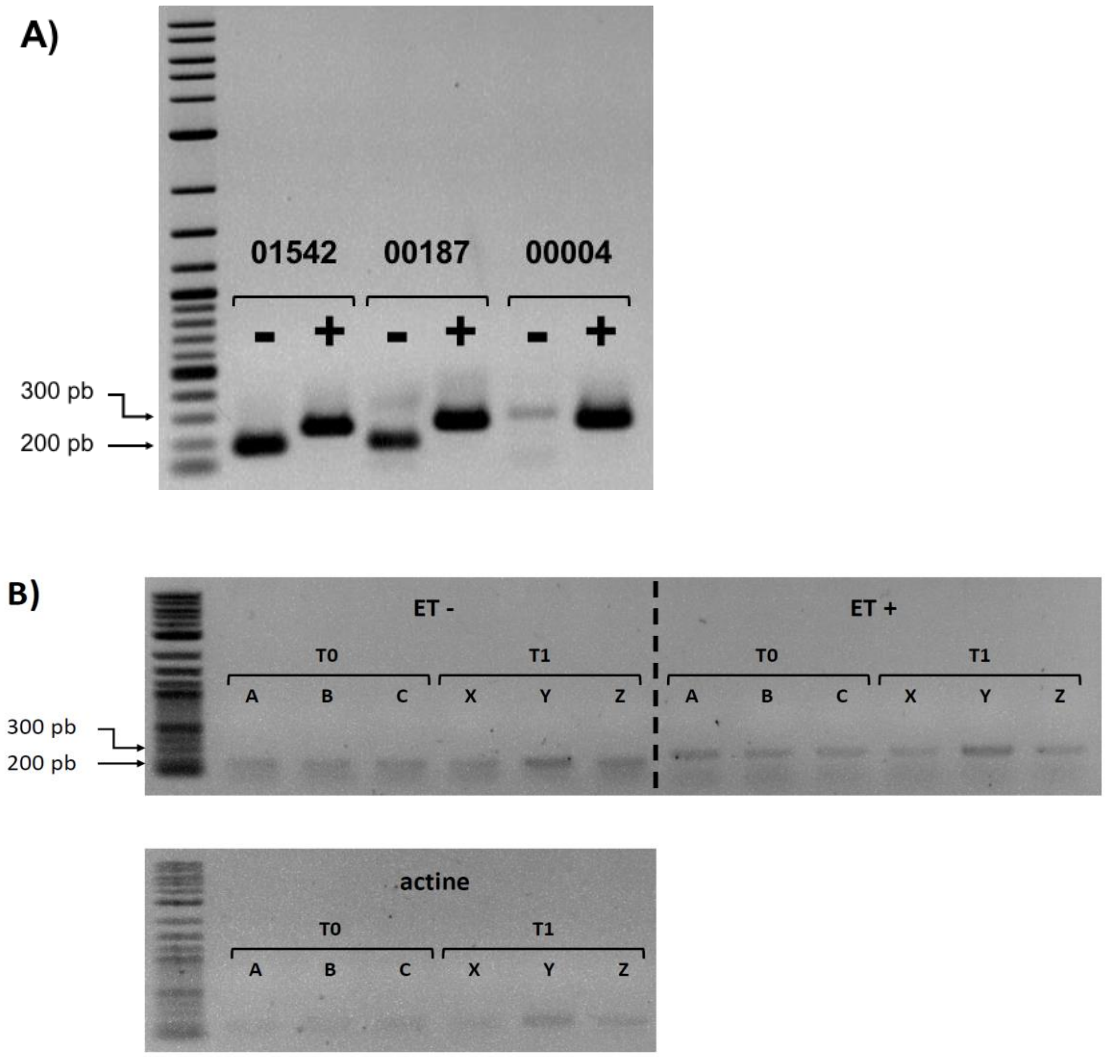


Figure 7.5.4 : validation expérimentale des polymorphismes de présence d'ETs.

A- ***Validation des couples de primers pour les 3 loci polymorphes sélectionnés.***

Chaque accolade correspond à un locus d'ET polymorphe et englobe deux couples de primers : un validant l'absence de l'ET dans l'échantillon (ET-) et l'autre validant la présence de l'ET à cette position (ET+) . Pour les 6 couples de primers, les tailles des amplicons correspondent aux tailles attendues (voir Tableau annexe 7.7.1). NB: pour les couples d'amorces validant la présence d'un ET, une des amorces se situe sur la séquence génomique flanquante du point d'insertion et l'autre se situe dans la séquence de l'ET. Ainsi, seule une partie de l'ET est amplifiée (voir Méthodes).

B- ***Validation par PCR du polymorphisme de présence / absence au locus 01542 pour deux points de cinétique.***

ET - , ET+ et Actine correspondent respectivement à la validation des couples de primers soutenant l'absence de l'ET au locus 01542 (TE-), la présence de l'ET à ce même locus (TE+) et la présence de l'actine (copie unique dans le génome, utilisé comme rapporteur de charge de matrice). Chaque couple de primer est testé en triplicat et ce pour deux points de cinétique (T0 : A, B, C et ; T1 : X, Y; Z) symbolisés par une accolade. Pour chaque couple de primer (ET-, ET+ et Actine), la taille des amplicons est cohérente avec les tailles attendues (voir Tableau annexe 7.7.1) et ce pour l'ensemble des réplicats et pour les deux points de cinétique.

Nous avons ensuite réalisé une deuxième expérience de PCR menée cette fois sur des échantillons issus de deux points de cinétique et en triplicat biologique. Nous avons réalisé cette expérience en conditions non saturantes (20 cycles) pour pouvoir ensuite procéder à un dosage Qubit des produits d'amplification et ainsi juger des quantités relatives de la présence / absence d'ET dans les échantillons. La migration sur gel effectué pour 01542 (voir Figure 7.5.4-B) ainsi que les dosages Qubit réalisés pour l'ensemble des loci (voir Tableau 7.5.2 pour les valeurs moyennes, voir le Tableau annexe 7.7.3 pour les données brutes) confirment le fait que les formes avec et sans ETs coexistent au sein de l'isolat morelos et ce pour l'ensemble des réplicats aux deux points de cinétiques.

Tableau 7.5.2 : valeurs moyennes de dosage Qubit en triplicat de la quantité d'ADN avec et sans ETs pour les 3 loci sur 2 points de cinétique.

Les valeurs présentées correspondent aux valeurs moyennes de dosage d'ADN à partir de produit d'analyses PCR effectuées en triplicats biologiques (A, B et C pour T0 ; X, Y et Z pour T1) sur les deux points de cinétique T0 (2015) et T1(2019).

Pour les loci 00187 et 00004, le contrôle "actine" est le même (pour chaque échantillon respectivement), l'expérience de PCR ayant été réalisée simultanément pour ces deux loci.

Q(x) -- (moyenne par triplicat biologique): quantités d'ADN (ng/microL) présente pour un couple d'amorce à l'issue de l'amplification PCR (20 cycles).

iniQ(x) -- (moyenne par triplicat biologique): quantités d'ADN (ng/microL) initialement présente pour le couple d'amorce x ; calculé selon la formule suivante : iniQ(x) = Q(x)/(facteur d'amplification (x) ^ nb. cycle) avec x un couple d'amorce, Q(x) la quantité d'ADN quantifié, et nb. cycle le nombre de cycle réalisé lors de la PCR (20 cycle pour l'ensemble des expériences).

normQ(x) -- (moyenne par triplicat biologique): quantité d'ADN initiale (ng/microL) normalisée par la quantité d'actine initiale moyenne calculé selon la formule suivante : norm(x) = iniQ(x)/iniQ(Actine) (pas d'unité)

normQ(TE+) / (normQ(TE-) + normQ(TE+)) -- (moyenne par triplicat biologique) proportion (d'ADN) pour lequel l'ET est présent à un point de cinétique donné .NB: Il est ici considéré que la somme des quantités d'ADN pour lequel l'ET est présent (TE+) ou absent (TE-) est égale à la quantité totale d'ADN

*fréquence expérimentale de présence de l'ET (%) -- (moyenne par triplicat biologique): évaluée selon la formule suivante : (normQ(TE+) / (normQ(TE-) + normQ(TE+)))*100 ; estimation expérimentale de la fréquence moyenne de présence de l'ET dans l'isolat pour un point de cinétique donné.*

		T0 (2015)		T1 (2019)		T1 - T0 (taux d'accroissement moyen (%))
		moyenne	erreur type de la moyenne	moyenne	erreur type de la moyenne	
01542	initQ(TE-)	1,101E-06	1,648E-08	1,242E-06	1,615E-07	-
	initQ(TE+)	1,329E-07	7,025E-09	1,764E-07	3,451E-08	
	initQ(Actine)	2,767E-06	2,803E-08	2,503E-06	4,728E-08	
	normQ(TE-)	3,980E-01	7,612E-03	4,982E-01	7,261E-02	
	normQ(TE+)	4,806E-02	2,936E-03	7,102E-02	1,516E-02	
	normQ(TE+) / (normQ(TE-) + normQ(TE+))	1,076E-01	4,078E-03	1,232E-01	1,280E-02	
	fréquence expérimentale de présence de l'ET (%)	**10,76**		**12,32**		**+ 1,56 (14,58)**
00187	initQ(TE-)	9,037E-06	3,479E-07	1,112E-05	1,065E-06	-
	initQ(TE+)	1,126E-06	2,298E-08	1,516E-06	4,056E-07	
	initQ(Actine)	2,853E-06	1,578E-07	4,221E-06	9,829E-07	
	normQ(TE-)	8,022E+00	1,697E-01	8,059E+00	1,385E+00	
	normQ(TE+)	3,962E-01	1,502E-02	3,593E-01	3,120E-02	
	normQ(TE+) / (normQ(TE-) + normQ(TE+))	4,716E-02	2,661E-03	4,543E-02	9,887E-03	
	fréquence expérimentale de présence de l'ET (%)	**4,72**		**4,54**		**- 0.18 (3,67)**
00004	initQ(TE-)	6,108E-07	2,075E-08	7,375E-07	7,975E-08	-
	initQ(TE+)	1,674E-06	6,124E-08	2,092E-06	4,155E-07	
	initQ(Actine)	2,853E-06	1,578E-07	4,221E-06	9,829E-07	
	normQ(TE-)	2,147E-01	5,910E-03	1,889E-01	3,125E-02	
	normQ(TE+)	5,895E-01	2,845E-02	5,161E-01	6,128E-02	
	normQ(TE+) / (normQ(TE-) + normQ(TE+))	7,327E-01	4,528E-03	7,334E-01	1,782E-02	
	fréquence expérimentale de présence de l'ET (%)	**73,27**		**73,34**		**+ 0.07 (0,10)**

Après avoir validé qu'il existe bel et bien une hétérogénéité de présence d'ET à plusieurs loci, nous avons souhaité quantifier expérimentalement les fréquences de ces ETs dans l'isolat aux deux points de cinétique pour ensuite pouvoir comparer l'évolution de ces fréquences au cours du temps par rapport aux prédictions bio-informatiques. Le défi consistait à réaliser l'estimation expérimentale des fréquences à partir d'un mélange d'ADN issu d'un pool d'individus. Nous avons tenté deux approches de quantification d'ADN basées sur l'utilisation des couples d'amorces précédemment testés caractérisant la présence ou l'absence d'un ET aux différents loci.

Nous avons réalisé une première expérience de dosage QuBit des formes ET+ et ET- selon la philosophie suivante : Nous avons pour chaque locus réalisé une expérience de PCR en conditions non saturantes (20 cycles ; aucun plateau d'amplification atteint). Nous avons ensuite utilisé l'efficacité des amorces pour calculer la quantité d'ADN initialement présente dans l'échantillon pour une forme données (ET+, ET- ou Actine) à partir de la quantité estimée dans le produit de PCR. La quantité d'ADN présente dans l'échantillon pouvant varier d'un échantillon à un autre, nous avons utilisé la quantité d'ADN amplifiée pour l'actine comme base de normalisation, cette région étant en copie unique dans le génome, et théoriquement présente chez la totalité des individus de l'échantillon. Ensuite, étant donné i) que nous avions précédemment démontré que chacune des région amplifiée était spécifique et unique, et ii) qu'un ET est soit présent soit absent (état binaire) à une position donnée du génome de chaque individu; nous avons fait l'hypothèse que la somme de la quantité d'ADN des formes ET+ et ET- dans un échantillon devait être égale à la quantité total d'ADN amplifié de manière unique pour chaque individus de l'échantillon (e.g. la quantité d'ADN amplifiée pour l'actine). Ainsi selon notre hypothèse, (Q(ET+) + Q(ET-))/Q(actine) = 1. De plus, l'amplification de chaque forme étant unique et spécifique au sein de chaque génome, la quantité d'ADN ET+ et ET- est directement proportionnelle à la proportion de copies d'ADN ayant l'ET présent ou non à cette position et donc au nombre d'individus dans l'échantillon pour qui l'ET est présent. Pour chaque ET, nous avons donc calculé la fréquence de présence de l'ET (%) dans l'échantillon comme suit : $f = (Q(ET+)/(Q(ET-) + Q(ET+))*100$.

Comme nous pouvons l'observer dans le Tableau 7.5.2, les fréquences calculées expérimentalement selon ces méthodologies sont très éloignées des fréquences estimées in silico (voir Tableau 7.5.1) ; en particulier en ce qui concerne les loci 01542 et 00187. Ceci s'explique d'ores et déjà par le fait que pour chaque échantillon $Q(ET+) + Q(ET-) << Q(actine)$ alors que ces deux quantités devraient en théorie être équivalentes. Le ratio $(Q(ET+)/(Q(ET-) + Q(ET+))*100$ n'est donc pas directement représentatif de la fréquence de présence des ETs analysés dans les échantillons.

Par ailleurs, nous pouvons aussi observer qu'il existe une variabilité importante des fréquences estimées entre les réplicats d'un même point de cinétique. Ainsi, si ces résultats confirment bien qu'il

existe une hétérogénéité de présence d'ET pour chaque échantillon aux loci étudiés, il est clair que cette expérience ne permet pas d'estimer avec suffisamment de précision les fréquences de présence de ces ETs dans l'isolat a un point de cinétique donné. Il est donc impossible de conclure quant à l'évolution de ces fréquences au cours du temps. On notera néanmoins que pour les 3 loci, le taux d'accroissement calculé à partir de ces fréquences estimées expérimentalement pour les deux points de cinétique suit la même tendance (signe de la différence) que l'évolution de fréquence prédite in silico.

Sans dresser une liste exhaustive, plusieurs facteurs, parfois cumulatifs, pourraient expliquer pourquoi ce type d'approche n'a pas permis d'évaluer les variations de fréquences au cours du temps. Un premier facteur serait un manque de précision général du dosage QuBit par rapport à celui requis pour ce type d'expérience. Un deuxième facteur pourrait être un manque de précision dans l'estimation de l'efficacité des couples d'amorces. En effet, comme 20 cycles d'amplification ont été réalisés, un manque de précision dans l'estimation de l'efficacité des couples d'amorces pourrait avoir des répercussions importantes sur le calcul de la quantité initiale d'ADN. En effet notre méthodologie, pour tenir compte des efficacités variables des amorces, prend en compte leur valeurs (voir légende Tableau 7.5.2); mais cela nous met à la merci d'imprécisions sur cette mesure. Enfin, il est aussi envisageable qu'une ou plusieurs des régions amplifiées ne soient pas uniques dans le génome. En effet, *M. incognita* est un organisme triploïde, ce qui augmente la probabilité que plusieurs régions soient strictement identiques dans le génome. Ainsi, si les loci détectés se trouvent dans ces régions (ET et région flanquante), la quantité d'ADN estimée serait donc un multiple du nombre de copies. Si tel était le cas, il serait impossible de différencier ces copies par migration sur gel ou séquençage (Sanger comme c'est le cas ici). Notons que dans l'assemblage du génome ces loci ne sont pas présents en de multiples copies mais s'ils correspondent à de longs fragments identiques, ils ont pu être collapsés lors de l'assemblage.

En parallèle, nous avons aussi réalisé une expérience test de PCR quantitative portant uniquement sur le locus 01542, ce locus étant celui pour lequel la plus grande variation de fréquence au cours du temps avait été prédite (+25,1 ; voir Table 7.5.1). La philosophie de cette expérience était ici d'utiliser la différence du nombre de cycles d'amplification nécessaire à la détection des formes TE+ et TE- comme rapporteur de la fréquence de l'ET au sein d'un échantillon. Cette variation a été calculée via la méthode du k^(- delta delta Ct) avec k correspondant à l'efficacité moyenne des deux couples d'amorces (ET+ et ET-).

Le défi de cette expérience concerne la précision de l'estimation du nombre de cycles nécessaire à la détection (Cq). En effet, nous faisons ici l'hypothèse que la quantité d'ADN correspondant aux

formes ET+ et ET- est directement proportionnelle au nombre d'individus présentant ou non l'ET à cette position. Or dans le cas idéal où l'efficacité de l'amorce est optimale (i.e 2), chaque cycle double la quantité d'ADN présente dans l'échantillon. Ainsi, la précision nécessaire à la détection d'une variation de fréquence est proportionnelle à la quantité d'individus ayant cette forme dans l'échantillon et à l'étendue de la variation de fréquence. Par exemple, dans le cas où un ET serait présent à une fréquence de 10% dans l'échantillon à T0 puis présent à une fréquence de 30% à T1, on devrait observer une variation théorique de 1.58 cycles entre les deux points de cinétique (30/10=2^x). En revanche, pour passer de 72.8% à 97.9%, soit une différence de 25.1% comme c'est le cas pour 01542, il faudrait être précis à moins de 0.42 cycle prêt (97.9/72.8 = 2^x).

Les résultats de cette expérience sont détaillés dans le Tableau annexe 7.7.4. En résumé, sur l'ensemble des échantillons nous pouvons constater que l'écart type moyen du nombre de cycle entre les réplicats techniques d'un échantillon (e.g réplicats techniques ET+, ET- et actine d'un réplicat biologique) est de 0.48 cycle, ce qui est au-delà de la tolérance maximale de 0.42 cycle pour pouvoir caractériser cette variation de fréquences.

Ainsi, sans même avoir à comparer les réplicats biologiques entre eux, nous pouvons d'ores et déjà conclure que la précision obtenue lors de cette expérience n'est pas suffisante pour pouvoir quantifier une évolution de fréquence de présence d'ET de cet ordre de grandeur au cours du temps. Ce locus étant celui pour lequel la plus grande variation de fréquences au cours du temps avait été prédite, il a donc été décidé de ne pas réaliser cette expérience pour les deux autres loci.

6 - Discussion

Comme présenté dans le chapitre VI, une hétérogénéité intra-isolat de la présence d'ET avait été prédite pour l'ensemble des isolats (12 au total) de *M. incognita* étudiés dont Morelos, l'isolat étudié ici (Kozlowski et al., 2020). Dans l'analyse actuelle, nous avons pu confirmer expérimentalement que des polymorphismes intra-isolat de fréquence de présence d'ET aux différents loci existent au sein de cet isolat alors que *M. incognita* est un organisme à reproduction strictement asexuée et dont la descendance est supposée très peu polymorphe, voire clonale. Ce résultat est d'autant plus étonnant que pour chaque isolat, l'ensemble des individus sont initialement issus de la masse d'œufs d'une seule et même femelle.

Dans l'hypothèse selon laquelle la diversité génétique observée serait une réalité biologique, elle pourrait avoir été favorisée par le mode de reproduction de cet organisme. En effet, une analyse de diversité génétique menée sur une population isolée (insulaire) de *Ruta microcarpa*, une plante à fleur à reproduction végétative, a permis de montrer que, chez cette espèce, la clonalité semble avoir un effet positif sur la diversité génétique en augmentant la diversité allélique, le polymorphisme et l'hétérozygotie (Meloni et al., 2013). Cependant, si à court terme la clonalité pourrait augmenter la diversité génétique, l'absence totale de reproduction sexuée pourrait, sur le long terme, mener à une population monoclonale optimale dans un environnement stable mais peu apte à s'adapter à un changement environnemental.

Il est intéressant de noter que chez l'isolat Morelos de *M. incognita*, ce phénomène pourrait avoir été amplifié artificiellement au cours du maintien de cet organisme en conditions contrôlées au sein du laboratoire. En effet, pour cet isolat, à chaque génération (*i.e* tous les 2-3 mois), un prélèvement de nématode est réalisé pour chaque lignée, plusieurs lignées étant maintenues en parallèle. La réduction de la taille de la population due à l'échantillonnage pourrait créer un goulot d'étranglement génétique et avoir pour effet de faciliter la dérive génétique, ce qui augmente l'hétérogénéité génomique au sein de la part restante de la lignée (Blumenstiel, 2019). En réalité, étant donné la fréquence des échantillonnages et donc des goulots d'étranglement potentiels engendrés (1 goulot par génération et par lignée), on peut faire l'hypothèse que ce mécanisme pourrait être une des sources principales de diversité génétique observée au sein de cet isolat et donc de la présence d'ETs avec des fréquences intermédiaires.

On notera néanmoins, que cette explication n'est probablement pas suffisante puisqu'on observe aussi une hétérogénéité de présence intra-isolat chez l'ensemble des isolats Brésiliens analysé précédemment (Kozlowski et al., 2020) alors qu'ils n'ont vraisemblablement pas subi ce type de biais ou du moins pas dans les mêmes fréquences lors de leur maintien.

Afin de tester ce biais de goulots d'étranglement au sein de l'isolat Morelos, nous avons réalisé des expériences de quantification de fréquence d'ET en triplicat biologique, chaque triplicat étant issu d'une lignée différente. Les résultats de dosage suggèrent qu'il existe une variabilité inter-réplicats (e.g. inter-lignées); en particulier en ce qui concerne le matériel génétique de 2019 pour lequel nous disposons d'une traçabilité plus claire. Un biais de goulots d'étranglement qui pourrait au moins en partie expliquer des variations de fréquence d'ETs au sein de l'isolat n'est donc pas exclu. Néanmoins, compte tenu du manque de précision de la technique de quantification utilisée (Qubit) (Nakayama et al., 2016) et du suivi des lignées (pas adapté à une analyse d'évolution expérimentale), nous ne pouvons trancher définitivement sur ce point. Des analyses complémentaires devront être réalisées dans le futur.

7 - Annexes

Tableau annexe 7.7.1 : amorces de PCR utilisées pour rapporter la présence ou l'absence d'un ET
Des couples spécifiques d'amorces ont été créés pour rapporter la présence ou l'absence d'un ET à 3 loci polymorphes. Pour chaque site polymorphe, 2 couples d'amorces ont été créés : un rapportant la présence de l'ET (étiquette « TE+ » dans le nom du couple d'amorce) ou son absence (« TE- »). Chaque couple est composé d'une amorce sens (« F ») et d'une amorce antisens (« R »)

Couples d'amorce	Séquence	Taille attendue de l'amplicon (pb)
01542_TE-_F	ACTTGGTACCAGCACACGGAGCTAA	172
01542_TE-_R	GGTGGTAAGGCGAAGAACGTTCGAG	
01542_TE+_F	ACTTGGTACCAGCACACGGAGCTAA	241
01542_TE+_R	TGCGACCTAGCTCGTCGAGA	
00187_TE-_F	TGAAAAGGATCAGTCGGTGG	171
00187_TE-_R	ACAAGTTTTTAGTTTTTCCTCAAC	
00187_TE+_F	GCTGAAAAATTCGTGCAAGTGC	260
00187_TE+_R	GGCACAAAGTTTACCCAACCA	
00004_TE-_F	ATTGTTAACCCAAAACGGCG	282
00004_TE-_R	AAAGGATTTAGTACCCTCCC	
00004_TE+_F	TGAAAGCCCAATCCCGTTCC	253
00004_TE+_R	GGTGTCATTCAGCCCTGCTC	
Actine_F	AAGATGGATGAAGAGGTAGCCGCCC	150
Actine_R	TGGAAAAACGGCACGAGGAGCA	

Tableau annexe 7.7.2 : efficacité des couples d'amorces et facteur d'amplification

Couple d'amorce	Efficacité (%)	Facteur d'amplification
01542_TE-_F	108.3	2.08
01542_TE-_R		
01542_TE+_F	130.3	2.30
01542_TE+_R		
00187_TE-_F	83	1.83
00187_TE-_R		
00187_TE+_F	109	2.09
00187_TE+_R		
00004_TE-_F	106.4	2.06
00004_TE-_R		
00004_TE+_F	100.6	2.01
00004_TE+_R		
Actine_F	98	1.98
Actine_R		

Tableau annexe 7.7.3 : dosage Qubit en triplicat de la quantité d'ADN avec et sans ETs pour les 3 loci sur 2 points de cinétique.

Q(x) -- quantités d'ADN (ng/microL) présente dans l'échantillon à l'issue de l'amplification PCR (20 cycles)

iniQ(x) -- quantités d'ADN (ng/microL) initialement présente dans l'échantillon calculé selon la formule suivante : iniQ(x) = Q(x)/(facteur d'amplification (x) ^ nb. cycle) avec x un couple d'amorce, Q(x) la quantité d'ADN quantifié dans l'échantillon, et nb. cycle le nombre de cycle réalisé lors de la PCR (20 cycle pour l'ensemble des expériences).

normQ(x) -- quantités d'ADN initiale (ng/microL) normalisé par la quantité initiale d'actine dans l'echantillon. Calculé selon la formule suivante : norm(x) = iniQ(x)/iniQ(Actine) (pas d'unité). Cette normalisation rend les échantillons comparables.

normQ(TE+) / (normQ(TE-) + normQ(TE+)) -- proportion (d'ADN) dans l'échantillon pour lequel l'ET est présent. .NB: Il est ici considéré que la somme des quantités d'ADN pour lequel l'ET est présent ou absent est égale à la quantité totale d'ADN

*fréquence expérimentale de présence de l'ET (%) -- (normQ(TE+) / (normQ(TE-) + normQ(TE+)))*100 ; estimation expérimentale de la frequence de présence de l'ET dans l'échantillon.*

Les analyses PCR effectuées sur les deux points de cinétique T0 (2015) et T1(2019) ont été réalisées en triplicas biologiques pour deux points de cinétique (A, B et C pour T0 ; X, Y et Z pour 2019)

		2015			2019		
		A	**B**	**C**	**D**	**E**	**F**
01542	Q(TE-)	2,59	2,54	2,46	2,36	3,58	2,62
	Q(TE+)	2,52	2,14	2,18	2,82	4,14	2,12
	Q(Actine)	2,34	2,42	2,36	2,14	2,08	2,22
	initQ(TE-)	1,127E-06	1,106E-06	1,071E-06	1,027E-06	1,558E-06	1,140E-06
	initQ(TE+)	1,468E-07	1,247E-07	1,270E-07	1,643E-07	2,412E-07	1,235E-07
	initQ(Actine)	2,728E-06	2,822E-06	2,752E-06	2,495E-06	2,425E-06	2,589E-06
	normQ(TE-)	4,132E-01	3,918E-01	3,891E-01	4,117E-01	6,425E-01	4,405E-01
	normQ(TE+)	5,382E-02	4,419E-02	4,616E-02	6,585E-02	9,947E-02	4,772E-02
	normQ(TE+) / (normQ(TE-) + normQ(TE+))	1,152E-01	1,014E-01	1,061E-01	1,379E-01	1,341E-01	9,774E-02
	fréquence expérimentale de présence de l'ET (%)	11,52	10.14	10.61	13.79	13.41	9.77

00187	Q(TE-)	1,66	1,48	1,67	1,67	2,32	1,93
	Q(TE+)	2,86	2,74	2,94	2,26	5,76	3,48
	Q(Actine)	2,54	2,18	2,62	2,02	4,88	3,96
	initQ(TE-)	9,356E-06	8,341E-06	9,412E-06	9,412E-06	1,308E-05	1,088E-05
	initQ(TE+)	1,131E-06	1,083E-06	1,163E-06	8,937E-07	2,278E-06	1,376E-06
	initQ(Actine)	2,962E-06	2,542E-06	3,055E-06	2,355E-06	5,690E-06	4,617E-06
	normQ(TE-)	8,273E+00	7,699E+00	8,096E+00	1,053E+01	5,741E+00	7,905E+00
	normQ(TE+)	3,819E-01	4,263E-01	3,806E-01	3,794E-01	4,003E-01	2,980E-01
	normQ(TE+) / (normQ(TE-) + normQ(TE+))	4,412E-02	5,246E-02	4,490E-02	3,477E-02	6,518E-02	3,633E-02
	fréquence expérimentale de présence de l'ET (%)	4,41	5,24	4,49	3,47	6,51	3,63

00004	Q(TE-)	1,21	1,08	1,18	1,12	1,64	1,43
	Q(TE+)	2,08	1,85	1,89	1,71	3,34	2,22
	Q(Actine)	2,54	2,18	2,62	2,02	4,88	3,96
	initQ(TE-)	6,389E-07	5,703E-07	6,231E-07	5,914E-07	8,660E-07	7,551E-07
	initQ(TE+)	1,795E-06	1,597E-06	1,631E-06	1,476E-06	2,883E-06	1,916E-06
	initQ(Actine)	2,962E-06	2,542E-06	3,055E-06	2,355E-06	5,690E-06	4,617E-06
	normQ(TE-)	2,157E-01	2,244E-01	2,040E-01	2,511E-01	1,522E-01	1,635E-01
	normQ(TE+)	6,062E-01	6,282E-01	5,340E-01	6,267E-01	5,067E-01	4,150E-01
	normQ(TE+) / (normQ(TE-) + normQ(TE+))	7,375E-01	7,368E-01	7,236E-01	7,139E-01	7,690E-01	7,173E-01
	fréquence expérimentale de présence de l'ET (%)	73,8	73,4	72,4	71,4	76,9	71,3

Tableau annexe 7.4 : résultats de PCR quantitative au locus 01542 (avec et sans ET) sur deux points de cinétique.

Les réplicats biologiques B et C concernent le point de cinétique T0 (2015). Les réplicats biologiques X, Y et Z concernent le point de cinétique T1 (2019). L'expérience de PCR quantitative n'a pas fonctionnée pour l'ensemble du réplicat biologique A (absence d'amplification). Les résultats concernant ce réplicat biologique ne sont donc pas représentés ici. Pour une raison similaire, certains réplicats techniques ne sont pas représentés non plus.

Pour chaque couple d'amorce, nous avons calculé le cq moyen à partir des valeurs des réplicats techniques. Pour le calcul du delta Ct, nous avons utilisé l'actine comme contrôle endogène. Les valeurs de delta Ct correspondent à Cq avr(X)/cq avr(actine) ; X représentant les couples d'amorce TE+ ou TE-. Les valeurs de delta delta CT ont été obtenus en calculant : delta CT (TE-) – delta CT (TE+). Le k^-delta delta CT à été calculé en prenant k=2.19 (moyenne des efficacités des couples de primers TE- (efficacité : 2.08) et TE+ (efficacité : 2.30) du locus 01542).

réplicats biologique	amorce	réplicats technique	Cq	Cq avr	ΔCt	ΔΔCt	2,19^-ΔΔCt
B	TE-	1	25,15	25,47	6,74	-4,15	25,91
		2	25				
		3	26,27				
	TE+	1	30,1	29,63	10,89		
		2	29,15				
	Actin	1	18,71	18,73	-		
		2	18,8				
		3	18,68				
C	TE-	1	22,84	23,9	5,2	-6,68	187,73
		3	24,95				
	TE+	1	29,74	30,57	11,88		
		2	30,56				
		3	31,42				
	Actin	1	18,87	18,69	-		
		2	18,35				
		3	18,84				
X	TE-	1	26,35	26,67	7,62	-6,1	119,29
		2	26,9				
		3	26,76				
	TE+	1	33,06	32,77	13,72		
		2	33,2				
		3	32,04				
	Actin	1	19,02	19,05	-		
		2	18,92				
		3	19,21				
Y	TE-	2	22,89	22,86	5,74	-4,73	40,89
		3	22,82				
	TE+	1	28,21	27,59	10,47		
		2	27,45				
		3	27,1				
	Actin	1	17,05	17,12	-		
		2	16,98				
		3	17,32				
Z	TE-	2	24	23,63	5,56	-2,96	10,18
		3	23,26				
	TE+	2	26,6	26,59	8,52		
		3	26,58				
	Actin	1	17,65	18,07	-		
		2	17,94				
		3	18,63				

VIII - Discussion générale.

Les *Meloidogyne* sont des nématodes phytoparasites constituant un groupe d'espèces diversifié en termes de trait biologiques (modes de reproduction, gamme de plantes hôtes) et de traits d'histoire de vie (hybridation, niveau de ploïdie). A ce jour, une centaine d'espèces de *Meloidogyne* ont été décrites et présentent une vaste gamme de modes de reproduction. Curieusement, il a pu être observé que les espèces les plus nuisibles se reproduisent de manière strictement asexuée. Il a aussi été constaté que certaines de ces espèces à reproduction asexuée arrivent à contourner les défenses de plantes hôtes en un nombre de génération restreint, ce qui pourrait constituer un paradoxe évolutif. En effet, il est majoritairement admis qu'en l'absence de brassage génétique et de recombinaison, le pouvoir adaptatif d'un organisme est en théorie réduit. Les *Meloidogyne* les plus dommageables étant polyploïdes et hybrides, il a été émis l'hypothèse selon laquelle disposer de plusieurs génomes en un et de gènes en copies multiples très divergentes pourrait être lié à leur capacité à parasiter une large gamme d'hôtes dans de nombreux endroits à travers le monde (Blanc-Mathieu et al. 2017). En effet, les phénotypes des espèces hybrides peuvent surpasser ceux des espèces parentales et c'est ce que l'on appelle communément des phénotypes transgressifs (Dittrich-Reed and Fitzpatrick, 2013).
Bien que l'hybridation puisse effectivement être impliquée dans la large gamme d'hôtes et la vaste répartition géographique de ces espèces, cela ne peut en aucun cas expliquer la capacité de ces espèces à s'adapter à de nouveaux hôtes ou à contourner un gène de résistance en quelques générations.
Un ou plusieurs facteurs combinés permettant de créer rapidement de la diversité génétique dans les individus et populations sont donc nécessairement à l'œuvre afin de soutenir cette adaptabilité en l'absence de sexe.

Les Éléments Transposables (ETs), par leur capacité à se déplacer et à se multiplier dans les génomes, sont des acteurs reconnus de la dynamique des génomes et peuvent en outre directement être mis en relation avec plusieurs des facteurs de plasticité déjà étudiés pour ces espèces. Au cours de mes travaux de thèse, je me suis donc intéressé aux rôles des ETs dans la plasticité génomique et la diversité génétique au sein des espèces de *Meloidogyne*.

A - Contenu et activité des ETs à l'échelle du genre *Meloidogyne* : lien avec des traits de vie et des traits biologiques.

Je me suis employé à dresser un état des lieux du contenu en ETs entre différentes espèces du genre *Meloidogyne* pour avoir une idée de la dynamique globale des ETs au sein de ce groupe d'espèces.

Lien entre contenu en ETs, mode de reproduction, niveau de ploïdie et hybridation

Au cours de mes travaux de thèse, j'ai eu l'opportunité d'évaluer le contenu en ETs (charge et composition) de 7 espèces de nématodes du genre *Meloidogyne* au travers de plusieurs types d'analyses : *M. incognita, M. javanica, M.arenaria, M. enterolobii, M. luci, M. floridensis* et *M.graminicola*. Nous allons ici discuter de la relation potentielle entre le contenu en ETs (charge et composition) du génome de ces espèces et les variations de certains traits biologiques observés chez ces espèces, mais aussi leur histoire évolutive.

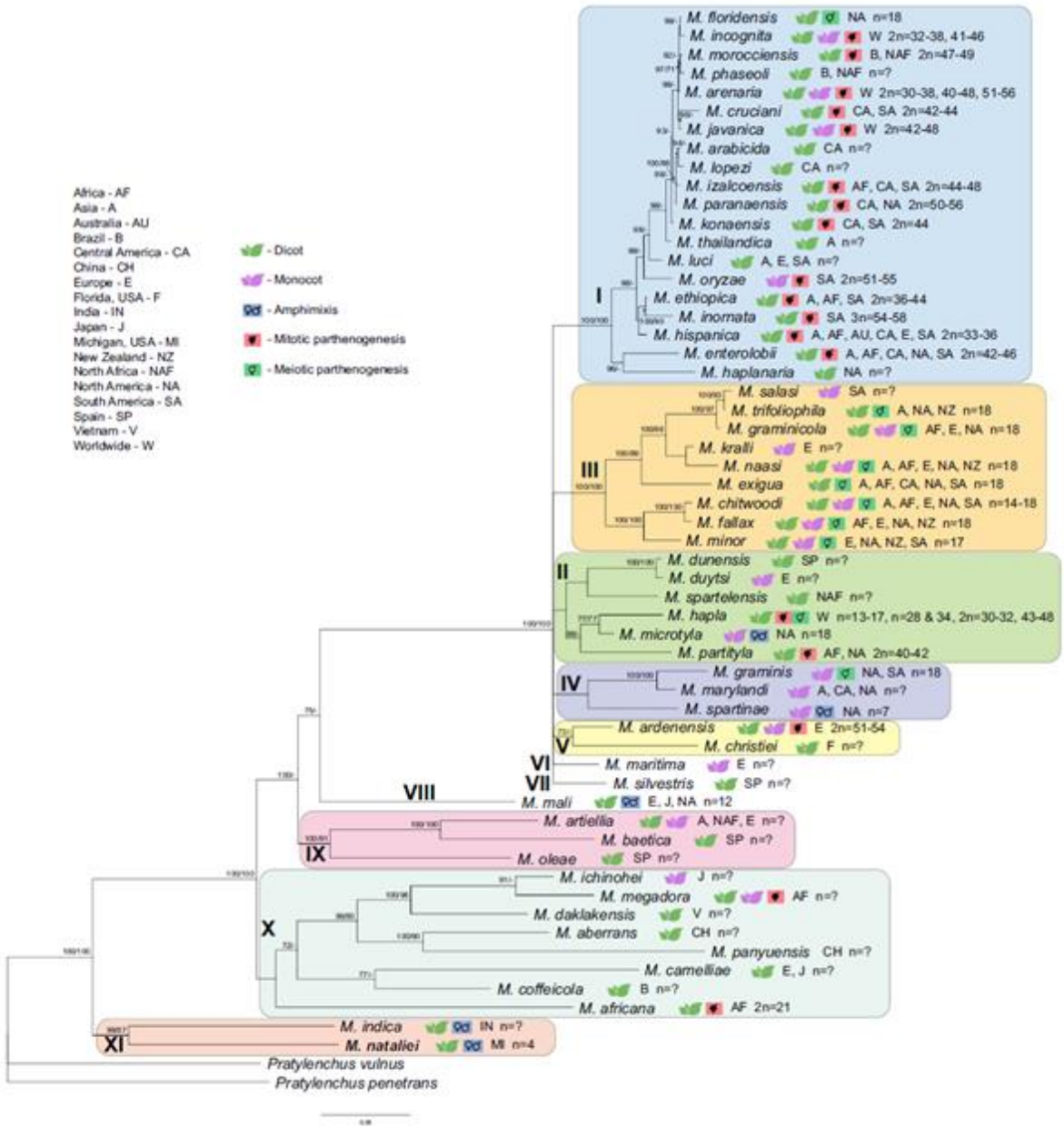


Figure 8.A.1 : arbre phylogénétique des espèces du genre Meloidogyne.

Figure issue de (Álvarez-Ortega et al., 2019).(Figure 4).

Arbre consensus majoritaire (50%) obtenu par inférence bayésienne (modèle GTR+I+G) à partir des alignements multiples des séquences de l'ARNr 18S, de l'ARNr ITS1, des segments d'expansion D2-D3 de l'ARNr 28S, du gène COI et l'ARNr COII-16S. La topologie n'est pas résolue pour les nœuds dont la valeur de support est inférieure à 70%. Les valeurs de support sont indiquées comme suit : valeur des probabilités postérieures dans l'analyse d'inférence bayésienne/valeur bootstrap de l'analyse de maximum de vraisemblance. n représente le nombre de chromosomes de l'espèce (n = ? -information sur le nombre de chromosomes inconnue).

La charge en ETs, i.e. la proportion du génome occupée par des ETs, varie de manière importante entre les espèces étudiées (facteur maximal de 6,67x : *M. graminicola* (1,50 % de génome occupé en moyenne) *vs. M. arenaria* (8,51 % en moyenne)). Cette variabilité ne semble pas être imputable à un biais méthodologique puisque pour trois espèces (*M. incognita, M. enterolobii*, et *M.graminicola*), deux types d'analyses ont été réalisées (une reposant sur l'utilisation de génomes assemblés et l'autre reposant sur les données de séquençage non assemblées) et les résultats obtenus se sont révélés cohérents entre les deux méthodologies. La variabilité de charge en ETs prédite entre les espèces de *Meloidogyne* étudiées semble donc être une réalité biologique.

La charge en ETs des espèces à reproduction asexuée obligatoire (*M. incognita, javanica, arenaria, enterolobii,* et *luci*; dans le clade I de la Figure 8.A.1) est largement plus importante que celle de l'espèce sexuée facultative du clade III *M. graminicola* (facteur minimum de 2,48: *M. graminicola v.s. M. incognita*). Ce résultat est cohérent avec une précédente analyse réalisée sur 4 espèces de *Meloidogyne* : *M. hapla, M. incognita, M. javanica* et *M. arenaria* (Blanc-Mathieu et al., 2017). Dans cette étude, il est décrit que *M. hapla* (clade II), une espèce à reproduction sexuée facultative, présente environ 46 % d'ETs en moins dans son génome que les trois espèces parthénogénétiques mitotiques (17% de génome occupé pour *M. hapla vs.* 27-30% pour *M. incognita, M. javanica* et *M. arenaria*). Mis ensemble, ces résultats pourraient toutefois laisser penser qu'il existe un lien entre la charge en ETs dans les génomes de *Meloidogyne* et leur mode de reproduction ; les espèces à reproduction sexuée (facultative) *M. hapla* et *M. graminicola* présentant une proportion plus faible du génome occupé par les ETs que les espèces asexuées obligatoires. Ceci serait cohérent avec la théorie selon laquelle les organismes se reproduisant de manière asexuée devraient accumuler des ETs étant donné que l'ensemble de leur génome ne peut les évacuer par recombinaison. Il serait donc tentant de conclure que la reproduction strictement asexuée induirait une expansion de la charge en ET chez les *Meloidogyne*.

Cependant, la distribution phylogénétique de ces espèces et d'autres facteurs pouvant eux aussi influer sur la charge en ETs doivent être considérés avant de pouvoir conclure.

En effet, on peut noter que ce regroupement d'espèces par mode de reproduction suit aussi leur phylogénie (Figure 8.A.1). Toutes les espèces à reproduction asexuée analysées dans ma thèse appartiennent à un seul et même clade (I). Les deux espèces capables de se reproduire sexuellement, *M. graminicola* (l'espèce présentant la plus faible charge d'ETs) et *M. hapla* sont respectivement les seuls représentants des clades III et II. Ainsi, les différences observées pourraient simplement révéler des divergences entre le clade I et les autres clades, sans nécessairement être liées au mode de reproduction.

Par ailleurs, il est aussi important de noter que toutes les espèces à reproduction asexuée obligatoire analysées ici(*M.incognita, javanica, arenaria, luci* et *enterolobii,* clade I), ainsi que *M. floridensis* dont le mode reproduction parthénogénétique exact n'est pas connu (parthénogénèse méiotique pouvant potentiellement être clonale) (Handoo, 2004), ont été décrites comme vraisemblablement hybrides et polyploïdes (Blanc-Mathieu et al., 2017; Koutsovoulos et al., 2020; Susic et al. 2020 ; Szitenberg et al., 2017). Au contraire, les espèces à reproduction sexuée facultative précédemment évoquées (*M. hapla* et *M. graminicola*) sont quant à elles diploïdes.

Or la polyploïdie et l'hybridation inter-espèce sont deux des processus connus comme pouvant être à l'origine d'une explosion (« burst») soudaine et rapide du nombre de copies d'un ou plusieurs ETs dans un génome. En effet, suite à hybridation, les ETs peuvent se retrouver dans un nouvel environnement génomique "naïf" vis à vis de ceux-ci et proliférer en l'absence de mécanismes de contrôle (résumé en détail dans (Belyayev, 2014)). Ainsi, la différence de charge observée entre les *Meloidogyne* du clade I par rapport aux deux autres espèces, des clades II et III, pourrait être due à leur origine allo-polyploïde plutôt qu'à leur mode de reproduction. Il est à noter que l'ensemble des espèces du clade I étudiées dans cette thèse possèdent des génomes mitochondriaux extrêmement proches, suggérant une origine maternelle commune récente (Blanc-Mathieu et al. 2017). Il est possible qu'un événement d'hybridation fondateur commun ait eu lieu à la base du clade I puis que des hybridations successives et indépendantes aient été à l'origine de la diversification des espèces à l'intérieur du clade I. Ces événements d'hybridation indépendants pourraient être responsables de la variabilité inter-espèce importante en termes de charge d'ETs au sein de ce clade.

En ce qui concerne la composition en ET, i.e. la diversité de classes et d'ordres d'ETs rencontrés ainsi que leur proportion relative dans les génomes, on peut observer que les transposons à ADN sont largement majoritaires dans l'ensemble des génomes de *Meloidogyne* étudiés, et ce quelle que soit la méthodologie utilisée. De manière intéressante, il semble que cette caractéristique soit commune à la plupart des espèces de nématodes pour lesquelles le contenu en ETs a été étudié (Bessereau, 2006; Szitenberg et al., 2016).

L'analyse de la répartition relative par ordre d'ET dans les génomes de *Meloidogyne* a montré qu'à quelques variations près, l'ensemble des espèces du clade I analysées présentent un profil de répartition par ordre d'ET similaire. De manière intéressante, *M. graminicola* présente un profil de répartition différent, principalement riche en Helitrons, et ce pour les deux souches analysées. *M. graminicola* et les espèces du clade I étant issus de lignées différentes, il est donc envisageable qu'une part de la composition en ETs décrite chez ces espèces soit la résultante de l'histoire évolutive de ces

deux lignées, ce qui est un phénomène courant dans le vivant puisque décrit tant chez les vertébrés (Chalopin et al., 2015) que chez les invertébrés (Petersen et al., 2019).

Mis ensemble, les résultats obtenus suggèrent qu'une part non négligeable de la variation de contenu en ETs (charge et composition) actuellement observée entre les deux lignées d'espèces analysées suit leurs histoires évolutives respectives. On peut en effet faire l'hypothèse que le profil de répartition des ET observé chez les espèces du clade I a été hérité d'un ancêtre commun.

Par ailleurs, le fait que les espèces du clade I présentent une charge d'ET plus importante que *M. graminicola* et dans des proportions différentes, suggèrent qu'un évènement d'amplification de certains ET (en particulier TIR) aurait eu lieu chez l'ancêtre commun des espèces du clade I. Il est intéressant de noter que ce phénomène de multiplication rapide d'un ou plusieurs ETs (« burst ») est souvent associé à la genèse de nouveaux groupes phylogénétiques (résumé en détail dans (Belyayev, 2014)). Ceci serait en accord avec l'idée d'une hybridation fondatrice du clade I suivie d'une expansion de TEs. Néanmoins la nature de la causalité entre expansion d'ETs et spéciation reste incertaine et hypothétique, un burst d'ET pouvant être une cause ou une conséquence (ou les deux) de cette radiation d'espèces.

Toutefois, l'ensemble des hypothèses émises précédemment ne peuvent réellement être confirmées faute de données suffisantes. En effet, ces hypothèses ont majoritairement été formulées en comparant un groupe d'espèces monophylétiques (clade I) à une seule espèce d'un autre clade (*M. graminicola*). Le regroupement par mode de reproduction et par niveau d' allo-ploïdie suivant précisément la phylogénie, il n'est pas possible en l'état de distinguer l'impact relatif de ces traits biologiques avec l'histoire évolutive des clades sur le paysage en ETs.

Afin de vraiment juger de l'impact respectif du mode de reproduction, de l'hybridation et du niveau de ploïdie sur la dynamique des ETs dans les génomes des *Meloidogyne*, il serait nécessaire d'acquérir des données sur plus d'espèces constituant autant de réplicats que possible des différents facteurs étudiés. Au sein du clade III dont est issu *M. graminicola*, il serait intéressant d'obtenir des données sur *M. fallax* et *M. exigua*, deux autres espèces parthénogénétiques méiotiques (Figure 8.A.1). De telles données constitueraient des réplicats de choix à la fois pour ce mode de reproduction et ce clade de l'arbre des *Meloidogyne*. Par ailleurs, il serait aussi nécessaire d'ajouter des espèces issues d'autres clades. A ce titre, *M. hapla* (2 sous-groupes : un parthénogénétique mitotique, l'autre parthénogénétique méiotique), *M. microtyla* (reproduction sexuée obligatoire) et *M. partityla* (parthénogénétique mitotique), toutes issues du clade II, constituent des organismes de choix. Bien

qu'un génome de *M. hapla* parthénogénétique méiotique soit disponible (Opperman et al., 2008), celui-ci a été généré avec des technologies de séquençage de première génération (i.e. Sanger), ne permettant pas une comparaison avec les autres génomes. Étant donné qu'elles font partie du même groupe, comparer ces espèces permettrait de se concentrer sur le rôle du mode de reproduction sur la dynamique des ETs en limitant l'impact de l'héritage phylogénétique profond. Afin de pleinement pouvoir étudier la gamme de mode de reproduction décrite chez les *Meloidogyne,* il faudrait acquérir des données sur *M. indica* et *M. natalei*, deux espèces se reproduisant exclusivement de manière sexuée ayant de surcroît une origine phylogénétique bien distincte des espèces précédemment citées (clade XI). Afin de distinguer les effets de l'allopolyploïdie et de la reproduction asexuée obligatoire, il serait nécessaire de séquencer le génome d'une espèce parthénogénétique mitotique non polyploïde. A ce titre, *M. africana* (clade X) est une espèce candidate de choix puisque décrite comme mitotique mais avec un nombre de chromosomes (*2n=22*) plus faible que les espèces polyploïdes (*xn>40-50*).

Enfin, afin d'étendre la question sur le lien entre éléments transposables et mode de reproduction, il serait ensuite judicieux de comparer les résultats obtenus chez les *Meloidogyne* avec ceux obtenus chez d'autres groupes d'organismes présentant eux aussi le même type de gamme de mode de reproduction. Une première étape pourrait consister à étudier d'autres groupes d'espèces de nématodes comme ceux du genre *Panagrolaimus* qui comprennent des espèces parthénogénétiques et amphimictiques et pour lesquels des génomes ont déjà été séquencés (Schiffer et al., 2019). Dans un second temps, il serait ensuite intéressant d'étudier d'autres groupes d'organismes, bien plus éloignés, présentant eux aussi des couples d'espèces proches à reproduction sexuée/aexuée comme c'est le cas chez les phasmes et les rotifères.

Activité des ETs au sein du genre *Meloidogyne*

Nous avons vu au travers des résultats précédents que les paysages en ETs, à la fois en termes de charge et de composition, montraient des variations entre les différentes espèces de *Meloidogyne*; en particulier entre groupes d'espèces. Cela suggère que ces ETs ont au moins été actifs à une période de l'histoire évolutive de ces espèces. Nous allons ici discuter de la dynamique globale de cette activité au cours du temps et du lien que celle-ci pourrait avoir avec des traits de vie et des traits biologiques rencontrés au sein du genre *Meloidogyne*.

Globalement chez les *Meloidogyne*, on peut observer que la plupart des copies d'ETs sont très proches de leurs consensus et donc vraisemblablement "jeunes", avec une activité récente à l'échelle du genre. Néanmoins, on peut aussi constater que le profil de répartition de l'âge des ETs varie entre les

espèces du clade I et *M. graminicola*. Le profil d'activité de *M. graminicola*, confirmé par deux méthodologies différentes et retrouvé chez deux souches distinctes, suggère qu'une amplification très soudaine du nombre d'ETs (en particulier des Helitrons) pourrait avoir eu lieu au sein de cette espèce. Or il a récemment été proposé que *M. graminicola* serait une espèce hybride homoploïde (Phan et al., 2020). Il est donc possible que l'augmentation soudaine du nombre de copies d'Helitrons prédite soit liée à cet événement d'hybridation n'affectant pas le niveau de ploïdie. Cependant, il est important de noter que la charge globale en ETs du génome de cette espèce est relativement faible en comparaison des autres *Meloidogyne* (venant du clade I). Le profil observé pourrait donc au contraire être révélateur d'un mécanisme de répression efficace permettant de rapidement contrôler l'activité des ETs. *M. graminicola* étant par ailleurs une espèce parthénogénétique facultative capable de faire de la recombinaison méiotique, il est possible que les ETs soient ensuite éliminés plus facilement du génome et/ou ne soient pas transmis à la descendance. Le profil observé serait alors le résultat d'une déplétion des ETs les plus anciens plutôt que d'un burst très récent. Afin de déterminer si le profil d'activité des ETs observé chez *M. graminicola* est propre à cette espèce ou à la lignée dont elle fait partie, il serait nécessaire d'acquérir des données sur d'autres espèces du Clade III comme cela à déjà été proposé plus haut dans cette discussion.

Bien que présentant un profil caractéristique d'une activité récente, les espèces du clade I présentent un profil d'activité plus progressif que celui de *M. graminicola*. Par ailleurs, il semble que plusieurs ordres d'ETs aient récemment subi une expansion chez ces espèces, aboutissant à une charge globale plus importante dans le génome. Cette différence de profil d'activité est donc ici encore cohérente avec la phylogénie des espèces de *Meloidogyne* puisqu'elle coïncide avec la différence de composition d'ETs constatée entre les espèces du Clade I et *M. graminicola* (Clade III). Ici encore, cette rapide augmentation du nombre de copies d'ETs pourrait être expliquée par une hausse d'activité des ETs lors d'évènements d'hybridations chez les ancêtres communs respectifs de ces deux groupes d'espèces.

Comme cela a précédemment été mis en avant, les ETs ont probablement été récemment actifs au sein des génomes des espèces de *Meloidogyne*. Le fait que l'on trouve une majorité de jeunes copies traduit le fait que des ETs se sont récemment multipliés et déplacés dans les génomes. De ce fait, les ET ont donc probablement participé à la plasticité génomique de ces espèces et il est envisageable que cette activité ait eu des répercussions fonctionnelles.

B - Apport des ETs à la plasticité génomique et relation avec la gamme d'hôtes : le cas de *M. incognita*.

M. incognita est souvent décrite comme étant l'espèce de *Meloidogyne* la plus nuisible et de manière générale l'un des plus grands ravageurs de culture (Trudgill and Blok, 2001). De manière intéressante, il a pu être observé en conditions contrôlées et en champ que *M. incognita* est capable de contourner les défenses de sa plante hôte en un nombre de générations restreint alors qu'il s'agit d'une espèce parthénogénétique mitotique (reproduction strictement asexuée) (Castagnone-Sereno et al., 1994; Castagnone-Sereno et al., 2019; Tzortzakakis et al., 2014). Un ou plusieurs facteurs permettant de créer rapidement de la plasticité génomique sont donc nécessairement à l'œuvre afin d'expliquer cette adaptabilité paradoxale en absence de recombinaison. Chez *M. javanica*, une autre espèce à reproduction strictement asexuée, la comparaison entre une lignée avirulente (i.e. incapable d'infecter les plants de tomates portant un gène de résistance aux nématodes) et une autre lignée virulente (qui a surmonté cette résistance) a mené à l'hypothèse que l'excision d'un transposon à ADN TIR aurait induit la délétion d'un gène d'avirulence, reconnu par l'hôte et aurait ainsi joué un rôle dans le contournement de la résistance (Gross and Williamson, 2011). Ainsi les ETs, de par leur capacité à se déplacer et à se multiplier dans les génomes, constituent donc un des mécanismes potentiels permettant d'expliquer l'adaptabilité de ces espèces en absence de reproduction sexuée.

Dans cette idée, après avoir annoté en détail le contenu en ETs de *M. incognita*, j'ai réalisé une analyse populationnelle entre 12 isolats géographiques présentant des variations de gamme d'hôtes afin i) de mieux caractériser l'activité des ETs au sein de cette espèce, et ii) de définir si l'activité des ET pouvait être mise en relation avec des variations de compatibilité d'hôtes entre ces isolats.

J'ai identifié plusieurs milliers de loci pour lesquels les fréquences de présence d'ETs varient entre les différents isolats, signe que les ETs participent activement à la diversité génétique entre isolats chez cette espèce. En utilisant les fréquences d'ETs comme signal phylogénétique, j'ai montré que le contenu en ETs des isolats étudiés suit leur niveau de divergence génomique mesurée à l'aide de SNPs. Les groupes phylogénétiques obtenus ne montrent pas de superposition avec les différentes gammes d'hôtes de ces 12 isolats. Aucun profil général de répartition de fréquence ne peut être mis en lien avec des regroupements par gamme d'hôte; il n'existe donc pas de lien entre les variations du contenu global en ETs dans le génome de cette espèce et les différences en termes de facultés parasitaires observées chez ces isolats. Les variations de fréquences d'ETs prises dans leur ensemble ne semblent donc pas corréler avec une adaptation à certaines plantes hôtes. De plus, nous n'avons pas non plus observé de

corrélation avec leur distribution géographique. Cette observation est en accord avec l'hypothèse selon laquelle l'évolution des ET est majoritairement neutre dans les génomes (Arkhipova, 2018).

Bien qu'à l'échelle de l'ensemble du paysage, les variations en fréquences d'ETs suivent la phylogénie et non les différentes gammes d'hôtes, j'ai tout de même souhaité vérifier si la variation en fréquence de certaines copies d'ETs particulières pouvait être liée à ces gammes d'hôtes. Même à cette échelle, aucun lien n'a pu être clairement établi. On notera néanmoins que le caractère « race d'hôte » analysé ne concerne que la capacité d'une espèce à infecter 4 plantes d'intérêt agronomique (Hartman and Sasser, 1985). Il n'est donc pas exclu qu'il existe bel et bien des associations entre la plasticité génomique apportée par les ET et leur capacité à infecter d'autres plantes ou à s'adapter à certains milieux.

Par ailleurs, de nombreuses néo-insertions ont pu être détectées entre ces différents isolats et une majeure partie, en étant insérées à proximité ou à l'intérieur de gènes, pourraient avoir de potentiels impacts fonctionnels. De plus, certaines de ces néo-insertions sont retrouvées à des fréquences proches de la fixation chez certains isolats et sont absentes chez d'autres. Fait intéressant, certains de ces ETs fixés sont communs à plusieurs isolats, et la répartition de ces ETs fixés suit également l'histoire évolutive des isolats analysés, en étant présents chez certains groupes d'isolats et absents chez d'autres. Il est donc probable que ces néo-insertions aient été héritées après être apparues chez l'ancêtre commun des isolats concernés. Or, l'ensemble des isolats analysés proviennent du Brésil et il est probable qu'ils aient été disséminés par l'activité agricole humaine. Ceci suggère donc qu'en un laps de temps court, des néo-insertions seraient apparues et auraient augmenté en fréquence jusqu'à être fixées, le tout avant que les différents isolats ne divergent les uns des autres. Ce résultat semble indiquer que bien qu'une majorité des ETs semble évoluer de manière neutre chez cette espèce, un nombre non négligeable suivent une dynamique de propagation rapide dans les populations, suggérant que certains pourraient être sous pression de sélection positive, ou du moins ne pas être contrôlés par le génome hôte.

C – Diversité génétique au sein d'un individu clonal

Au cours de l'analyse populationnelle précédemment discutée, j'ai pu constater que chacun des 12 isolats de *M. incognita* étudiés présentait indépendamment des ETs en fréquences intermédiaires (i.e fréquence comprise entre 25% et 75%), suggérant qu'il existe une hétérogénéité de présence des ETs entre individus au sein des isolats. Pour l'un de ces isolats (Morelos), nous avons pu confirmer expérimentalement que cette hétérogénéité de présence d'ET est une réalité biologique. Ce résultat est d'autant plus surprenant que i) *M. incognita* est un organisme parthénogénétique mitotique (confirmé en cytogénétique par l'absence d'observation d'évènements de fusion des noyaux mâles et femelles (Triantaphyllou, 1981), et in silico par l'absence de signes de recombinaison méiotique (Koutsovoulos et al., 2020)), et ii) que cet isolat a originellement été constitué à partir de la masse d'œufs d'une seule femelle.

La diversité génétique actuellement observée pourrait être la conséquence combinée de variations génétiques présentes dans la masse d'œufs originelle ou accumulée depuis, et de goulots d'étranglements successifs ayant eu pour effet de redistribuer la fréquence de ces variations dans la population. Afin de tester l'hypothèse de l'existence d'une variabilité génétique à l'intérieur d'une masse d'œuf (issue d'une seule femelle), il serait intéressant de réaliser une analyse de séquençage du génome en single-cell sur une centaine d'oeufs individuellements a minima, lorsque la technologie le permettra.

Par ailleurs, en plus de cette variabilité en fréquence observée entre isolats à un temps donné, l'analyse de dynamique des fréquences d'ETs polymorphes au cours du temps au sein d'un isolat de *M. incognita* (isolat morelos) a permis de montrer que certains ETs décrits dans le génome de référence varient en fréquence au cours du temps. Une majorité de ces sites polymorphes suivent une dynamique de diminution de fréquence au cours du temps, signe que ces ETs auraient tendance à progressivement être expulsés du génome de cet isolat. Ceci pourrait sembler étonnant car dans les chapitres précédents nous avons présenté la méiose comme mécanisme permettant d'éliminer plus efficacement les ETs et la méiose est absente chez *M. incognita*. Cependant deux théories contraires s'opposent concernant les effets de la méiose sur la charge en ETs. Une hypothèse alternative voit en effet la méiose comme une opportunité pour les ETs de se transmettre 'sexuellement' d'un génotype à un autre alors que chez les asexués les ETs seraient "piégés" dans la seule lignée descendante d'un clone. Les ETs seraient donc plus rapidement éliminés par sélection purifiante ou dérive chez les asexués (Wright and Finnegan, 2001). En accord avec cette théorie, une expérience récente de comparaison de l'évolution de la charge d'ETs au fil du temps chez des populations à reproduction sexuées ou asexuées de *Saccharomyces*

cerevisiae a montré que la charge d'ETs diminue rapidement dans les populations à reproduction asexuée (Bast et al., 2019). Par ailleurs, à l'aide de simulations, les auteurs de cette même étude ont pu montrer que cette réduction du nombre d'ETs se produit très probablement par l'augmentation des taux d'excision. Or, dans le génome de *M. incognita*, les transposons à ADN, i.e. les transposons les plus susceptibles de s'exciser, sont majoritaires. Ce phénomène pourrait donc être amplifié par la nature des ETs présents dans le génome de *M. incognita*.

On pondèrera néanmoins ce résultat par le fait que ces diminutions de fréquence au cours du temps dans l'isolat Morelos ne touchent en réalité qu'une minorité des annotations d'ETs décrites dans le génome de référence. Par ailleurs, bien qu'elles concernent moins d'ETs, les augmentations en fréquences sont plus intenses, ce qui laisse supposer que certains de ces loci pourraient être sous pression de sélection positive.

Face à ces résultats contradictoires, il est difficile de trancher quant à la dynamique à court terme des ETs au sein de cet isolat.

Outre un manque de précision dans l'estimation *in silico*, certains biais méthodologiques pourraient expliquer les variations de fréquences observées au cours du temps au sein de cet isolat. Le biais principal, serait une accumulation de goulots d'étranglement lors des prélèvements de nématodes qui amplifierait artificiellement la diversité génétique dans la population (dérive génétique). En effet, chaque prélèvement entraîne une réduction variable mais a priori non négligeable de la taille de la population, ce qui pourrait avoir pour effet de redistribuer les fréquences de présence des ETs dans la population.

Nous avons tenté d'évaluer l'importance de ce biais par la comparaison de lignées du même isolat au cours du temps. Néanmoins, ni les techniques de quantification utilisées, ni la traçabilité des lignées ne nous ont permis d'infirmer ou de confirmer cette hypothèse de manière certaine. Ceci est peut-être une conséquence du fait que ces organismes étant supposément clonaux, les manipulations de maintien de lignées en laboratoire n'ont jamais été pensées en termes de variabilité de la population maintenue. Une expérience d'évolution expérimentale structurée serait nécessaire afin de tester cette hypothèse.

En ce qui concerne l'évaluation des fréquences de présence au sein des isolats, l'alternative la plus "simple", mais aussi la plus lourde et pas encore mise en œuvre actuellement, consisterait à réaliser un séquençage « single individual » pour pouvoir valider la présence/absence des ETs par individu et ensuite pouvoir calculer les fréquences de ces ETs dans les échantillons, i.e isolats ou lignées.

IX - Perspectives générales

Plusieurs suggestions de futures analyses en lien direct avec les questions abordées au cours de cette thèse ont déjà été proposées au fil de la précédente discussion. Néanmoins, je souhaite ici rajouter plusieurs suggestions complémentaires qui pourraient directement venir enrichir cette étude.

Certaines questions n'ont pu être abordées au cours de cette thèse faute de disposer d'un génome suffisamment contigu et/ou de posséder des données permettant d'y répondre.

Un génome plus contigu permettrait en effet de s'intéresser à la présence ou non d'une compartimentation des ETs dans le génome, avec par exemple des régions riches en ETs mais pauvres en gènes et d'autres régions pauvres en ETs mais riches en gènes, des différences en composition de GC pouvant éventuellement y être associées. C'est le cas chez des phytopathogènes filamenteux tels que des champignons et des oomycètes chez lesquels ont été décrits des génomes ayant des compartiments à plusieurs (deux voire trois) vitesses (Faino et al., 2016).

En outre, chez *M. incognita* il a été proposé que des variations du nombre de copies de gènes (CNV), et en particulier des délétions convergentes, auraient permis à cet organisme d'acquérir un phénotype virulent, c'est à dire de contourner la résistance d'une plante (Castagnone-Sereno et al., 2019). Compte tenu de leur rôle structural reconnu, un génome de haute qualité permettrait donc de déterminer si les ETs pourraient jouer un rôle passif (en favorisant de la recombinaison illégitime) ou actif (en embarquant des gènes lors de leur mouvements) sur des variations structurales du type CNV.

L'acquisition de données d'expression (RNAseq) supplémentaires permettraient de dresser un portrait plus clair du rôle des ETs dans la plasticité génomique de ces organismes. Par exemple, avoir de telles données pour chacun des isolats brésiliens de *M. incognita* précédemment étudiés fournirait un autre caractère observable pour définir si l'activité des ETs décrite au sein de ces isolats auraient de potentiels impacts fonctionnels en lien avec les variations de gamme d'hôte observées chez cette espèce. Par ailleurs, l'acquisition de données sur des populations virulentes et avirulentes permettrait aussi d'explorer plus en détails le lien entre ETs et contournement de la résistance par un hôte. Une seconde étape consisterait ensuite à analyser des données d'expression par stade de développement pour ces mêmes populations afin de quantifier l'impact des ETs sur la régulation d'expression de gènes liés au parasitisme ou à la reconnaissance par l'hôte au cours du développement.

Il serait aussi intéressant de pouvoir quantifier l'activité des ETs a un instant *t*, i.e déterminer à quel rythme les ETs transposent, afin de voir si la dynamique « instantanée » de cette activité peut être mise en relation avec un stress et/ou le parasitisme. Ce type de données permettrait par exemple de voir si on observe une hausse d'activité des ETs lors de l'inoculation de nématodes avirulents sur une plante résistante (cet hôte résistant pouvant constituer un stress pour le nématode). Au cours de ma thèse, j'ai commencé à aborder cette question via l'analyse de données de séquençage eccDNA (résultats non présentés) disponibles pour *M. incognita*. La philosophie générale de cette technique est la suivante. Lors d'un évènement de transposition, il arrive que l'ET se circularise avant d'être réintégré ailleurs dans le génome (eccDNA pour « extra chromosomal circular DNA »). Le séquençage eccDNA consiste à isoler et séquencer spécifiquement ces fragments d'ADN circulaire, ce qui permettrait en théorie ensuite de caractériser l'activité des ETs comme cela a pu être montré dans le génome du riz (Lanciano et al., 2017). Chez *M. incognita*, les différentes approches entreprises ne se sont pas révélées concluantes, la quantité de données séquencées correspondant réellement à des ETs étant minoritaires. Ce résultat est cohérent avec le fait que chez la levure, plus de 23% du génome se circularise, ces séquences correspondant tant à des séquences répétées que non répétées (Møller et al., 2016). La technique de séquençage eccDNA n'est donc a priori pas complètement adaptée à la caractérisation d'activité des ETs faute de spécificité. On notera néanmoins que comme les données que j'ai eu en ma possession étaient issues d'un crible pilote, ces résultats sont à prendre avec du recul.

Des travaux précédents ont montré qu'il existait au final que très peu de variations nucléotidiques (ie SNPs) entre différents isolats de *M. incognita*, bien que ceux-ci puissent présenter des phénotypes différents en termes de gamme d'hôtes (Koutsovloulos et al. 2019). De plus, ces quelques variations nucléotidiques ne se superposent pas avec les différentes gammes d'hôtes. Une partie des travaux réalisés au cours de ma thèse montrent la même chose au niveau des ETs. Il semblerait donc, que globalement, peu de mutations au sens large, qu'elles soient à petite échelle (SNP) ou plus large échelle (ETs) existent entre les différentes lignées de *M. incognita*. Par ailleurs, à ce jour aucun lien n'a pu être établi entre ces mutations et un caractère adaptatif (bien que très peu aient été mesurés). *M. incognita* semble donc capable de s'adapter rapidement sans que des mutations sous-tendent nécessairement cette adaptation. Il est donc tout à fait naturel d'imaginer qu'il puisse y avoir une composante épigénétique soutenant cette adaptabilité. Chez *M. incognita*, des recherches actives sont menées au sein du laboratoire pour déterminer s'il existe une composante épigénétique liée au contournement des défenses de la plante. Compte tenu de la relation intime qu'il existe entre les ETs et les régulations épigénétiques d'expression de gènes, il sera intéressant dans le futur de recouper les résultats obtenus afin d'évaluer la dynamique entre ces deux facteurs de plasticité sur l'adaptabilité de cet organisme.

Enfin une dernière perspective, plus éloignée du sujet original, serait d'étudier les transferts horizontaux d'ET (THET) chez les *Meloidogyne*. En effet, les THET sont considérés comme un moyen pour les ETs de garantir leur persistance à long terme, en passant d'hôtes capables de réprimer leur activité à des hôtes naïfs dans lesquels ils peuvent produire de nouvelles copies (Gilbert et al., 2010). Fait intéressant, l'introduction d'un ET « étranger » a souvent pour conséquence son amplification rapide dans le génome « naïf » (Belyayev, 2014), ce qui pourrait donc aussi être un facteur supplémentaire ou alternatif permettant d'expliquer certains profils d'activité observés comme l'explosion du nombre de copies d'Helitrons chez *M. graminicola*. L'analyse de l'apport des THET chez les *Meloidogyne* pourrait être étudiée à deux niveaux. Le premier consisterait à identifier d'éventuels transferts entre nématodes. En effet, plusieurs espèces de *Meloidogyne* sont souvent présentes simultanément dans un même champ/dans une même plante. Il n'est donc pas exclu que de tels évènements de transferts aient pu avoir lieu par exemple par l'intermédiaire d'un organisme du microbiote racinaire. Un deuxième type de transfert, plus hypothétique, pourrait aussi avoir eu lieu entre les nématodes et les plantes infectées (i.e transfert horizontal d'ETs inter-règne), les deux organismes étant en constante interaction du fait du mode de vie parasitaire du nématode. Il a en effet été suggéré que ce type d'évènement (transfert d'un rétrotransposon Penelope) pourrait avoir lieu entre les arthropodes et les conifères (Lin et al., 2016).

X – Annexes

A - Liste des publications réalisées au cours de la thèse

Premier auteur

Kozlowski, D.K., Hassanaly-Goulamhoussen, R., Da-Rocha, M., Koutsovoulos, G., Bailly-Bechet, M., Danchin, E.G., 2020. Transposable Elements are an evolutionary force shaping genomic plasticity in the parthenogenetic root-knot nematode Meloidogyne incognita (preprint). Biorxiv Evolutionary Biology. https://doi.org/10.1101/2020.04.30.069948 (Recommanded by PCI Evolutionary biology : https://doi.org/10.24072/pci.evolbiol.100106)

Collaborations

Simion, P., Narayan, J., Houtain, A., Derzelle, A., Baudry, L., Nicolas, E., Cariou, M., Guiglielmoni, N., **Kozlowski, D.K.**, Gaudray, F.R., Terwagne, M., Virgo, J., Noel, B., Wincker, P., Danchin, E.G., Marbouty, M., Hallet, B., Koszul, R., Limasset, A., Flot, J.-F., Van Doninck, K., 2020. Homologous chromosomes in asexual rotifer *Adineta vaga* suggest automixis (preprint). Biorxiv Evolutionary Biology. https://doi.org/10.1101/2020.06.16.155473

Phan, N.T., Orjuela, J., Danchin, E.G.J., Klopp, C., Perfus-Barbeoch, L., **Kozlowski, D.K**., Koutsovoulos, G.D., Lopez-Roques, C., Bouchez, O., Zahm, M., Besnard, G., Bellafiore, S., 2020. Genome structure and content of the rice root-knot nematode (*Meloidogyne graminicola*). Ecol. Evol. ece3.6680. https://doi.org/10.1002/ece3.6680

Koutsovoulos, G.D., Marques, E., Arguel, M., Duret, L., Machado, A.C.Z., Carneiro, R.M.D.G., **Kozlowski, D.K.**, Bailly-Bechet, M., Castagnone-Sereno, P., Albuquerque, E.V.S., Danchin, E.G.J., 2020. Population genomics supports clonal reproduction and multiple independent gains and losses of parasitic abilities in the most devastating nematode pest. Evol. Appl. 13, 442–457. https://doi.org/10.1111/eva.12881

Koutsovoulos, G.D., Poullet, M., Elashry, A., **Kozlowski, D.K.**, Sallet, E., Da Rocha, M., Perfus-Barbeoch, L., Martin-Jimenez, C., Frey, J.E., Ahrens, C.H., Kiewnick, S. & Danchin, E.G.J., 2020. The polyploid genome of the mitotic parthenogenetic root-knot nematode *Meloidogyne enterolobii* (in press). Scientific Data.

*Anthonomus grandi*s genome paper (in preparation). Arraes, F., [. . .], **Kozlowski, D.K.**, [. . .].

Phytophthora parasitica genome paper (in preparation). Panabieres, F., [. . .], **Kozlowski, D.K.**, [. . .].

B - Suppléments de l’article (Kozlowski et al., 2020)

Supplementary material:

Transposable Elements are an evolutionary force shaping genomic plasticity in the parthenogenetic root-knot nematode *Meloidogyne incognita*

Djampa KL KOZLOWSKI, Rahim HASSANALY-GOULAMHOUSSEN, Martine DA-ROCHA, Georgios KOUTSOVOULOS, Marc BAILLY-BECHET*, Etienne GJ DANCHIN*.

* co-last authors
Affiliation : Université Côte d'Azur, INRAE, CNRS, ISA, Sophia Antipolis, France

Table S1: Per-order summary of *M.incognita* draft TE annotations.

Autonomous TE orders (*) regroup elements known to present transposition machinery and thus able to transpose by themselves. On the opposite, non-autonomous orders (**) regroup elements lacking transposition machinery and therefore relying on autonomous elements to transpose. "Class 1 & 2 like" regroup elements for which homology-based evidence is sufficient to support an assignment to class I (retro) or II (DNA-transposon), but insufficient to assign a known order. "PotHostGenesOrOther" classification regroups elements which most likely correspond to duplicated genes. "Unclassif." elements are repetitive elements without sufficient evidence to be classified as class I (retro) or II (DNA-transposon). "Class 1 & 2 like", "PotHostGenesOrOther", and "Unclassif." are removed in the canonical TE annotations.

	order autonomous (*) / non-autonomous (**)	**nb. of features**	**total length (bp)**	**genome percentage (%)**	**median length (bp)**	**median identity with consensus (%)**
Retro - transposon	SINE (**)	19	6,618	0.004	258.0	87.6
	LARD (**)	217	132,969	0.072	244.0	92.35
	TRIM (**)	2,466	1,240,016	0.676	468.0	76.3
	LINE (*)	970	822,008	0.448	477.0	76.7
	LTR (*)	2,878	2,702,453	1.472	429.5	77.8
DNA - transposon	Helitron (*)	152	282,819	0.154	742.0	78.1
	Maverick (*)	17,684	9,553,119	5.205	364.0	74.8
	MITE (**)	12,435	5,126,098	2.793	363.0	88.5
	TIR (**)	11,094	5,389,275	2.936	379.0	85.0
Others	CLASS_1_LIKE	11,053	6,737,590	3.671	522.0	74.1
	CLASS_2_LIKE	77	34,339	0.019	497.0	98.7
	potHostGenesOrOther	26,225	12,185,975	6.640	359.0	75.1
	unclassif	8,811	4,212,017	2.295	390.0	79.0
	Total	**94,081**	**48,425,296**	**26.385**		

Table S2: Per-order summary of *C.elegans* draft TE annotations.

	order autonomous (*) / non-autonomous (**)	**nb. of features**	**total length (bp)**	**genome percentage (%)**	**median length (bp)**	**median identity with consensus (%)**
Retro - transposon	SINE (**)	85	51,197	0.051	479.0	89.7
	LARD (**)	14	17,043	0.017	572.5	87.8
	TRIM (**)	3,324	2,184,226	2.178	485.0	79.1
	LINE (*)	519	480,089	0.479	538.0	96.4
	LTR (*)	246	215,384	0.215	509.0	96.15
DNA - transposon	Helitron (*)	2,865	2,103,981	2.098	547.0	77.2
	Maverick (*)	26	39,843	0.040	680.0	95.25
	MITE (**)	4,274	1,752,665	1.748	322.0	81.4
	TIR (**)	3,840	2,499,195	2.492	413.0	90.45
Others	CLASS_1_LIKE	46	19,873	0.020	385.5	85.325
	CLASS_2_LIKE	5,607	1,678,689	1.674	230.0	81.3
	potHostGenesOrOther	742	497,310	0.496	407.0	75.0
	unclassif	372	314,317	0.313	729.5	92.0
	Total	**21,960**	**11,853,812**	**11.820**		

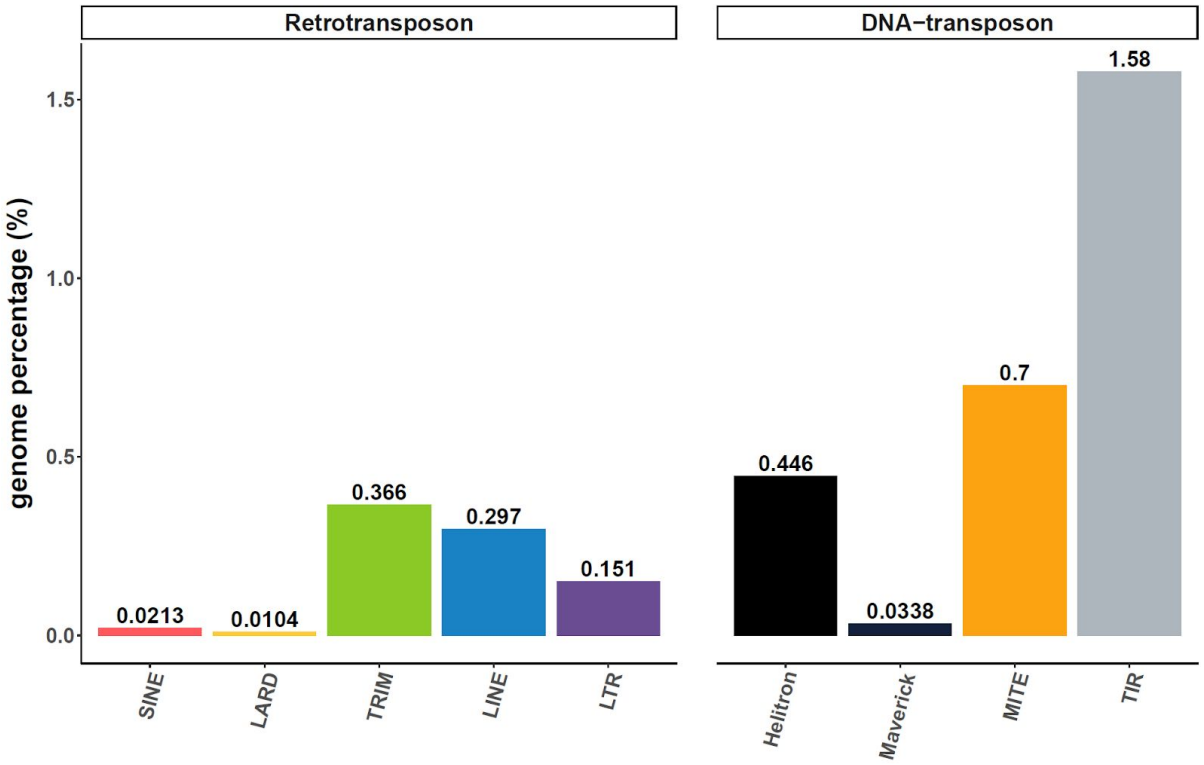


Fig S1: Canonical TE annotations distribution in the *C. elegans* genome

Genome percentage is based on a *C. elegans* genome size of 100,286,401 bp.

Table S3: Per-order summary of *C.elegans* canonical TE annotations.

	order autonomous (*) / non-autonomous (**)	**nb. of features**	**total length (bp)**	**genome percentage (%)**	**median length (bp)**	**median identity with consensus (%)**
Retro - transposon	SINE (**)	23	21,342	0.021	908.0	98.1
	LARD (**)	3	10,417	0.010	3969.0	99.7
	TRIM (**)	294	366,742	0.366	744.5	90.6
	LINE (*)	184	297,840	0.297	1252.5	98.7
	LTR (*)	124	151,145	0.151	617.5	97.75
DNA - transposon	Helitron (*)	267	447,385	0.446	1514.0	96.1
	Maverick (*)	14	33,884	0.034	1399.5	97.6
	MITE (**)	1,101	702,012	0.700	521.0	95.0
	TIR (**)	1,475	1,582,321	1.578	815.0	97.1
	Total	**3,485**	**3,613,088**	**3.603**		

Table S4: *M. incognita* per-order summary of copies % identity with their consensus.

	Min.	1st Quantile	Median	Mean	3rd Quantile	Max.
Helitron	85.3	92.0	94.4	93.4	95.8	97.7
LARD	92.6	96.1	97.1	96.9	97.9	99
LINE	85.9	95.4	96.6	96.3	98.8	100
LTR	85.3	94.8	97.0	96.3	98.4	100
Maverick	85.0	90.3	95.3	93.7	97.2	99.8
MITE	85.0	92.8	96.2	95.3	98.4	100
SINE	93.4	99.3	99.7	98.9	99.8	100
TIR	85.0	93.7	97.3	96.0	99.2	100
TRIM	85.2	95.7	97.7	96.8	98.8	99.9

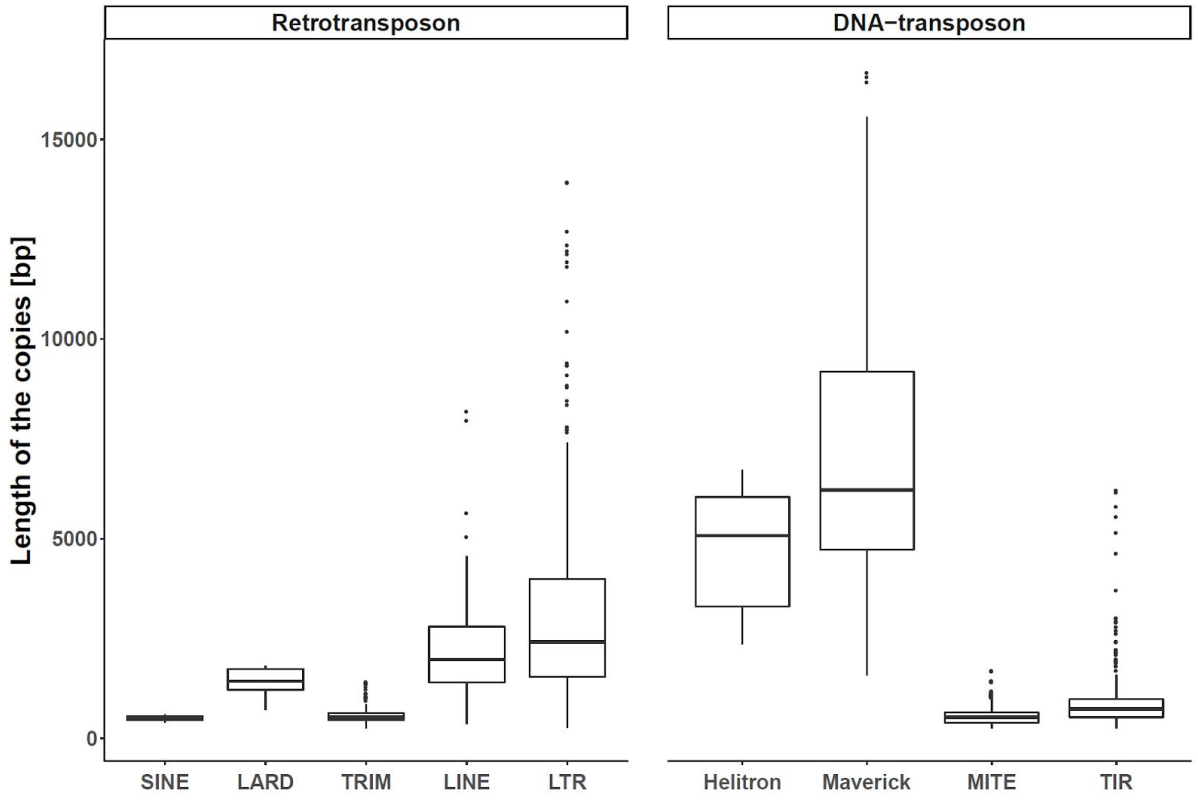


Fig S2: Distribution of TE lengths per-order (*M. incognita*).

Box plots per order of the distribution of the canonical TE annotations' length .

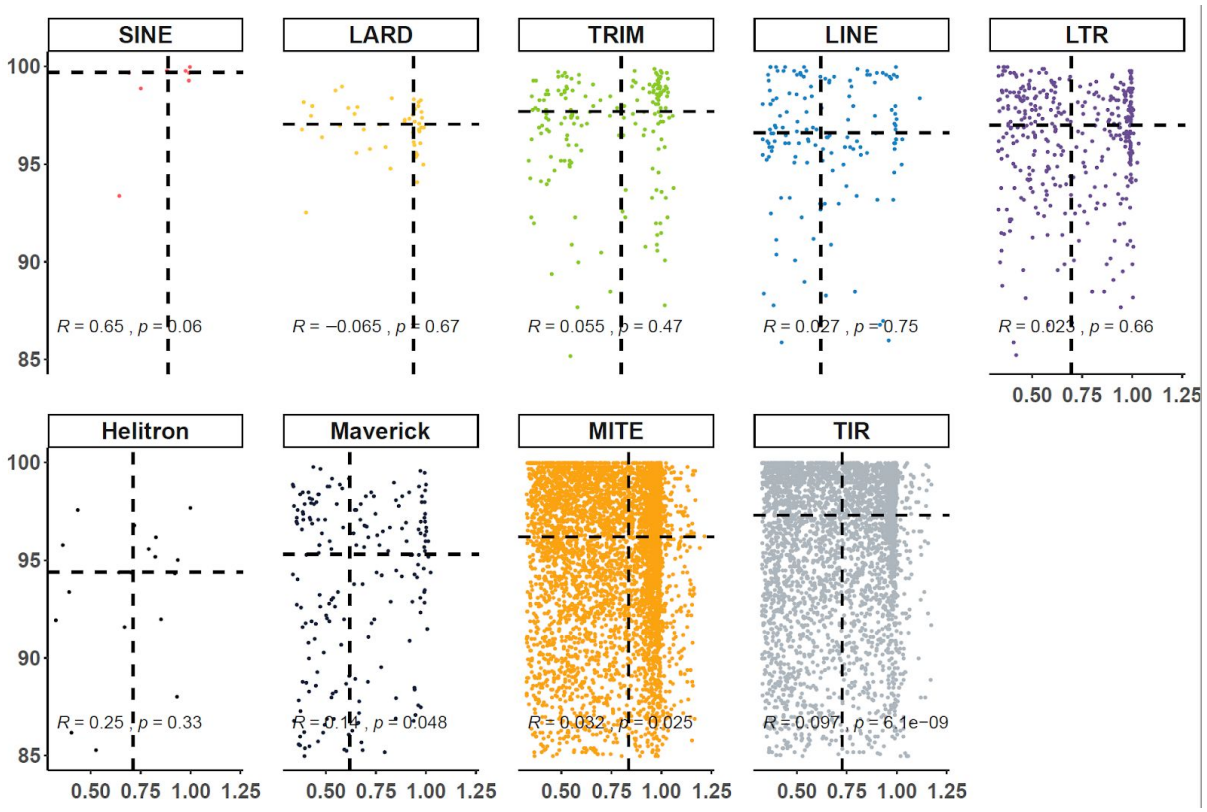
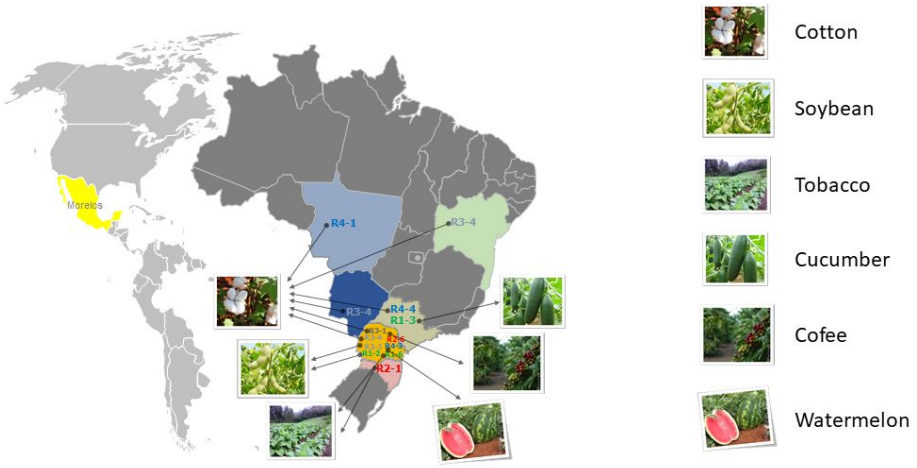


Fig S3: *M. incognita* per TE copy % identity with its consensus in function of the proportion of consensus covered.

Data are splitted in panels according to TE orders. For each panel, Y-axis represents the percentage of identity a copy shares with its consensus. X-axis represents the coverage of the TE consensus (proportion). Coverage values > 100% correspond to cases for which the copy includes a nested sequence regarding the TE consensus sequence (other TE, repeats, other). Each point represents a TE locus (*i.e.* a TE copy). Dashed lines represent the per order median value of both the identity percentage (horizontal line) and the proportion of coverage (vertical line). R value represents the correlation coefficient (Pearson) computed for each order, and p is the associated p-value.

Table S5: canonical TE annotations with putative transposition machinery

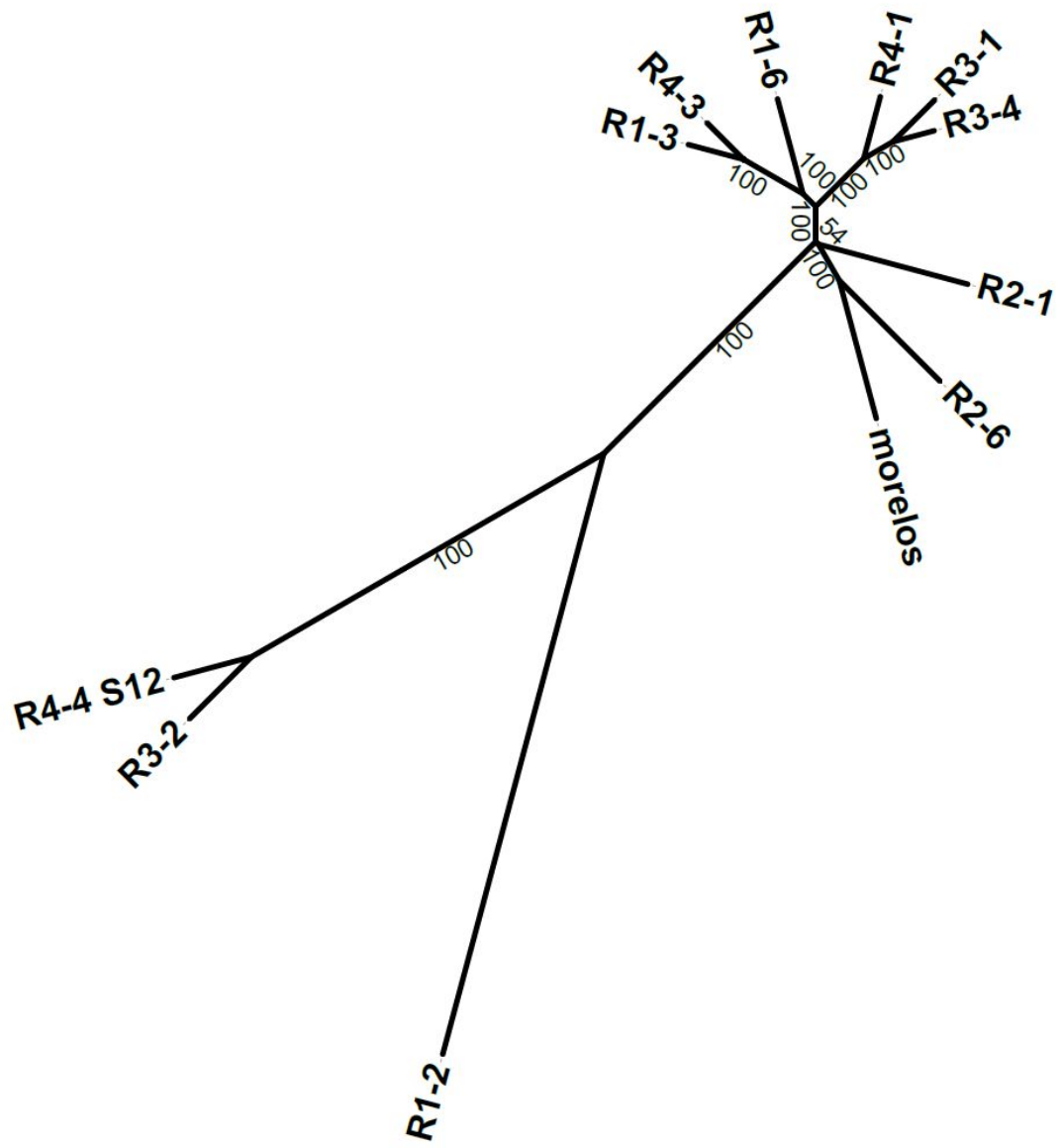
	autonomous (*) / non-autonomous (**) orders	nb. of annotations with putative transposition machinery	nb. of annotations with substantially expressed putative transposition machinery
retro - transposon	SINE (**)	0	0
	LARD (**)	0	0
	TRIM (**)	0	0
	LINE (*)	54	26
	LTR (*)	147	45
DNA - transposon	Helitron (*)	17	3
	Maverick (*)	63	26
	MITE (**)	0	0
	TIR (*)	30	6



Adapted from Evolutionary Applications, Volume: 13, Issue: 2, Pages: 442-457, First published: 19 October 2019, DOI: (10.1111/eva.12881)

Fig S4: Isolates geographical distribution and host plants.

American continent map showing the geographical distribution for all isolates used in the study. Expanded map of Brazil shows the states where the 11 isolates sequenced in (Koutsovoulos et al. 2020) were collected. Each state is highlighted with a different colour. The crops from which the samples were isolated are illustrated by photographs, which are pointed by arrows coming from the name of the respective isolate.


Tree scale: 0.01
R1-6
R4-1
R4-3
R3-1
R1-3
R3-4
100
100
100
100
100
54
100
R2-1
100
R2-6
morelos
100
R4-4 S12
R3-2
R1-2

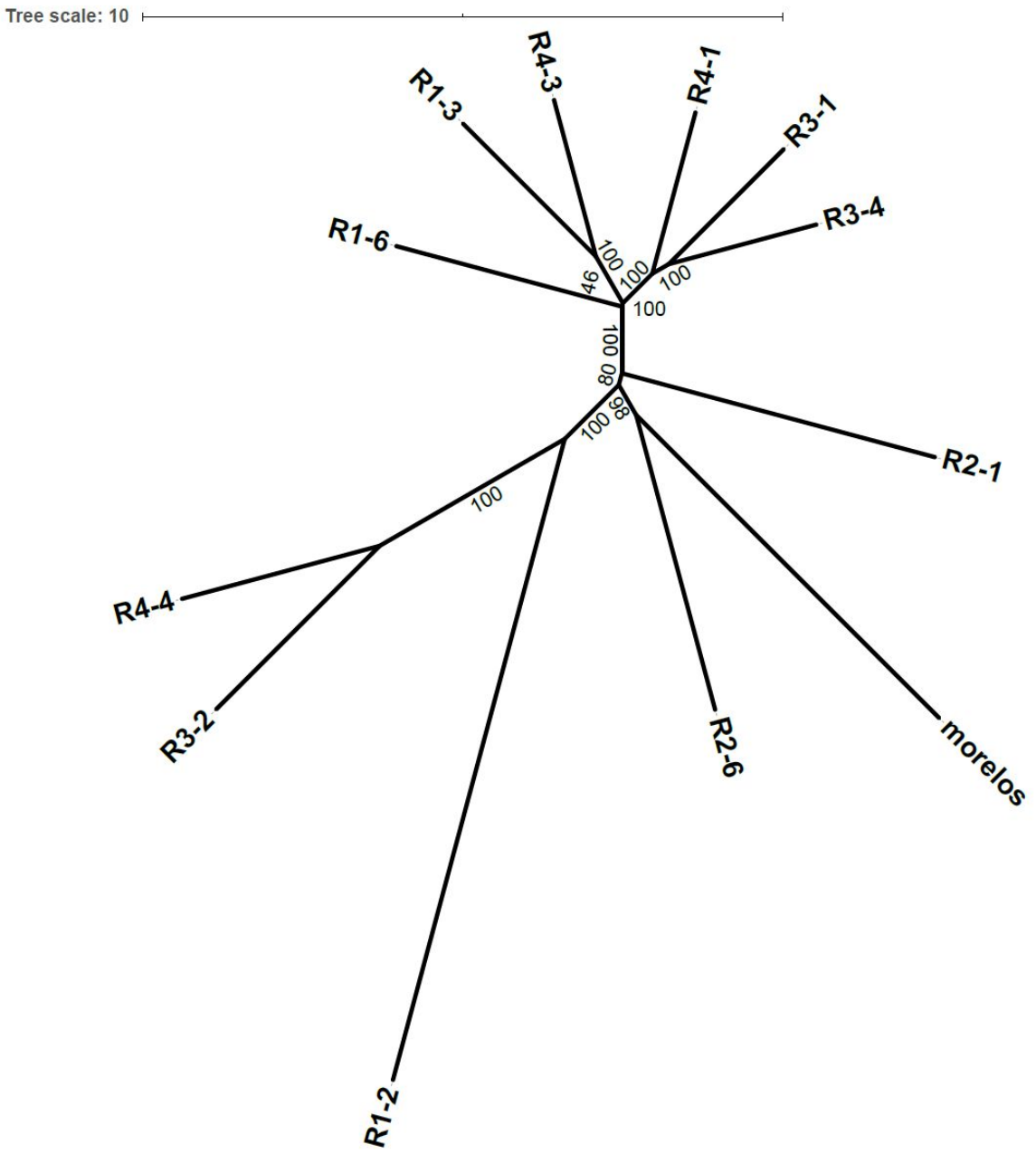


Fig S5: Phylogenetic tree for *M. incognita* isolates with branch length.

Top Phylogenetic tree is based on SNP present in coding sequences. Maximum Likelihood (ML) tree reconstruction. Bottom Phylogenetic tree based on TE-frequencies euclidean distances between isolates. Neighbor-Joining (NJ) tree reconstruction. For both trees, the branch length is displayed. The scale is displayed in the top-left corner of each sub-figure. Values in black (bootstrap values) represent nodes support values, 100 being the maximal value.

Table S6: TE repartition per orders in the reference annotation and between polymorphisms types.

Ref-annotation line represents the per-order number of elements in the reference genome annotation. The sum of "non-polymorphic ref." and "polymorphic-ref" is not equal to the number of reference annotations due to filtering criteria. See sup. Fig S8 for detailed explanations.

	SINE	LARD	TRIM	LINE	LTR	Helitron	Maverick	MITE	TIR
ref-annotations	9	45	174	145	373	18	189	5085	3595
non-polymorphic ref-annotations	8	35	154	128	322	16	179	3602	2657
polymorphic ref-annotations	1	6	14	13	37	1	7	1194	818
neo-insertions	0	0	0	10	11	0	0	192	74
extra-detection	0	0	4	6	12	2	16	97	69

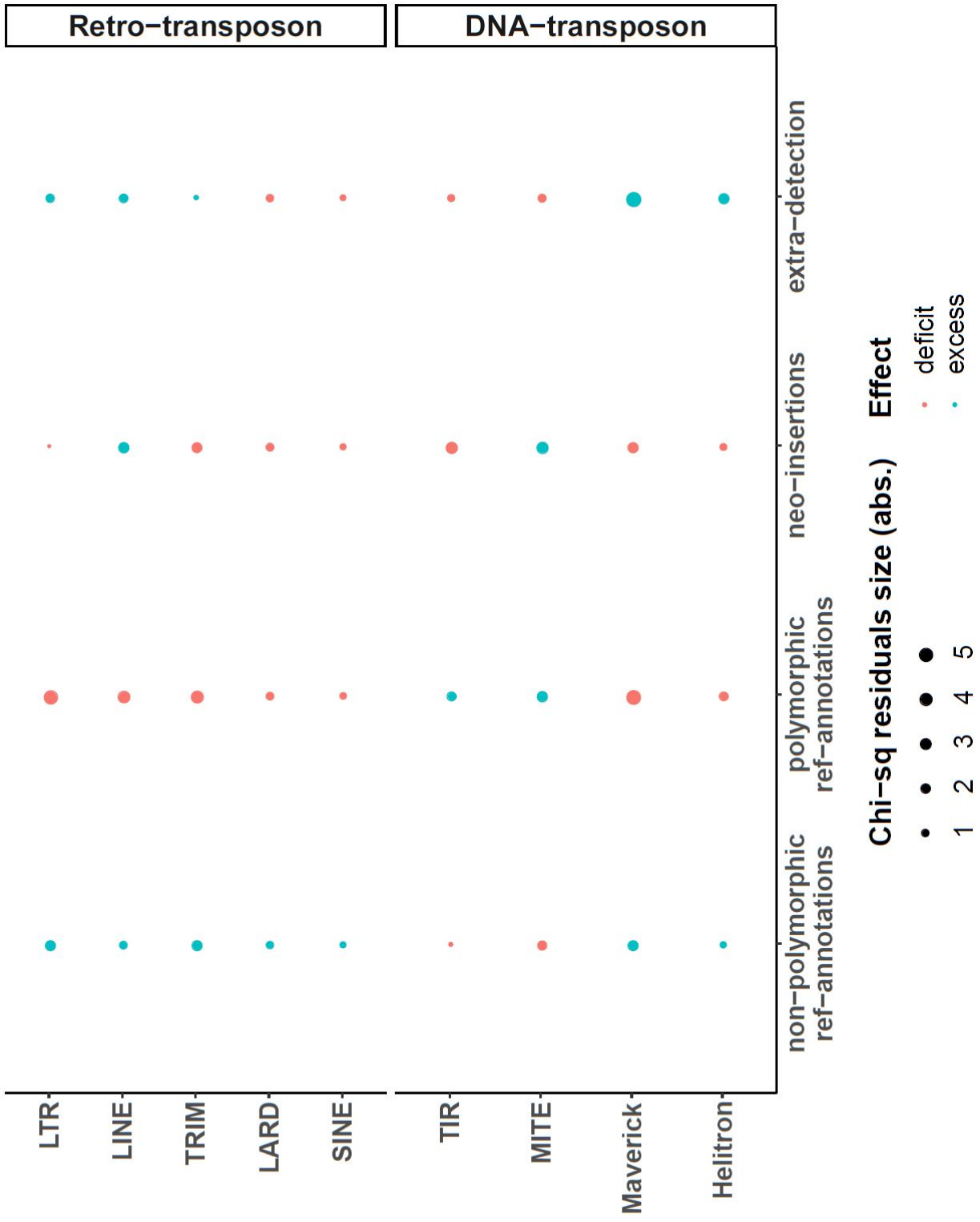


Fig S6: Relative abundance of the TE copies (count) per polymorphism types and TE-orders.

Each point represents a chi-square residual value. Chi-square residuals are the distance from the expected distribution under the homogeneity hypothesis. They are used here as a proxy to estimate the relative abundance (count) per polymorphism type and TE order and

how each combination differs from the expected distribution. For each point, the wider is the surface, the higher is the distance from the expectation. Red points represent a deficit compared to the expectation while the blue points represent an excess.

Table S7: Number of HCPTEs copies per-consensus.

consensus	order	nb. of HCPTEs copies
DTX-comp_mincV3XDN-B-R1459-Map20	TIR	8
DTX-incomp_mincV3XDN-B-R11531-Map10	TIR	2
DTX-incomp_mincV3XDN-B-R271-Map10	TIR	1
DTX-incomp_mincV3XDN-B-R3892-Map20	TIR	1
DXX-MITE_mincV3XDN-B-G1048-Map15	MITE	1
DXX-MITE_mincV3XDN-B-G305-Map9	MITE	1
DXX-MITE_mincV3XDN-B-R14125-Map7	MITE	1
DXX-MITE_mincV3XDN-B-R306-Map20	MITE	10
DXX-MITE_mincV3XDN-B-R321-Map20	MITE	1
DXX-MITE_mincV3XDN-B-R3266-Map20	MITE	1
DXX-MITE_mincV3XDN-B-R3611-Map9	MITE	4
RIX-comp_mincV3XDN-B-R6875-Map20_reversed	LINE	1
RIX-incomp_mincV3XDN-B-R4613-Map9	LINE	1

Table S8: Orthologs to genes potentially impacted by HCPTEs.

Entries with bold font are *M. incognita*'s genes with orthologs in other *Meloidogyne species* or other Plant Parasitic Nematode (PPN) genus only. Entries are sorted by gene name. Genes highlighted in yellow are the genes potentially impacted by HCPTEs which have been selected for experimental validation.

Gene	Wormbase gene trees orthologs	Which species	Tree URL
Minc3s00005g00347	154	in many nematodes and other animals	https://parasite.wormbase.org/Meloidogyne_incognita_prjeb8714/Gene/Compara_Tree?g=Minc3s00005g00347;r=FXSY01000005.1:267231-279946;t=Minc3s00005g00347
Minc3s00005g00348	168	in many nematodes and other animals	https://parasite.wormbase.org/Meloidogyne_incognita_prjeb8714/Gene/Compara_Tree?g=Minc3s00005g00348;r=FXSY01000005.1:271509-282281;t=Minc3s00005g00348
Minc3s00026g01668	14	**Meloidogyne-specific: incognita, arenaria, javanica, floridensis**	https://parasite.wormbase.org/Meloidogyne_incognita_prjeb8714/Gene/Compara_Tree?g=Minc3s00026g01668;r=FXSY01000026.1:125149-126932;t=Minc3s00026g01668;collapse=
Minc3s00137g05752	122	in many nematodes and other animals	https://parasite.wormbase.org/Meloidogyne_incognita_prjeb8714/Gene/Compara_Tree?g=Minc3s00137g05752;r=FXSY01000137.1:70079-73404;t=Minc3s00137g05752
Minc3s00157g06330	135	in many nematodes and other animals	https://parasite.wormbase.org/Meloidogyne_incognita_prjeb8714/Gene/Compara_Tree?g=Minc3s00157g06330;r=FXSY01000157.1:83470-88312;t=Minc3s00157g06330
Minc3s00199g07364	203	in many nematodes and other animals	https://parasite.wormbase.org/Meloidogyne_incognita_prjeb8714/Gene/Compara_Tree?g=Minc3s00199g07364;r=FXSY01000199.1:14729-17937;t=Minc3s00199g07364
Minc3s00199g07365	149	nematode specific but many nematodes	https://parasite.wormbase.org/Meloidogyne_incognita_prjeb8714/Gene/Compara_Tree?g=Minc3s00199g07365;r=FXSY01000199.1:18300-23780;t=Minc3s00199g07365
Minc3s00201g07425	188	in many nematodes and other animals	https://parasite.wormbase.org/Meloidogyne_incognita_prjeb8714/Gene/Compara_Tree?g=Minc3s00201g07425;r=FXSY01000201.1:30179-31671;t=Minc3s00201g07425
Minc3s00201g07426		tRNA (non-coding), widely conserved in nematodes	
Minc3s00201g07427	5	nematode specific, mainly Meloidogyne but also Chromadorea and Dirofilaria	https://parasite.wormbase.org/Meloidogyne_incognita_prjeb8714/Gene/Compara_Tree?g=Minc3s00201g07427;r=FXSY01000201.1:31822-32328;t=Minc3s00201g07427

Minc3s00301g09724	129	in many nematodes and other animals	https://parasite.wormbase.org/Meloidogyne_incognita_prjeb8714/Gene/Compara_Tree?g=Minc3s00301g09724;r=FXSY01000301.1:27845-35780;t=Minc3s00301g09724
Minc3s00450g12515	5	**Meloidogyne-specific: incognita, arenaria.**	https://parasite.wormbase.org/Meloidogyne_incognita_prjeb8714/Gene/Compara_Tree?g=Minc3s00450g12515;r=FXSY01000450.1:51949-52954;t=Minc3s00450g12515;collapse=
Minc3s00621g15225	9	**Meloidogyne-specific: incognita, arenaria, javanica, floridensis, enterolobii, hapla, graminicola.**	https://parasite.wormbase.org/Meloidogyne_incognita_prjeb8714/Gene/Compara_Tree?g=Minc3s00621g15225;r=FXSY01000621.1:38374-38735;t=Minc3s00621g15225
Minc3s00667g15847	17	nematode specific, all Plant Parasitic Nematodes (PPN) except A. nanus	https://parasite.wormbase.org/Meloidogyne_incognita_prjeb8714/Gene/Compara_Tree?g=Minc3s00667g15847;r=FXSY01000667.1:10892-13619;t=Minc3s00667g15847
Minc3s00751g16867	13	**Meloidogyne-specific: incognita, arenaria, javanica, floridensis, enterolobii.**	https://parasite.wormbase.org/Meloidogyne_incognita_prjeb8714/Gene/Compara_Tree?g=Minc3s00751g16867;r=FXSY01000751.1:15531-16499;t=Minc3s00751g16867
Minc3s00905g18730	251	in many nematodes and other animals	https://parasite.wormbase.org/Meloidogyne_incognita_prjeb8714/Gene/Compara_Tree?g=Minc3s00905g18730;r=FXSY01000905.1:1630-5121;t=Minc3s00905g18730
Minc3s00905g18731	5	**Meloidogyne-specific: incognita, arenaria, javanica, floridensis, enterolobii.**	https://parasite.wormbase.org/Meloidogyne_incognita_prjeb8714/Gene/Compara_Tree?g=Minc3s00905g18731;r=FXSY01000905.1:5454-6420;t=Minc3s00905g18731
Minc3s00909g18773	14	**Meloidogyne-specific: incognita, arenaria, javanica, floridensis, enterolobii, graminicola.**	https://parasite.wormbase.org/Meloidogyne_incognita_prjeb8714/Gene/Compara_Tree?g=Minc3s00909g18773;r=FXSY01000909.1:23309-24625;t=Minc3s00909g18773
Minc3s00965g19365	160	in many nematodes and other animals	https://parasite.wormbase.org/Meloidogyne_incognita_prjeb8714/Gene/Compara_Tree?g=Minc3s00965g19365;r=FXSY01000965.1:6357-15984;t=Minc3s00965g19365
Minc3s00988g19605	3	**Meloidogyne-specific: incognita, arenaria, javanica.**	https://parasite.wormbase.org/Meloidogyne_incognita_prjeb8714/Gene/Compara_Tree?g=Minc3s00988g19605;r=FXSY01000988.1:15653-17968;t=Minc3s00988g19605
Minc3s01127g20975	4	**Meloidogyne-specific: incognita, arenaria, javanica, floridensis, hapla.**	https://parasite.wormbase.org/Meloidogyne_incognita_prjeb8714/Gene/Compara_Tree?g=Minc3s01127g20975;r=FXSY01001127.1:7481-15500;t=Minc3s01127g20975
Minc3s01138g21099	6	**Meloidogyne-specific: incognita, arenaria, javanica, floridensis, enterolobii, hapla, graminicola.**	https://parasite.wormbase.org/Meloidogyne_incognita_prjeb8714/Gene/Compara_Tree?g=Minc3s01138g21099;r=FXSY01001138.1:37292-39709;t=Minc3s01138g21099

Minc3s01318g22714	10	**PPN-specific: i) Meloidogyne: incognita, arenaria, javanica, floridensis, enterolobii, graminicola; ii) Globobodera: rostochiensis; iii) Ditylenchus: destructor**	https://parasite.wormbase.org/Meloidogyne_incognita_prjeb8714/Gene/Compara_Tree?g=Minc3s01318g22714;r=FXSY01001318.1:1931-3523;t=Minc3s01318g22714
Minc3s01455g23950	0	no gene tree at all: gene specific to Meloidogyne incognita	
Minc3s01827g26567	3	**Meloidogyne-specific: incognita, arenaria, javanica, floridensis, enterolobii.**	https://parasite.wormbase.org/Meloidogyne_incognita_prjeb8714/Gene/Compara_Tree?g=Minc3s01827g26567;r=FXSY01001827.1:2587-2859;t=Minc3s01827g26567;collapse=9293401
Minc3s02496g30324	170	in many nematodes and other animals	https://parasite.wormbase.org/Meloidogyne_incognita_prjeb8714/Gene/Compara_Tree?g=Minc3s02496g30324;r=FXSY01002496.1:10099-17217;t=Minc3s02496g30324
Minc3s03567g34213	78	Present in many animals then only Meloidogyne	https://parasite.wormbase.org/Meloidogyne_incognita_prjeb8714/Gene/Compara_Tree?g=Minc3s03567g34213;r=FXSY01003567.1:9085-10645;t=Minc3s03567g34213;collapse=14989368,14989316,14989313,14989310,14988329

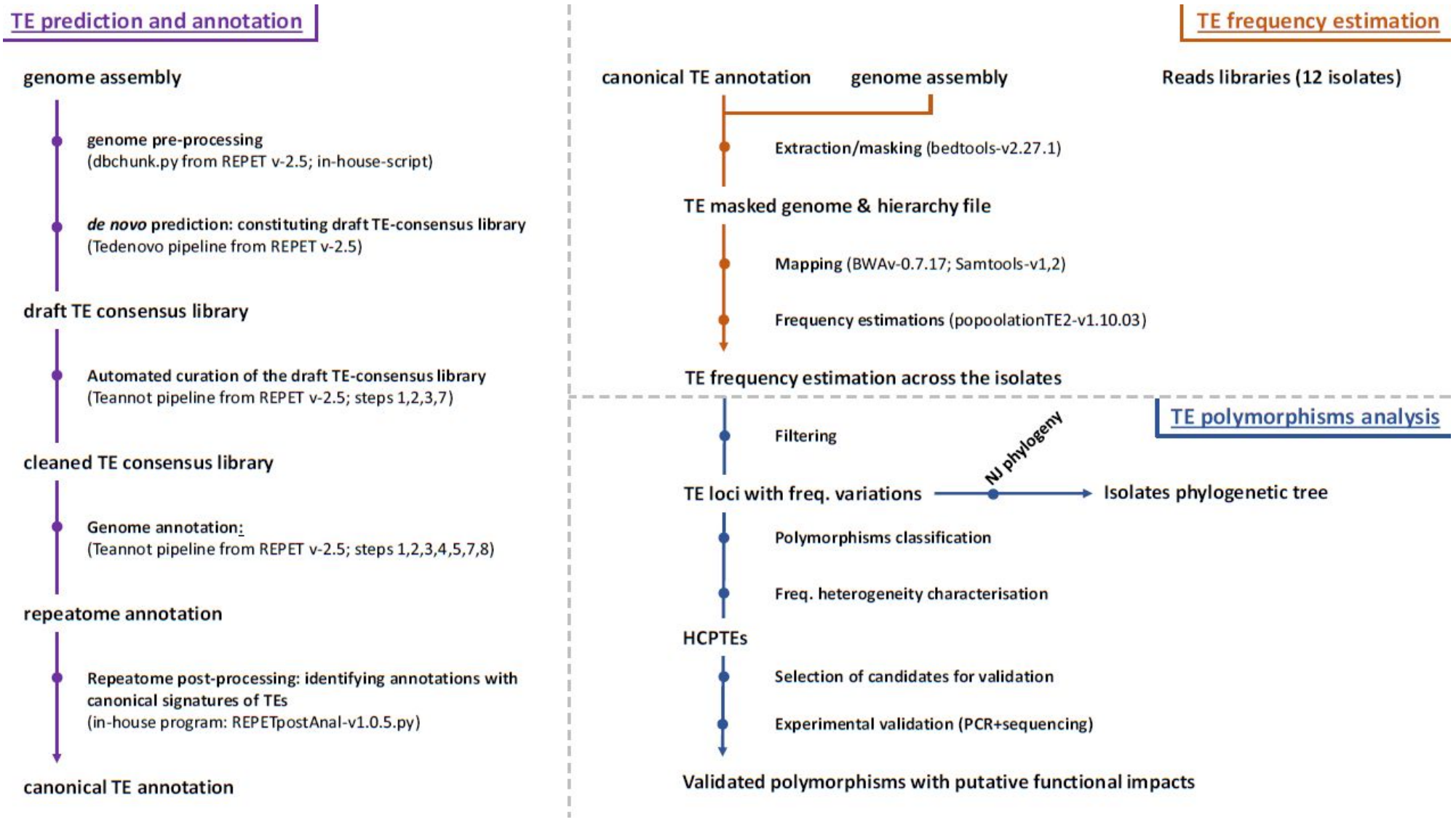

TE prediction and annotation
genome assembly
genome pre-processing
(dbchunk.py from REPET v-2.5; in-house-script)
de novo prediction: constituting draft TE-consensus library
(Tedenovo pipeline from REPET v-2.5)
draft TE consensus library
Automated curation of the draft TE-consensus library
(Teannot pipeline from REPET v-2.5; steps 1,2,3,7)
cleaned TE consensus library
Genome annotation:
(Teannot pipeline from REPET v-2.5; steps 1,2,3,4,5,7,8)
repeatome annotation
Repeatome post-processing: identifying annotations with canonical signatures of TEs
(in-house program: REPETpostAnal-v1.0.5.py)
canonical TE annotation
TE frequency estimation
canonical TE annotation
genome assembly
Reads libraries (12 isolates)
Extraction/masking (bedtools-v2.27.1)
TE masked genome & hierarchy file
Mapping (BWAv-0.7.17; Samtools-v1,2)
Frequency estimations (popoolationTE2-v1.10.03)
TE frequency estimation across the isolates
TE polymorphisms analysis
Filtering
NJ phylogeny
TE loci with freq. variations
Isolates phylogenetic tree
Polymorphisms classification
Freq. heterogeneity characterisation
HCPTEs
Selection of candidates for validation
Experimental validation (PCR+sequencing)
Validated polymorphisms with putative functional impacts

Fig S7: workflow overview.

The current analysis encompasses 3 pipelines: the TE prediction and annotation, the TE frequency estimation, and the TE polymorphisms analysis. Each step's workflow is represented in a separated panel. Each step of each sub-pipeline is explained in detail in Methods. All the scripts are available in (Kozlowski 2020). "Polymorphisms classification" and "Freq. heterogeneity characterisation" steps of the TE-polymorphism pipeline are detailed as a decision tree in Fig S8

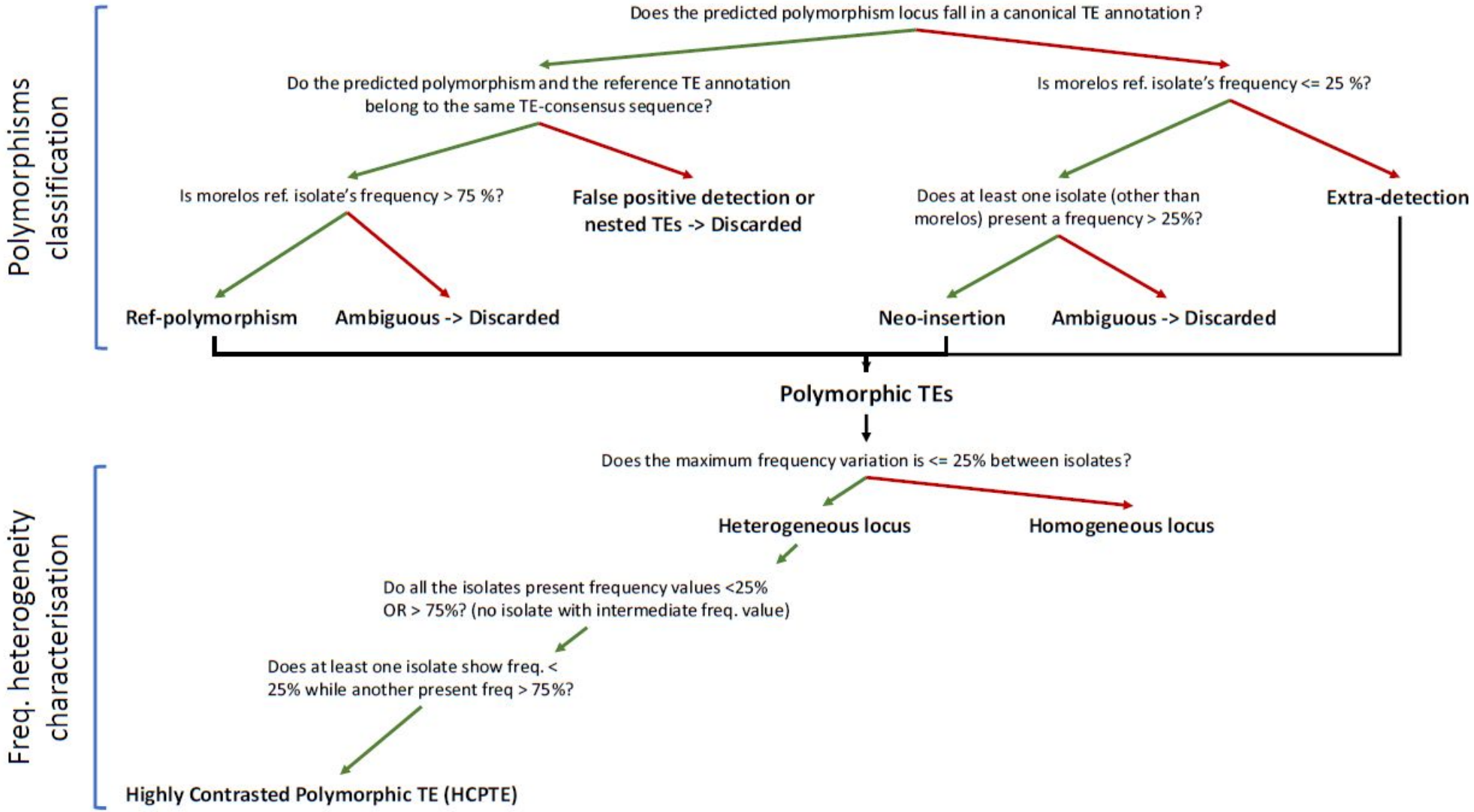

Polymorphisms classification
Does the predicted polymorphism locus fall in a canonical TE annotation ?
Do the predicted polymorphism and the reference TE annotation belong to the same TE-consensus sequence?
Is morelos ref. isolate's frequency <= 25 %?
Is morelos ref. isolate's frequency > 75 %?
False positive detection or nested TEs -> Discarded
Does at least one isolate (other than morelos) present a frequency > 25%?
Extra-detection
Ref-polymorphism
Ambiguous -> Discarded
Neo-insertion
Ambiguous -> Discarded
Polymorphic TEs
Does the maximum frequency variation is <= 25% between isolates?
Heterogeneous locus
Homogeneous locus
Freq. heterogeneity characterisation
Do all the isolates present frequency values <25% OR > 75%? (no isolate with intermediate freq. value)
Does at least one isolate show freq. < 25% while another present freq > 75%?
Highly Contrasted Polymorphic TE (HCPTE)

FIG S8: decision trees for polymorphisms classification and frequency heterogeneity characterisation.

This figure details as a decision tree the "Polymorphisms classification" and "Freq. heterogeneity characterisation" steps of the TE-polymorphism pipeline from the Fig S7. Green arrows represent a positive answer. The red ones represent a negative answer.

Table S9: pairwise blastn of locus 1 sequencing results.

insertion predicted	subject	query	sequence	%identity	query cover (%)	e-value	average % identity
N	morelos	R2-1	F	98.68	99	6,00E-75	99.34
			R	100	91	2,00E-73	
N	morelos	R2-6	F	100	93	8,00E-73	95.3
			R	90.6	89	4,00E-52	
N	R2-6	R2-1	F	98.64	96	1,00E-71	94.45
			R	90.26	92	4,00E-52	
Y	R1-2	R4-4	F	92.73	72	0,00E+00	96.25
			R	99.77	99	0,00E+00	
Y	R3-2	R1-2	F	98.52	98	0,00E+00	98.88
			R	99.24	89	0,00E+00	
Y	R4-4	R3-2	F	92.16	93	0,00E+00	95.7
			R	99.24	98	0,00E+00	

Table S10: Reads libraries accession numbers & statistics

Lib. name	acess. nb. (SRA)	nb. reads (P-E)	read length (bp)	% GC
morelos	ERS1696677	76077411	2*150	28
R1-2	SRX4373671	76359269	2*150	29
R1-3	SRX4373672	75542522	2*150	28
R1-6	SRX4373673	75033425	2*150	28
R2-1	SRX4373674	75065658	2*150	29
R2-6	SRX4373675	75300726	2*150	29
R3-1	SRX4373676	74468408	2*150	30
R3-2	SRX4373677	74671928	2*150	28
R3-4	SRX4373678	74620706	2*150	28
R4-1	SRX4373679	75063890	2*150	28
R4-3	SRX4373680	75235737	2*150	29
R4-4	SRX4373681	74987959	2*150	28

Table S11: PCR primers targeting 5 candidate locus for TE insertion

Primer Pair	Sequence	Amplicon size without insertion (bp)	Amplicon size with insertion (bp)
Locus1-F	CTTAGGTTTTTGACTGCGTCTGCCAT	180	973
Locus1-R	CAGATGCATTGCGGTGACGTTCTT		
Locus2-F	GGGGGTCAGATTACCCTCTATTATGGCA	761	1870
Locus2-R	CCTCTCCCATCACTCTCACAACCCA		
Locus3-F	CCGTCGGCGGGATCCCTGATATAAA	690	1814
Locus3-R	TTATCGGTTTCAACCCCGACCGAAC		
Locus4-F	GGTGGTGTTGTTGCTGGAATTACTAACC	981	1781
Locus4-R	GACAAACGTTGGAGCACGTTATGCTCG		
Locus5-F	GGAACAGTCAGCGGTGTCGGAAATC	1005	2080
Locus5-R	GTGTATGCTTCAGAACCCAGACGGGGA		
actin-F (ctrl +)	AAGATGGATGAAGAGGTAGCCGCCC	-	1667
actin-R (ctrl +)	ACTCTTGCTTGCTGATCCACCTTGA		

References

Koutsovoulos GD, Marques E, Arguel M-J, Duret L, Machado ACZ, Carneiro RMDG, Kozlowski DK, Bailly-Bechet M, Castagnone-Sereno P, Albuquerque EVS, et al. 2020. Population genomics supports clonal reproduction and multiple independent gains and losses of parasitic abilities in the most devastating nematode pest. *Evol. Appl.* 13:442–457.

Kozlowski D. 2020. TE polymorphisms detection and analysis with PopoolationTE2. *Portail Data INRAE* [Internet]. Available from: https://doi.org/10.15454/EWJCT8

XI – Bibliographie

Abad, P., Gouzy, J., Aury, J.-M., Castagnone-Sereno, P., Danchin, E.G.J., Deleury, E., Perfus-Barbeoch, L., Anthouard, V., Artiguenave, F., Blok, V.C., Caillaud, M.-C., Coutinho, P.M., Dasilva, C., De Luca, F., Deau, F., Esquibet, M., Flutre, T., Goldstone, J.V., Hamamouch, N., Hewezi, T., Jaillon, O., Jubin, C., Leonetti, P., Magliano, M., Maier, T.R., Markov, G.V., McVeigh, P., Pesole, G., Poulain, J., Robinson-Rechavi, M., Sallet, E., Ségurens, B., Steinbach, D., Tytgat, T., Ugarte, E., van Ghelder, C., Veronico, P., Baum, T.J., Blaxter, M., Bleve-Zacheo, T., Davis, E.L., Ewbank, J.J., Favery, B., Grenier, E., Henrissat, B., Jones, J.T., Laudet, V., Maule, A.G., Quesneville, H., Rosso, M.-N., Schiex, T., Smant, G., Weissenbach, J., Wincker, P., 2008. Genome sequence of the metazoan plant-parasitic nematode Meloidogyne incognita. Nat. Biotechnol. 26, 909–915. https://doi.org/10.1038/nbt.1482

Abascal, F., Tress, M.L., Valencia, A., 2015. Alternative splicing and co-option of transposable elements: the case of TMPO/LAP2α and ZNF451 in mammals. Bioinformatics 31, 2257–2261. https://doi.org/10.1093/bioinformatics/btv132

Acuna, R., Padilla, B.E., Florez-Ramos, C.P., Rubio, J.D., Herrera, J.C., Benavides, P., Lee, S.-J., Yeats, T.H., Egan, A.N., Doyle, J.J., Rose, J.K.C., 2012. Adaptive horizontal transfer of a bacterial gene to an invasive insect pest of coffee. Proc. Natl. Acad. Sci. 109, 4197–4202. https://doi.org/10.1073/pnas.1121190109

Álvarez-Ortega, S., Brito, J.A., Subbotin, Sergei.A., 2019. Multigene phylogeny of root-knot nematodes and molecular characterization of Meloidogyne nataliei Golden, Rose & Bird, 1981 (Nematoda: Tylenchida). Sci. Rep. 9, 11788. https://doi.org/10.1038/s41598-019-48195-0

Andrews S., 2010. FastQC: a quality control tool for high throughput sequence data.

Arkhipova, I.R., 2017. Using bioinformatic and phylogenetic approaches to classify transposable elements and understand their complex evolutionary histories. Mob. DNA 8, 19. https://doi.org/10.1186/s13100-017-0103-2

Ashley, J., Cordy, B., Lucia, D., Fradkin, L.G., Budnik, V., Thomson, T., 2018. Retrovirus-like Gag Protein Arc1 Binds RNA and Traffics across Synaptic Boutons. Cell 172, 262-274.e11. https://doi.org/10.1016/j.cell.2017.12.022

Bao, W., Kojima, K.K., Kohany, O., 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob. DNA 6, 11. https://doi.org/10.1186/s13100-015-0041-9

Bao, Z., 2002. Automated De Novo Identification of Repeat Sequence Families in Sequenced Genomes. Genome Res. 12, 1269–1276. https://doi.org/10.1101/gr.88502

Barbary, A., Djian-Caporalino, C., Palloix, A., Castagnone-Sereno, P., 2015. Host genetic resistance to root-knot nematodes, *Meloidogyne* spp., in Solanaceae: from genes to the field: Host genetic resistance to root-knot nematodesin Solanaceae. Pest Manag. Sci. 71, 1591–1598. https://doi.org/10.1002/ps.4091

Bar-Zvi, D., Lupo, O., Levy, A.A., Barkai, N., 2017. Hybrid vigor: The best of both parents, or a genomic clash? Curr. Opin. Syst. Biol. 6, 22–27. https://doi.org/10.1016/j.coisb.2017.08.004

Bast, J., Jaron, K.S., Schuseil, D., Roze, D., Schwander, T., 2019. Asexual reproduction reduces transposable element load in experimental yeast populations. eLife 8, e48548. https://doi.org/10.7554/eLife.48548

Beare, P.A., Unsworth, N., Andoh, M., Voth, D.E., Omsland, A., Gilk, S.D., Williams, K.P., Sobral, B.W., Kupko, J.J., Porcella, S.F., Samuel, J.E., Heinzen, R.A., 2009. Comparative Genomics Reveal Extensive Transposon-Mediated Genomic Plasticity and Diversity among Potential Effector Proteins within the Genus Coxiella. Infect. Immun. 77, 642–656. https://doi.org/10.1128/IAI.01141-08

Belyayev, A., 2014. Bursts of transposable elements as an evolutionary driving force. J. Evol. Biol. 27, 2573–2584. https://doi.org/10.1111/jeb.12513

Bengtsson, B.O., 2009. Asex and Evolution: A Very Large-Scale Overview, in: Schön, I., Martens, K., Dijk, P. (Eds.), Lost Sex. Springer Netherlands, Dordrecht, pp. 1–19. https://doi.org/10.1007/978-90-481-2770-2_1

Benson, G., 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27, 573–580. https://doi.org/10.1093/nar/27.2.573

Bessereau, J.-L., 2006. Transposons in C. elegans. WormBook. https://doi.org/10.1895/wormbook.1.70.1

Bhattacharyya, M.K., Smith, A.M., Ellis, T.H.N., Hedley, C., Martin, C., 1990. The wrinkled-seed character of pea described by Mendel is caused by a transposon-like insertion in a gene encoding starch-branching enzyme. Cell 60, 115–122. https://doi.org/10.1016/0092-8674(90)90721-P

Blanc-Mathieu, R., Perfus-Barbeoch, L., Aury, J.-M., Da Rocha, M., Gouzy, J., Sallet, E., Martin-Jimenez, C., Bailly-Bechet, M., Castagnone-Sereno, P., Flot, J.-F., Kozlowski, D.K., Cazareth, J., Couloux, A., Da Silva, C., Guy, J., Kim-Jo, Y.-J., Rancurel, C., Schiex, T., Abad, P., Wincker, P., Danchin, E.G.J., 2017. Hybridization and polyploidy enable genomic plasticity without sex in the most devastating plant-parasitic nematodes. PLOS Genet. 13, e1006777. https://doi.org/10.1371/journal.pgen.1006777

Blumenstiel, J.P., 2019. Birth, School, Work, Death, and Resurrection: The Life Stages and Dynamics of Transposable Element Proliferation. Genes 10, 336. https://doi.org/10.3390/genes10050336

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bonchev, G., Parisod, C., 2013. Transposable elements and microevolutionary changes in natural populations. Mol. Ecol. Resour. 13, 765–775. https://doi.org/10.1111/1755-0998.12133

Bourgeois, Y., Boissinot, S., 2019. On the Population Dynamics of Junk: A Review on the Population Genomics of Transposable Elements. Genes 10, 419. https://doi.org/10.3390/genes10060419

Bourque, G., Burns, K.H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H.L., Macfarlan, T.S., Mager, D.L., Feschotte, C., 2018. Ten things you should know about transposable elements. Genome Biol. 19, 199. https://doi.org/10.1186/s13059-018-1577-z

Butelli, E., Licciardello, C., Zhang, Y., Liu, J., Mackay, S., Bailey, P., Reforgiato-Recupero, G., Martin, C., 2012. Retrotransposons Control Fruit-Specific, Cold-Dependent Accumulation of

Anthocyanins in Blood Oranges. Plant Cell 24, 1242–1255. https://doi.org/10.1105/tpc.111.095232

Casacuberta, E., 2017. Drosophila: Retrotransposons Making up Telomeres. Viruses 9, 192. https://doi.org/10.3390/v9070192

Casacuberta, E., González, J., 2013. The impact of transposable elements in environmental adaptation. Mol. Ecol. 22, 1503–1517. https://doi.org/10.1111/mec.12170

Castagnone-Sereno, P., 2006. Genetic variability and adaptive evolution in parthenogenetic root-knot nematodes. Heredity 96, 282–289. https://doi.org/10.1038/sj.hdy.6800794

Castagnone-Sereno, P., Danchin, E.G.J., Perfus-Barbeoch, L., Abad, P., 2013. Diversity and Evolution of Root-Knot Nematodes, Genus *Meloidogyne* : New Insights from the Genomic Era. Annu. Rev. Phytopathol. 51, 203–220. https://doi.org/10.1146/annurev-phyto-082712-102300

Castagnone-Sereno, P., Mulet, K., Danchin, E.G.J., Koutsovoulos, G.D., Karaulic, M., Da Rocha, M., Bailly-Bechet, M., Pratx, L., Perfus-Barbeoch, L., Abad, P., 2019. Gene copy number variations as signatures of adaptive evolution in the parthenogenetic, plant-parasitic nematode *Meloidogyne incognita*. Mol. Ecol. 28, 2559–2572. https://doi.org/10.1111/mec.15095

Castagnone-Sereno, P., Wajnberg, E., Bongiovanni, M., Leroy, F., Dalmasso, A., 1994. Genetic variation inMeloidogyne incognita virulence against the tomatoMi resistance gene: evidence from isofemale line selection studies. Theor. Appl. Genet. 88, 749–753. https://doi.org/10.1007/BF01253980

Chalopin, D., Volff, J.-N., Galiana, D., Anderson, J.L., Schartl, M., 2015. Transposable elements and early evolution of sex chromosomes in fish. Chromosome Res. 23, 545–560. https://doi.org/10.1007/s10577-015-9490-8

Cook, L.M., 2003. The Rise and Fall of the *Carbonaria* Form of the Peppered Moth. Q. Rev. Biol. 78, 399–417. https://doi.org/10.1086/378925

Cordaux, R., Batzer, M.A., 2009. The impact of retrotransposons on human genome evolution. Nat. Rev. Genet. 10, 691–703. https://doi.org/10.1038/nrg2640

Cordaux, R., Gilbert, C., 2017. Evolutionary Significance of Wolbachia-to-Animal Horizontal Gene Transfer: Female Sex Determination and the f Element in the Isopod Armadillidium vulgare. Genes 8, 186. https://doi.org/10.3390/genes8070186

Cuypers, T.D., Hogeweg, P., 2014. A Synergism between Adaptive Effects and Evolvability Drives Whole Genome Duplication to Fixation. PLoS Comput. Biol. 10, e1003547. https://doi.org/10.1371/journal.pcbi.1003547

Danchin, E.G.J., Rosso, M.-N., Vieira, P., de Almeida-Engler, J., Coutinho, P.M., Henrissat, B., Abad, P., 2010. Multiple lateral gene transfers and duplications have promoted plant parasitism ability in nematodes. Proc. Natl. Acad. Sci. 107, 17651–17656. https://doi.org/10.1073/pnas.1008486107

de Meeûs, T., Prugnolle, F., Agnew, P., 2007. Asexual reproduction: Genetics and evolutionary aspects. Cell. Mol. Life Sci. 64, 1355–1372. https://doi.org/10.1007/s00018-007-6515-2

Dechaud, C., Volff, J.-N., Schartl, M., Naville, M., 2019. Sex and the TEs: transposable elements in sexual development and function in animals. Mob. DNA 10, 42. https://doi.org/10.1186/s13100-019-0185-0

Deng, N., Zhou, H., Fan, H., Yuan, Y., 2017. Single nucleotide polymorphisms and cancer susceptibility. Oncotarget 8, 110635–110649. https://doi.org/10.18632/oncotarget.22372

Dennis, M.Y., Eichler, E.E., 2016. Human adaptation and evolution by segmental duplication. Curr. Opin. Genet. Dev. 41, 44–52. https://doi.org/10.1016/j.gde.2016.08.001

Dittrich-Reed, D.R., Fitzpatrick, B.M., 2013. Transgressive Hybrids as Hopeful Monsters. Evol. Biol. 40, 310–315. https://doi.org/10.1007/s11692-012-9209-0

Drezen, J.-M., Gauthier, J., Josse, T., Bézier, A., Herniou, E., Huguet, E., 2017. Foreign DNA acquisition by invertebrate genomes. J. Invertebr. Pathol. 147, 157–168. https://doi.org/10.1016/j.jip.2016.09.004

Edgar, R.C., Myers, E.W., 2005. PILER: identification and classification of genomic repeats. Bioinformatics 21, i152–i158. https://doi.org/10.1093/bioinformatics/bti1003

Ellegren, H., Galtier, N., 2016. Determinants of genetic diversity. Nat. Rev. Genet. 17, 422–433. https://doi.org/10.1038/nrg.2016.58

Elliott, T.A., Gregory, T.R., 2015. Do larger genomes contain more diverse transposable elements? BMC Evol. Biol. 15, 69. https://doi.org/10.1186/s12862-015-0339-8

Ellison, C.E., Bachtrog, D., 2013. Dosage Compensation via Transposable Element Mediated Rewiring of a Regulatory Network. Science 342, 846–850. https://doi.org/10.1126/science.1239552

Ewing, A.D., 2015. Transposable element detection from whole genome sequence data. Mob. DNA 6, 24. https://doi.org/10.1186/s13100-015-0055-3

Eyres, I., Boschetti, C., Crisp, A., Smith, T.P., Fontaneto, D., Tunnacliffe, A., Barraclough, T.G., 2015. Horizontal gene transfer in bdelloid rotifers is ancient, ongoing and more frequent in species from desiccating habitats. BMC Biol. 13, 90. https://doi.org/10.1186/s12915-015-0202-9

Faino, L., Seidl, M.F., Shi-Kunne, X., Pauper, M., van den Berg, G.C.M., Wittenberg, A.H.J., Thomma, B.P.H.J., 2016. Transposons passively and actively contribute to evolution of the two-speed genome of a fungal pathogen. Genome Res. 26, 1091–1100. https://doi.org/10.1101/gr.204974.116

Faust, G.G., Hall, I.M., 2012. YAHA: fast and flexible long-read alignment with optimal breakpoint detection. Bioinforma. Oxf. Engl. 28, 2417–2424. https://doi.org/10.1093/bioinformatics/bts456

Feiner, N., 2016. Accumulation of transposable elements in *Hox* gene clusters during adaptive radiation of *Anolis* lizards. Proc. R. Soc. B Biol. Sci. 283, 20161555. https://doi.org/10.1098/rspb.2016.1555

Feschotte, C., Pritham, E.J., 2007. DNA Transposons and the Evolution of Eukaryotic Genomes. Annu. Rev. Genet. 41, 331–368. https://doi.org/10.1146/annurev.genet.40.110405.090448

Finnegan, D.J., 1989. Eukaryotic transposable elements and genome evolution. Trends Genet. TIG 5, 103–107. https://doi.org/10.1016/0168-9525(89)90039-5

Fitzpatrick, B.M., Shaffer, H.B., 2007. Hybrid vigor between native and introduced salamanders raises new challenges for conservation. Proc. Natl. Acad. Sci. 104, 15793–15798. https://doi.org/10.1073/pnas.0704791104

Flot, J.-F., Hespeels, B., Li, X., Noel, B., Arkhipova, I., Danchin, E.G.J., Hejnol, A., Henrissat, B., Koszul, R., Aury, J.-M., Barbe, V., Barthélémy, R.-M., Bast, J., Bazykin, G.A., Chabrol, O., Couloux, A., Da Rocha, M., Da Silva, C., Gladyshev, E., Gouret, P., Hallatschek, O., Hecox-Lea, B., Labadie, K., Lejeune, B., Piskurek, O., Poulain, J., Rodriguez, F., Ryan, J.F., Vakhrusheva, O.A., Wajnberg, E., Wirth, B., Yushenova, I., Kellis, M., Kondrashov, A.S., Mark Welch, D.B., Pontarotti, P., Weissenbach, J., Wincker, P., Jaillon, O., Van Doninck, K., 2013. Genomic evidence for ameiotic evolution in the bdelloid rotifer Adineta vaga. Nature 500, 453–457. https://doi.org/10.1038/nature12326

Flutre, T., Duprat, E., Feuillet, C., Quesneville, H., 2011. Considering Transposable Element Diversification in De Novo Annotation Approaches. PLoS ONE 6, e16526. https://doi.org/10.1371/journal.pone.0016526

Franchini, L.F., Lopez-Leal, R., Nasif, S., Beati, P., Gelman, D.M., Low, M.J., de Souza, F.J.S., Rubinstein, M., 2011. Convergent evolution of two mammalian neuronal enhancers by sequential exaptation of unrelated retroposons. Proc. Natl. Acad. Sci. 108, 15270–15275. https://doi.org/10.1073/pnas.1104997108

Galis, F., Alphen, J.J.M., 2020. Parthenogenesis and developmental constraints. Evol. Dev. 22, 205–217. https://doi.org/10.1111/ede.12324

Gemmell, P., Hein, J., Katzourakis, A., 2019. The Exaptation of HERV-H: Evolutionary Analyses Reveal the Genomic Features of Highly Transcribed Elements. Front. Immunol. 10, 1339. https://doi.org/10.3389/fimmu.2019.01339

Gilbert, C., Schaack, S., Pace II, J.K., Brindley, P.J., Feschotte, C., 2010. A role for host–parasite interactions in the horizontal transfer of transposons across phyla. Nature 464, 1347–1350. https://doi.org/10.1038/nature08939

Gladyshev, E.A., Meselson, M., Arkhipova, I.R., 2008. Massive Horizontal Gene Transfer in Bdelloid Rotifers. Science 320, 1210–1213. https://doi.org/10.1126/science.1156407

Glémin, S., Galtier, N., 2012. Genome Evolution in Outcrossing Versus Selfing Versus Asexual Species, in: Anisimova, M. (Ed.), Evolutionary Genomics, Methods in Molecular Biology. Humana Press, Totowa, NJ, pp. 311–335. https://doi.org/10.1007/978-1-61779-582-4_11

González, J., Lenkov, K., Lipatov, M., Macpherson, J.M., Petrov, D.A., 2008. High Rate of Recent Transposable Element–Induced Adaptation in Drosophila melanogaster. PLoS Biol. 6, e251. https://doi.org/10.1371/journal.pbio.0060251

Goodier, J.L., Kazazian, H.H., 2008. Retrotransposons Revisited: The Restraint and Rehabilitation of Parasites. Cell 135, 23–35. https://doi.org/10.1016/j.cell.2008.09.022

Goubert, C., Modolo, L., Vieira, C., ValienteMoro, C., Mavingui, P., Boulesteix, M., 2015. De Novo Assembly and Annotation of the Asian Tiger Mosquito (Aedes albopictus) Repeatome with

dnaPipeTE from Raw Genomic Reads and Comparative Analysis with the Yellow Fever Mosquito (Aedes aegypti). Genome Biol. Evol. 7, 1192–1205. https://doi.org/10.1093/gbe/evv050

Gross, S.M., Williamson, V.M., 2011. Tm1: A Mutator/Foldback Transposable Element Family in Root-Knot Nematodes. PLoS ONE 6, e24534. https://doi.org/10.1371/journal.pone.0024534

Haegeman, A., Jones, J.T., Danchin, E.G.J., 2011. Horizontal Gene Transfer in Nematodes: A Catalyst for Plant Parasitism? Mol. Plant-Microbe Interactions® 24, 879–887. https://doi.org/10.1094/MPMI-03-11-0055

Han, K., Lee, J., Meyer, T.J., Remedios, P., Goodwin, L., Batzer, M.A., 2008. L1 recombination-associated deletions generate human genomic variation. Proc. Natl. Acad. Sci. U. S. A. 105, 19366–19371. https://doi.org/10.1073/pnas.0807866105

Hancks, D.C., Kazazian, H.H., 2016. Roles for retrotransposon insertions in human disease. Mob. DNA 7, 9. https://doi.org/10.1186/s13100-016-0065-9

Handoo, Z.A., Nyczepir, A.P., Esmenjaud, D., van der Beek, J.G., Castagnone-Sereno, P., Carta, L.K., Skantar, A.M., Higgins, J.A., 2004. Morphological, Molecular, and Differential-Host Characterization of Meloidogyne floridensis n. sp. (Nematoda: Meloidogynidae), a Root-Knot Nematode Parasitizing Peach in Florida. J. Nematol. 36, 20–35.

Hartman, K.M., Sasser, J.N., 1985. AND PERINEAL-PATTERN MORPHOLOGY. Adv. Treatise Meloidogyne Methodol. 2, 69.

Hoede, C., Arnoux, S., Moisset, M., Chaumier, T., Inizan, O., Jamilloux, V., Quesneville, H., 2014. PASTEC: An Automatic Transposable Element Classification Tool. PLoS ONE 9, e91929. https://doi.org/10.1371/journal.pone.0091929

Hoen, D.R., Bureau, T.E., 2015. Discovery of Novel Genes Derived from Transposable Elements Using Integrative Genomic Analysis. Mol. Biol. Evol. 32, 1487–1506. https://doi.org/10.1093/molbev/msv042

Hoen, D.R., Hickey, G., Bourque, G., Casacuberta, J., Cordaux, R., Feschotte, C., Fiston-Lavier, A.-S., Hua-Van, A., Hubley, R., Kapusta, A., Lerat, E., Maumus, F., Pollock, D.D., Quesneville, H., Smit, A., Wheeler, T.J., Bureau, T.E., Blanchette, M., 2015. A call for benchmarking transposable element annotation methods. Mob. DNA 6, 13. https://doi.org/10.1186/s13100-015-0044-6

Hof, A.E. van't, Campagne, P., Rigden, D.J., Yung, C.J., Lingley, J., Quail, M.A., Hall, N., Darby, A.C., Saccheri, I.J., 2016. The industrial melanism mutation in British peppered moths is a transposable element. Nature 534, 102–105. https://doi.org/10.1038/nature17951

Hotopp, J.C.D., Clark, M.E., Oliveira, D.C.S.G., Foster, J.M., Fischer, P., Torres, M.C.M., Giebel, J.D., Kumar, N., Ishmael, N., Wang, S., Ingram, J., Nene, R.V., Shepard, J., Tomkins, J., Richards, S., Spiro, D.J., Ghedin, E., Slatko, B.E., Tettelin, H., Werren, J.H., 2007. Widespread Lateral Gene Transfer from Intracellular Bacteria to Multicellular Eukaryotes. Science 317, 1753–1756. https://doi.org/10.1126/science.1142490

Huang, S., Tao, X., Yuan, S., Zhang, Yuhang, Li, P., Beilinson, H.A., Zhang, Ya, Yu, W., Pontarotti, P., Escriva, H., Le Petillon, Y., Liu, X., Chen, S., Schatz, D.G., Xu, A., 2016. Discovery of an

Active RAG Transposon Illuminates the Origins of V(D)J Recombination. Cell 166, 102–114. https://doi.org/10.1016/j.cell.2016.05.032

Huang, W., Li, L., Myers, J.R., Marth, G.T., 2012. ART: a next-generation sequencing read simulator. Bioinformatics 28, 593–594. https://doi.org/10.1093/bioinformatics/btr708

Hull, R.M., Cruz, C., Jack, C.V., Houseley, J., 2017. Environmental change drives accelerated adaptation through stimulated copy number variation. PLOS Biol. 15, e2001333. https://doi.org/10.1371/journal.pbio.2001333

Hunter, M.C., Smith, R.G., Schipanski, M.E., Atwood, L.W., Mortensen, D.A., 2017. Agriculture in 2050: Recalibrating Targets for Sustainable Intensification. BioScience 67, 386–391. https://doi.org/10.1093/biosci/bix010

Jepson, S.B., 1987. Identification of root-knot nematodes (Meloidogyne species). Identif. Root-Knot Nematodes Meloidogyne Species.

Jiang, N., Feschotte, C., Zhang, X., Wessler, S.R., 2004. Using rice to understand the origin and amplification of miniature inverted repeat transposable elements (MITEs). Curr. Opin. Plant Biol. 7, 115–119. https://doi.org/10.1016/j.pbi.2004.01.004

Jones, J.T., Haegeman, A., Danchin, E.G.J., Gaur, H.S., Helder, J., Jones, M.G.K., Kikuchi, T., Manzanilla-López, R., Palomares-Rius, J.E., Wesemael, W.M.L., Perry, R.N., 2013. Top 10 plant-parasitic nematodes in molecular plant pathology: Top 10 plant-parasitic nematodes. Mol. Plant Pathol. 14, 946–961. https://doi.org/10.1111/mpp.12057

Jurka, J., Klonowski, P., Dagman, V., Pelton, P., 1996. Censor—a program for identification and elimination of repetitive elements from DNA sequences. Comput. Chem. 20, 119–121. https://doi.org/10.1016/S0097-8485(96)80013-1

Kanazawa, A., Liu, B., Kong, F., Arase, S., Abe, J., 2009. Adaptive Evolution Involving Gene Duplication and Insertion of a Novel Ty1/copia-Like Retrotransposon in Soybean. J. Mol. Evol. 69, 164–175. https://doi.org/10.1007/s00239-009-9262-1

Kaneko-Ishino, T., Ishino, F., 2012. The role of genes domesticated from LTR retrotransposons and retroviruses in mammals. Front. Microbiol. 3. https://doi.org/10.3389/fmicb.2012.00262

Kapitonov, V.V., Jurka, J., 2008. A universal classification of eukaryotic transposable elements implemented in Repbase. G E N E Ti C S 2.

Kapitonov, V.V., Jurka, J., 2005. RAG1 Core and V(D)J Recombination Signal Sequences Were Derived from Transib Transposons. PLoS Biol. 3, e181. https://doi.org/10.1371/journal.pbio.0030181

Katoh, K., 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30, 3059–3066. https://doi.org/10.1093/nar/gkf436

Kawakami, K., 2007. Tol2: a versatile gene transfer vector in vertebrates. Genome Biol. 8 Suppl 1, S7. https://doi.org/10.1186/gb-2007-8-s1-s7

Keeling, P.J., Palmer, J.D., 2008. Horizontal gene transfer in eukaryotic evolution. Nat. Rev. Genet. 9, 605–618. https://doi.org/10.1038/nrg2386

Kim, S., Park, J., Yeom, S.-I., Kim, Y.-M., Seo, E., Kim, K.-T., Kim, M.-S., Lee, J.M., Cheong, K., Shin, H.-S., Kim, S.-B., Han, K., Lee, Jundae, Park, M., Lee, H.-A., Lee, Hye-Young, Lee, Y., Oh, S., Lee, J.H., Choi, Eunhye, Choi, Eunbi, Lee, S.E., Jeon, J., Kim, H., Choi, G., Song, H., Lee, JunKi, Lee, S.-C., Kwon, J.-K., Lee, Hea-Young, Koo, N., Hong, Y., Kim, R.W., Kang, W.-H., Huh, J.H., Kang, B.-C., Yang, T.-J., Lee, Y.-H., Bennetzen, J.L., Choi, D., 2017. New reference genome sequences of hot pepper reveal the massive evolution of plant disease-resistance genes by retroduplication. Genome Biol. 18, 210. https://doi.org/10.1186/s13059-017-1341-9

Klasson, L., Kumar, N., Bromley, R., Sieber, K., Flowers, M., Ott, S.H., Tallon, L.J., Andersson, S.G.E., Dunning Hotopp, J.C., 2014. Extensive duplication of the Wolbachia DNA in chromosome four of Drosophila ananassae. BMC Genomics 15, 1097. https://doi.org/10.1186/1471-2164-15-1097

Klein, S.J., O'Neill, R.J., 2018. Transposable elements: genome innovation, chromosome diversity, and centromere conflict. Chromosome Res. 26, 5–23. https://doi.org/10.1007/s10577-017-9569-5

Knight, C.G., Zitzmann, N., Prabhakar, S., Antrobus, R., Dwek, R., Hebestreit, H., Rainey, P.B., 2006. Unraveling adaptive evolution: how a single point mutation affects the protein coregulation network. Nat. Genet. 38, 1015–1022. https://doi.org/10.1038/ng1867

Koch, P., Platzer, M., Downie, B.R., 2014. RepARK—de novo creation of repeat libraries from whole-genome NGS reads. Nucleic Acids Res. 42, e80–e80. https://doi.org/10.1093/nar/gku210

Kofler, R., 2018. SimulaTE: simulating complex landscapes of transposable elements of populations. Bioinformatics 34, 1419–1420. https://doi.org/10.1093/bioinformatics/btx772

Kofler, R., Gómez-Sánchez, D., Schlötterer, C., 2016. PoPoolationTE2: Comparative Population Genomics of Transposable Elements Using Pool-Seq. Mol. Biol. Evol. 33, 2759–2764. https://doi.org/10.1093/molbev/msw137

Kokot, M., Długosz, M., Deorowicz, S., 2017. KMC 3: counting and manipulating k-mer statistics. Bioinformatics 33, 2759–2761. https://doi.org/10.1093/bioinformatics/btx304

Kolpakov, R., 2003. mreps: efficient and flexible detection of tandem repeats in DNA. Nucleic Acids Res. 31, 3672–3678. https://doi.org/10.1093/nar/gkg617

Koutsovoulos, G.D., Marques, E., Arguel, M., Duret, L., Machado, A.C.Z., Carneiro, R.M.D.G., Kozlowski, D.K., Bailly-Bechet, M., Castagnone-Sereno, P., Albuquerque, E.V.S., Danchin, E.G.J., 2020. Population genomics supports clonal reproduction and multiple independent gains and losses of parasitic abilities in the most devastating nematode pest. Evol. Appl. 13, 442–457. https://doi.org/10.1111/eva.12881

Koutsovoulos, G.D., Poullet, M., Ashry, A.E., Kozlowski, D.K., Sallet, E., Rocha, M.D., Martin-Jimenez, C., Perfus-Barbeoch, L., Frey, J.-E., Ahrens, C., Kiewnick, S., Danchin, E.G.J., 2019. The polyploid genome of the mitotic parthenogenetic root-knot nematode *Meloidogyne enterolobii* (preprint). Genomics. https://doi.org/10.1101/586818

Kozlowski, D., 2020. Transposable Elements prediction and annotation in the M. incognita genome. https://doi.org/10.15454/EPTDOS

Kozlowski, D.K., Hassanaly-Goulamhoussen, R., Da-Rocha, M., Koutsovoulos, G., Bailly-Bechet, M., Danchin, E.G., 2020. Transposable Elements are an evolutionary force shaping genomic plasticity in the parthenogenetic root-knot nematode *Meloidogyne incognita* (preprint). Evolutionary Biology. https://doi.org/10.1101/2020.04.30.069948

Krasileva, K.V., 2019. The role of transposable elements and DNA damage repair mechanisms in gene duplications and gene fusions in plant genomes. Curr. Opin. Plant Biol. 48, 18–25. https://doi.org/10.1016/j.pbi.2019.01.004

Kumar, S., Stecher, G., Suleski, M., Hedges, S.B., 2017. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Mol. Biol. Evol. 34, 1812–1819. https://doi.org/10.1093/molbev/msx116

Lanciano, S., Carpentier, M.-C., Llauro, C., Jobet, E., Robakowska-Hyzorek, D., Lasserre, E., Ghesquière, A., Panaud, O., Mirouze, M., 2017. Sequencing the extrachromosomal circular mobilome reveals retrotransposon activity in plants. PLOS Genet. 13, e1006630. https://doi.org/10.1371/journal.pgen.1006630

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359. https://doi.org/10.1038/nmeth.1923

Lee, J., Waminal, N.E., Choi, H.-I., Perumal, S., Lee, S.-C., Nguyen, V.B., Jang, W., Kim, N.-H., Gao, L., Yang, T.-J., 2017. Rapid amplification of four retrotransposon families promoted speciation and genome size expansion in the genus Panax. Sci. Rep. 7, 9045. https://doi.org/10.1038/s41598-017-08194-5

Lemos, B., Branco, A.T., Hartl, D.L., 2010. Epigenetic effects of polymorphic Y chromosomes modulate chromatin components, immune response, and sexual conflict. Proc. Natl. Acad. Sci. 107, 15826–15831. https://doi.org/10.1073/pnas.1010383107

Lerat, E., 2010. Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. Heredity 104, 520–533. https://doi.org/10.1038/hdy.2009.165

Lerman, D.N., Michalak, P., Helin, A.B., Bettencourt, B.R., Feder, M.E., 2003. Modification of Heat-Shock Gene Expression in Drosophila melanogaster Populations via Transposable Elements. Mol. Biol. Evol. 20, 135–144. https://doi.org/10.1093/molbev/msg015

Li, H., Durbin, R., 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinforma. Oxf. Engl. 26, 589–595. https://doi.org/10.1093/bioinformatics/btp698

Lin, X., Faridi, N., Casola, C., 2016. An Ancient Trans-Kingdom Horizontal Transfer of *Penelope* -like Retroelements from Arthropods to Conifers. Genome Biol. Evol. evw076. https://doi.org/10.1093/gbe/evw076

Liu, J., 2004. siRNAs targeting an intronic transposon in the regulation of natural flowering behavior in Arabidopsis. Genes Dev. 18, 2873–2878. https://doi.org/10.1101/gad.1217304

Loxdale, H.D., 2010. Rapid genetic changes in natural insect populations. Ecol. Entomol. 35, 155–164. https://doi.org/10.1111/j.1365-2311.2009.01141.x

Lu, X., Sachs, F., Ramsay, L., Jacques, P.-É., Göke, J., Bourque, G., Ng, H.-H., 2014. The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. Nat. Struct. Mol. Biol. 21, 423–425. https://doi.org/10.1038/nsmb.2799

Lynch, V.J., 2007. [No title found]. BMC Evol. Biol. 7, 2. https://doi.org/10.1186/1471-2148-7-2

Ma, B., Xin, Y., Kuang, L., He, N., 2019. Distribution and Characteristics of Transposable Elements in the Mulberry Genome. Plant Genome 12, 180094. https://doi.org/10.3835/plantgenome2018.12.0094

Maciver, S.K., 2016. Asexual Amoebae Escape Muller's Ratchet through Polyploidy. Trends Parasitol. 32, 855–862. https://doi.org/10.1016/j.pt.2016.08.006

Madlung, A., 2013. Polyploidy and its effect on evolutionary success: old questions revisited with new tools. Heredity 110, 99–104. https://doi.org/10.1038/hdy.2012.79

Makałowski, W., Gotea, V., Pande, A., Makałowska, I., 2019. Transposable Elements: Classification, Identification, and Their Use As a Tool For Comparative Genomics, in: Anisimova, M. (Ed.), Evolutionary Genomics, Methods in Molecular Biology. Springer New York, New York, NY, pp. 177–207. https://doi.org/10.1007/978-1-4939-9074-0_6

McClintock, B., 1953. Induction of Instability at Selected Loci in Maize. Genetics 38, 579–599.

Mee, J.A., Brauner, C.J., Taylor, E.B., 2011. Repeat Swimming Performance and Its Implications for Inferring the Relative Fitness of Asexual Hybrid Dace (Pisces: *Phoxinus*) and Their Sexually Reproducing Parental Species. Physiol. Biochem. Zool. 84, 306–315. https://doi.org/10.1086/659245

Meloni, M., Reid, A., Caujapé-Castells, J., Marrero, Á., Fernández-Palacios, J.M., Mesa-Coelo, R.A., Conti, E., 2013. Effects of clonality on the genetic variability of rare, insular species: the case of *Ruta microcarpa* from the Canary Islands. Ecol. Evol. 3, 1569–1579. https://doi.org/10.1002/ece3.571

Møller, H.D., Larsen, C.E., Parsons, L., Hansen, A.J., Regenberg, B., Mourier, T., 2016. Formation of Extrachromosomal Circular DNA from Long Terminal Repeats of Retrotransposons in *Saccharomyces cerevisiae*. G3amp58 GenesGenomesGenetics 6, 453–462. https://doi.org/10.1534/g3.115.025858

Morgan, H.D., Sutherland, H.G.E., Martin, D.I.K., Whitelaw, E., 1999. Epigenetic inheritance at the agouti locus in the mouse. Nat. Genet. 23, 314–318. https://doi.org/10.1038/15490

Muller, H.J., 1964. The relation of recombination to mutational advance. Mutat. Res. Mol. Mech. Mutagen. 1, 2–9. https://doi.org/10.1016/0027-5107(64)90047-8

Munoz-Lopez, M., Garcia-Perez, J., 2010. DNA Transposons: Nature and Applications in Genomics. Curr. Genomics 11, 115–128. https://doi.org/10.2174/138920210790886871

Nakayama, Y., Yamaguchi, H., Einaga, N., Esumi, M., 2016. Pitfalls of DNA Quantification Using DNA-Binding Fluorescent Dyes and Suggested Solutions. PloS One 11, e0150528. https://doi.org/10.1371/journal.pone.0150528

Naville, M., Henriet, S., Warren, I., Sumic, S., Reeve, M., Volff, J.-N., Chourrout, D., 2019. Massive Changes of Genome Size Driven by Expansions of Non-autonomous Transposable Elements. Curr. Biol. 29, 1161-1168.e6. https://doi.org/10.1016/j.cub.2019.01.080

Nikoh, N., Tanaka, K., Shibata, F., Kondo, N., Hizume, M., Shimada, M., Fukatsu, T., 2008. Wolbachia genome integrated in an insect chromosome: Evolution and fate of laterally transferred endosymbiont genes. Genome Res. 18, 272–280. https://doi.org/10.1101/gr.7144908

Novak, P., Neumann, P., Pech, J., Steinhaisl, J., Macas, J., 2013. RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. Bioinformatics 29, 792–793. https://doi.org/10.1093/bioinformatics/btt054

O'Donnell, K.A., Burns, K.H., 2010. Mobilizing diversity: transposable element insertions in genetic variation and disease. Mob. DNA 1, 21. https://doi.org/10.1186/1759-8753-1-21

Opperman, C.H., Bird, D.M., Williamson, V.M., Rokhsar, D.S., Burke, M., Cohn, J., Cromer, J., Diener, S., Gajan, J., Graham, S., Houfek, T.D., Liu, Q., Mitros, T., Schaff, J., Schaffer, R., Scholl, E., Sosinski, B.R., Thomas, V.P., Windham, E., 2008. Sequence and genetic map of Meloidogyne hapla: A compact nematode genome for plant parasitism. Proc. Natl. Acad. Sci. 105, 14802–14807. https://doi.org/10.1073/pnas.0805946105

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J.R.A., Hellinga, A.J., Lugo, C.S.B., Elliott, T.A., Ware, D., Peterson, T., Jiang, N., Hirsch, C.N., Hufford, M.B., 2019. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. 20, 275. https://doi.org/10.1186/s13059-019-1905-y

Ozias-Akins, P., Conner, J.A., 2020. Clonal Reproduction through Seeds in Sight for Crops. Trends Genet. 36, 215–226. https://doi.org/10.1016/j.tig.2019.12.006

Pace, J.K., Feschotte, C., 2007. The evolutionary history of human DNA transposons: Evidence for intense activity in the primate lineage. Genome Res. 17, 422–432. https://doi.org/10.1101/gr.5826307

Paganini, J., Campan-Fournier, A., Da Rocha, M., Gouret, P., Pontarotti, P., Wajnberg, E., Abad, P., Danchin, E.G.J., 2012. Contribution of Lateral Gene Transfers to the Genome Composition and Parasitic Ability of Root-Knot Nematodes. PLoS ONE 7, e50875. https://doi.org/10.1371/journal.pone.0050875

Pardue, M.-L., Rashkova, S., Casacuberta, E., DeBaryshe, P.G., George, J.A., Traverse, K.L., 2005. Two retrotransposons maintain telomeres in Drosophila. Chromosome Res. 13, 443–453. https://doi.org/10.1007/s10577-005-0993-6

Pastuzyn, E.D., Day, C.E., Kearns, R.B., Kyrke-Smith, M., Taibi, A.V., McCormick, J., Yoder, N., Belnap, D.M., Erlendsson, S., Morado, D.R., Briggs, J.A.G., Feschotte, C., Shepherd, J.D., 2018. The Neuronal Gene Arc Encodes a Repurposed Retrotransposon Gag Protein that Mediates Intercellular RNA Transfer. Cell 172, 275-288.e18. https://doi.org/10.1016/j.cell.2017.12.024

Patterson, E.L., Saski, C.A., Sloan, D.B., Tranel, P.J., Westra, P., Gaines, T.A., 2019. The Draft Genome of Kochia scoparia and the Mechanism of Glyphosate Resistance via Transposon-

Mediated EPSPS Tandem Gene Duplication. Genome Biol. Evol. 11, 2927–2940. https://doi.org/10.1093/gbe/evz198

Petersen, M., Armisén, D., Gibbs, R.A., Hering, L., Khila, A., Mayer, G., Richards, S., Niehuis, O., Misof, B., 2019. Diversity and evolution of the transposable element repertoire in arthropods with particular reference to insects. BMC Evol. Biol. 19, 11. https://doi.org/10.1186/s12862-018-1324-9

Phan, N.T., Orjuela, J., Danchin, E.G.J., Klopp, C., Perfus-Barbeoch, L., Kozlowski, D.K., Koutsovoulos, G.D., Lopez-Roques, C., Bouchez, O., Zahm, M., Besnard, G., Bellafiore, S., 2020. Genome structure and content of the rice root-knot nematode (*Meloidogyne graminicola*). Ecol. Evol. ece3.6680. https://doi.org/10.1002/ece3.6680

Piégu, B., Bire, S., Arensburger, P., Bigot, Y., 2015. A survey of transposable element classification systems – A call for a fundamental update to meet the challenge of their diversity and complexity. Mol. Phylogenet. Evol. 86, 90–109. https://doi.org/10.1016/j.ympev.2015.03.009

Platt, R.N., Vandewege, M.W., Kern, C., Schmidt, C.J., Hoffmann, F.G., Ray, D.A., 2014. Large Numbers of Novel miRNAs Originate from DNA Transposons and Are Coincident with a Large Species Radiation in Bats. Mol. Biol. Evol. 31, 1536–1545. https://doi.org/10.1093/molbev/msu112

Platt, R.N., Vandewege, M.W., Ray, D.A., 2018. Mammalian transposable elements and their impacts on genome evolution. Chromosome Res. 26, 25–43. https://doi.org/10.1007/s10577-017-9570-z

Quadrana, L., Etcheverry, M., Gilly, A., Caillieux, E., Madoui, M.-A., Guy, J., Bortolini Silveira, A., Engelen, S., Baillet, V., Wincker, P., Aury, J.-M., Colot, V., 2019. Transposition favors the generation of large effect mutations that may facilitate rapid adaption. Nat. Commun. 10, 3421. https://doi.org/10.1038/s41467-019-11385-5

Quesneville, H., Bergman, C.M., Andrieu, O., Autard, D., Nouaud, D., Ashburner, M., Anxolabehere, D., 2005. Combined Evidence Annotation of Transposable Elements in Genome Sequences. PLoS Comput. Biol. 1, e22. https://doi.org/10.1371/journal.pcbi.0010022

Quesneville, H., Nouaud, D., Anxolabéhère, D., 2003. Detection of New Transposable Element Families in Drosophila melanogaster and Anopheles gambiae Genomes. J. Mol. Evol. 57, S50–S59. https://doi.org/10.1007/s00239-003-0007-2

Quinlan, A.R., Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842. https://doi.org/10.1093/bioinformatics/btq033

Rabeling, C., Kronauer, D.J.C., 2013. Thelytokous Parthenogenesis in Eusocial Hymenoptera. Annu. Rev. Entomol. 58, 273–292. https://doi.org/10.1146/annurev-ento-120811-153710

Radke, D.W., Lee, C., 2015. Adaptive potential of genomic structural variation in human and mammalian evolution. Brief. Funct. Genomics 14, 358–368. https://doi.org/10.1093/bfgp/elv019

Ranallo-Benavidez, T.R., Jaron, K.S., Schatz, M.C., 2020. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat. Commun. 11, 1432. https://doi.org/10.1038/s41467-020-14998-3

Rice, W.R., 2002. Experimental tests of the adaptive significance of sexual recombination. Nat. Rev. Genet. 3, 241–251. https://doi.org/10.1038/nrg760

Rishishwar, L., Mariño-Ramírez, L., Jordan, I.K., 2016. Benchmarking computational tools for polymorphic transposable element detection. Brief. Bioinform. bbw072. https://doi.org/10.1093/bib/bbw072

Rogers, P.C., Gale, J.A., 2017. Restoration of the iconic Pando aspen clone: emerging evidence of recovery. Ecosphere 8, e01661. https://doi.org/10.1002/ecs2.1661

Rosso, M.-N., Favery, B., Piotte, C., Arthaud, L., De Boer, J.M., Hussey, R.S., Bakker, J., Baum, T.J., Abad, P., 1999. Isolation of a cDNA Encoding a β-1,4-endoglucanase in the Root-Knot Nematode *Meloidogyne incognita* and Expression Analysis During Plant Parasitism. Mol. Plant-Microbe Interactions® 12, 585–591. https://doi.org/10.1094/MPMI.1999.12.7.585

SanMiguel, P., Tikhonov, A., Jin, Y.-K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P.S., Edwards, K.J., Lee, M., Avramova, Z., Bennetzen, J.L., 1996. Nested Retrotransposons in the Intergenic Regions of the Maize Genome. Science 274, 765–768. https://doi.org/10.1126/science.274.5288.765

Schartl, M., Schmid, M., Nanda, I., 2016. Dynamics of vertebrate sex chromosome evolution: from equal size to giants and dwarfs. Chromosoma 125, 553–571. https://doi.org/10.1007/s00412-015-0569-y

Schiffer, P.H., Danchin, E.G.J., Burnell, A.M., Creevey, C.J., Wong, S., Dix, I., O'Mahony, G., Culleton, B.A., Rancurel, C., Stier, G., Martínez-Salazar, E.A., Marconi, A., Trivedi, U., Kroiher, M., Thorne, M.A.S., Schierenberg, E., Wiehe, T., Blaxter, M., 2019. Signatures of the Evolution of Parthenogenesis and Cryptobiosis in the Genomes of Panagrolaimid Nematodes. iScience 21, 587–602. https://doi.org/10.1016/j.isci.2019.10.039

Schmidt, J.M., Good, R.T., Appleton, B., Sherrard, J., Raymant, G.C., Bogwitz, M.R., Martin, J., Daborn, P.J., Goddard, M.E., Batterham, P., Robin, C., 2010. Copy Number Variation and Transposable Elements Feature in Recent, Ongoing Adaptation at the Cyp6g1 Locus. PLoS Genet. 6, e1000998. https://doi.org/10.1371/journal.pgen.1000998

Schrader, L., Kim, J.W., Ence, D., Zimin, A., Klein, A., Wyschetzki, K., Weichselgartner, T., Kemena, C., Stökl, J., Schultner, E., Wurm, Y., Smith, C.D., Yandell, M., Heinze, J., Gadau, J., Oettler, J., 2014. Transposable element islands facilitate adaptation to novel environments in an invasive species. Nat. Commun. 5, 5495. https://doi.org/10.1038/ncomms6495

Schrader, L., Schmitz, J., 2019. The impact of transposable elements in adaptive evolution. Mol. Ecol. 28, 1537–1549. https://doi.org/10.1111/mec.14794

Seberg, O., Petersen, G., 2009. A unified classification system for eukaryotic transposable elements should reflect their phylogeny. Nat. Rev. Genet. 10, 276–276. https://doi.org/10.1038/nrg2165-c3

Seidl, M.F., Thomma, B.P.H.J., 2017. Transposable Elements Direct The Coevolution between Plants and Microbes. Trends Genet. 33, 842–851. https://doi.org/10.1016/j.tig.2017.07.003

Seidl, M.F., Thomma, B.P.H.J., 2014. Sex or no sex: Evolutionary adaptation occurs regardless: Insights & Perspectives. BioEssays 36, 335–345. https://doi.org/10.1002/bies.201300155

Selmecki, A., Forche, A., Berman, J., 2010. Genomic Plasticity of the Human Fungal Pathogen Candida albicans. Eukaryot. Cell 9, 991–1008. https://doi.org/10.1128/EC.00060-10

Selmecki, A.M., Maruvka, Y.E., Richmond, P.A., Guillet, M., Shoresh, N., Sorenson, A.L., De, S., Kishony, R., Michor, F., Dowell, R., Pellman, D., 2015. Polyploidy can drive rapid adaptation in yeast. Nature 519, 349–352. https://doi.org/10.1038/nature14187

Serrato-Capuchina, A., Matute, D., 2018. The Role of Transposable Elements in Speciation. Genes 9, 254. https://doi.org/10.3390/genes9050254

Shang, Y., Yang, F., Schulman, A.H., Zhu, J., Jia, Y., Wang, J., Zhang, X.-Q., Jia, Q., Hua, W., Yang, J., Li, C., 2017. Gene Deletion in Barley Mediated by LTR-retrotransposon BARE. Sci. Rep. 7, 43766. https://doi.org/10.1038/srep43766

Simion, P., Narayan, J., Houtain, A., Derzelle, A., Baudry, L., Nicolas, E., Cariou, M., Guiglielmoni, N., Kozlowski, D.K., Gaudray, F.R., Terwagne, M., Virgo, J., Noel, B., Wincker, P., Danchin, E.G., Marbouty, M., Hallet, B., Koszul, R., Limasset, A., Flot, J.-F., Van Doninck, K., 2020. Homologous chromosomes in asexual rotifer *Adineta vaga* suggest automixis (preprint). Evolutionary Biology. https://doi.org/10.1101/2020.06.16.155473

Soltis, D.E., Visger, C.J., Soltis, P.S., 2014. The polyploidy revolution then...and now: Stebbins revisited. Am. J. Bot. 101, 1057–1078. https://doi.org/10.3732/ajb.1400178

Somvanshi, V.S., Ghosh, O., Budhwar, R., Dubay, B., Shukla, R.N., Rao, U., 2018. A comprehensive annotation for the root-knot nematode Meloidogyne incognita proteome data. Data Brief 19, 1073–1079. https://doi.org/10.1016/j.dib.2018.05.131

Stapley, J., Santure, A.W., Dennis, S.R., 2015. Transposable elements as agents of rapid adaptation may explain the genetic paradox of invasive species. Mol. Ecol. 24, 2241–2252. https://doi.org/10.1111/mec.13089

Steenwyk, J.L., Rokas, A., 2018. Copy Number Variation in Fungi and Its Implications for Wine Yeast Genetic Diversity and Adaptation. Front. Microbiol. 9, 288. https://doi.org/10.3389/fmicb.2018.00288

Stuart, T., Eichten, S.R., Cahn, J., Karpievitch, Y.V., Borevitz, J.O., Lister, R., 2016. Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation. eLife 5, e20777. https://doi.org/10.7554/eLife.20777

Suh, A., 2019. Genome Size Evolution: Small Transposons with Large Consequences. Curr. Biol. 29, R241–R243. https://doi.org/10.1016/j.cub.2019.02.032

Susič, N., Koutsovoulos, G.D., Riccio, C., Danchin, E.G.J., Blaxter, M.L., Lunt, D.H., Strajnar, P., Širca, S., Urek, G., Stare, B.G., 2020. Genome sequence of the root-knot nematode Meloidogyne luci. J. Nematol. 52, 1–5. https://doi.org/10.21307/jofnem-2020-025

Szitenberg, A., Cha, S., Opperman, C.H., Bird, D.M., Blaxter, M.L., Lunt, D.H., 2016. Genetic Drift, Not Life History or RNAi, Determine Long-Term Evolution of Transposable Elements. Genome Biol. Evol. 8, 2964–2978. https://doi.org/10.1093/gbe/evw208

Szitenberg, A., Salazar-Jaramillo, L., Blok, V.C., Laetsch, D.R., Joseph, S., Williamson, V.M., Blaxter, M.L., Lunt, D.H., 2017. Comparative Genomics of Apomictic Root-Knot Nematodes:

Hybridization, Ploidy, and Dynamic Genome Change. Genome Biol. Evol. 9, 2844–2861. https://doi.org/10.1093/gbe/evx201

Triantaphyllou, A.C., 1981. Oogenesis and the Chromosomes of the Parthenogenic Root-knot Nematode Meloidogyne incognita. J. Nematol. 13, 95–104.

Trudgill, D.L., Blok, V.C., 2001. A POMICTIC , P OLYPHAGOUS R OOT -K NOT N EMATODES : Exceptionally Successful and Damaging Biotrophic Root Pathogens. Annu. Rev. Phytopathol. 39, 53–77. https://doi.org/10.1146/annurev.phyto.39.1.53

Tzortzakakis, E.A., Conceição, I., Dias, A.M., Simoglou, K.B., Abrantes, I., 2014. Occurrence of a new resistant breaking pathotype of Meloidogyne incognita on tomato in Greece. J. Plant Dis. Prot. 121, 184–186. https://doi.org/10.1007/BF03356508

Venkatesh, B., Lee, A.P., Ravi, V., Maurya, A.K., Lian, M.M., Swann, J.B., Ohta, Y., Flajnik, M.F., Sutoh, Y., Kasahara, M., Hoon, S., Gangu, V., Roy, S.W., Irimia, M., Korzh, V., Kondrychyn, I., Lim, Z.W., Tay, B.-H., Tohari, S., Kong, K.W., Ho, S., Lorente-Galdos, B., Quilez, J., Marques-Bonet, T., Raney, B.J., Ingham, P.W., Tay, A., Hillier, L.W., Minx, P., Boehm, T., Wilson, R.K., Brenner, S., Warren, W.C., 2014. Elephant shark genome provides unique insights into gnathostome evolution. Nature 505, 174–179. https://doi.org/10.1038/nature12826

Volff, J.-N., 2006. Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. BioEssays 28, 913–922. https://doi.org/10.1002/bies.20452

Vrijenhoek, R.C., Parker, E.D., 2009. Geographical Parthenogenesis: General Purpose Genotypes and Frozen Niche Variation, in: Schön, I., Martens, K., Dijk, P. (Eds.), Lost Sex. Springer Netherlands, Dordrecht, pp. 99–131. https://doi.org/10.1007/978-90-481-2770-2_6

Welch, D.M., 2000. Evidence for the Evolution of Bdelloid Rotifers Without Sexual Reproduction or Genetic Exchange. Science 288, 1211–1215. https://doi.org/10.1126/science.288.5469.1211

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., Schulman, A.H., 2007. A unified classification system for eukaryotic transposable elements. Nat. Rev. Genet. 8, 973–982. https://doi.org/10.1038/nrg2165

Wright, S., Finnegan, D., 2001. Genome evolution: Sex and the transposable element. Curr. Biol. 11, R296–R299. https://doi.org/10.1016/S0960-9822(01)00168-3

Wu, C., Lu, J., 2019. Diversification of Transposable Elements in Arthropods and Its Impact on Genome Evolution. Genes 10, 338. https://doi.org/10.3390/genes10050338

Wu, Q., Smith, N., Zhang, D., Zhou, C., Wang, M.-B., 2018. Root-Specific Expression of a Jacalin Lectin Family Protein Gene Requires a Transposable Element Sequence in the Promoter. Genes 9, 550. https://doi.org/10.3390/genes9110550

Wybouw, N., Pauchet, Y., Heckel, D.G., Van Leeuwen, T., 2016. Horizontal Gene Transfer Contributes to the Evolution of Arthropod Herbivory. Genome Biol. Evol. 8, 1785–1801. https://doi.org/10.1093/gbe/evw119

Xin, Y., Ma, B., Xiang, Z., He, N., 2019. Amplification of miniature inverted-repeat transposable elements and the associated impact on gene regulation and alternative splicing in mulberry (Morus notabilis). Mob. DNA 10, 27. https://doi.org/10.1186/s13100-019-0169-0

Zeng, L., Kortschak, R.D., Raison, J.M., Bertozzi, T., Adelson, D.L., 2018. Superior ab initio identification, annotation and characterisation of TEs and segmental duplications from genome assemblies. PLOS ONE 13, e0193588. https://doi.org/10.1371/journal.pone.0193588

Zeyl, C., Mizesko, M., Visser, J.A.G.M. de, 2001. MUTATIONAL MELTDOWN IN LABORATORY YEAST POPULATIONS. Evolution 55, 909. https://doi.org/10.1554/0014-3820(2001)055[0909:MMILYP]2.0.CO;2

Résumé

Les nématodes à galles (genre *Meloidogyne*) sont parmi les parasites de plantes les plus nuisibles. Ces organismes se distinguent par la diversité de leurs modes de reproduction. Étonnement, il a été observé que les espèces les plus néfastes se reproduisent de manière strictement asexuée et certaines sont capables de contourner la résistance de la plante hôte en un nombre de génération restreint. Ainsi, bien qu'incapables de combiner des mutations bénéfiques provenant de différents individus, ces espèces peuvent s'adapter à des changements du milieu. L'adaptabilité et le succès parasitaire de ces espèces malgré l'absence de reproduction sexuée semblent paradoxaux et doivent reposer sur d'autres mécanismes capables de générer de la plasticité génétique.

Les éléments transposables (ETs) sont des fragments d'ADN capables de se déplacer et de se multiplier dans les génomes. De ce fait, les ETs peuvent avoir des répercussions fonctionnelles et structurales sur les génomes. Les ETs pourraient constituer un des mécanismes permettant de générer la diversité génétique nécessaire à l'adaptabilité chez *Meloidogyne*.

En réalisant une analyse de génomique comparative entre 7 espèces de *Meloidogyne*, j'ai montré que le contenu en ET actuellement observé chez ces espèces semble suivre leur histoire évolutive et la dérive entre espèces, plutôt que des traits d'histoires de vie tels que le mode de reproduction. Par ailleurs, cette analyse soutient une activité récente des ETs au sein de la plupart des espèces. Ces résultats suggèrent que bien que les ETs aient récemment été actifs au sein du genre *Meloidogyne*, leur dynamique dans les génomes semble spécifique à chaque espèce et nécessite donc une étude ciblée.

Dans cette optique, j'ai concentré mes efforts sur *M. incognita*, l'espèce à reproduction asexuée la plus préjudiciable pour l'agriculture. Dans un premier temps, j'ai annoté en détail le contenu en ET du génome de *M. incognita*. L'analyse du contenu en ET a confirmé que ces éléments ont probablement été récemment actifs dans le génome. Afin de mieux caractériser cette activité et ses potentiels effets, j'ai ensuite estimé la mobilité de ces ET via une analyse de génomique comparative portant sur 12 isolats géographiques. J'ai pu identifier plusieurs milliers de loci dans le génome où les fréquences de présence d'ETs varient entre les différents isolats. Par une approche phylogénétique, j'ai montré que ces variations de fréquence d'ET suivent l'histoire évolutive des isolats étudiés. Par rapport au génome de référence, j'ai prédit des néo-insertions d'ETs, certaines ayant un potentiel impact fonctionnel. Les validations expérimentales réalisées pour plusieurs de ces insertions confirment le rôle probable des ETs dans la plasticité du génome de cette espèce.

Lors de cette analyse, j'ai également identifié des ET présents à des fréquences intermédiaires (différentes de 0 ou 1) au sein de chaque isolat, signe d'une variabilité entre individus. Or *M. incognita* est un organisme supposé clonal et chaque isolat étudié est issu d'une seule femelle. En nous concentrant sur l'analyse d'un de ces isolats, nous avons validé expérimentalement plusieurs polymorphismes de présence, ce qui confirme qu'il existe une hétérogénéité génétique non négligeable au sein d'un même isolat. Par ailleurs, en comparant des données de séquençage issues du même isolat a deux points de cinétique différents, nous avons pu prédire que quelques ETs varient en fréquences au sein de l'isolat en un faible nombre de génération, ce qui sous-entend que ces ETs participent à la dynamique de la diversité génétique de cet organisme.

Ces résultats posent les bases pour de futures analyses visant à déterminer si l'activité des ETs joue un rôle actif dans la capacité d'espèces à s'adapter à leur environnement en absence de reproduction sexuée.