Plasticité des génomes des pucerons des céréales et de leur plante hôte : recherche in silico et in vitro des éléments transposables des superfamilles Tc1-mariner-IS630 et piggyBac

Maryem Bouallègue

NNT : 2017SACLS061

THESE DE DOCTORAT
DE
L'UNIVERSITE DE TUNIS EL MANAR
ET DE
L'UNIVERSITE PARIS-SACLAY

PREPAREE A L'UNIVERSITE DE TUNIS EL MANAR ET A L'UNIVERSITE PARIS-SUD

ECOLE DOCTORALE STVT SCIENCES ET TECHNOLOGIES DU VIVANT ET DE LA TERRE
Spécialité de doctorat : Sciences Biologiques

ECOLE DOCTORALE N° 577 SDSV STRUCTURE ET DYNAMIQUE DES SYSTEMES VIVANTS
Spécialité de doctorat : Sciences de la Vie et de la Santé

Par

Mme Maryem Bouallègue

Plasticité des génomes des pucerons des céréales et de leur plante hôte : Recherche *in silico* et *in vitro* des éléments transposables des superfamilles *Tc1-mariner-IS630* et *piggyBac*

Thèse présentée et soutenue à la salle des conférences Al-Khawarizmi, de la Faculté des Sciences de Tunis, le 27 Mars 2017 à 9h30

Composition du jury :

Président :	**Mme Amel Ben Ammar Gaaied**	Pr. Faculté des Sciences de Tunis
Rapporteurs :	**Mr Ali-Faouzi Gargouri**	Pr. Centre de Biotechnologie de Sfax
	Mme Cristina Vieira-Heddi	Pr. Université de Claude-Bernard Lyon1
Examinateur :	**Mme Mireille Bétermier**	DR au CNRS, Université Paris-Saclay
Directeurs de Thèse :	**Mr Mohamed Makni**	Pr. Faculté des Sciences de Tunis
	Mr Pierre Capy	Pr. Université Paris-Saclay

Que tout ce que les mortels font
Soit une oeuvre monumentale ;
Il faut un fleuve à tout Tantale,
A tout Sisyphe il faut un mont.

Mohamed Yosri Ben Hemdene

A Mes parents

Vous avez su m'inculquer le sens de la responsabilité et m'avez toujours appris à ne pas me contenter de la médiocrité ainsi qu'à chercher à me dépasser. C'est à travers vos encouragements, vos critiques et votre soutien que j'ai pu avancer et si j'en suis là aujourd'hui, ce n'est que le fruit de vos efforts et de vos sacrifices, en espérant pouvoir vous combler à mon tour et rester votre fierté. Tous les mots du monde ne sauraient exprimer l'infini amour, le profond respect et la reconnaissance éternelle que je vous porte.

A mon Mari

Ton irremplaçable et indispensable soutien, ton amour, ta patience et tes encouragements ont été ma source d'énergie et d'apaisement. C'est à travers tes avis et tes conseils en tant que « reviewer, œil de lynx» que j'ai pu me corriger et évoluer. Nos efforts et nos sacrifices ont payé par l'achèvement de « notre » thèse. J'exprime ma gratitude, mon amour et mon profond respect. Que dieu illumine notre chemin et nous apporte du bonheur.

A mes sœurs, à mon beau-frère et à mon neveu

Imen, « my second mom » et Ons, en plus de m'avoir procuré un accueil chaleureux et paisible, votre aide inconditionnelle, vos encouragements et tous ces moments partagés ont amplement contribué à mon avancement et à mon épanouissement. J'espère pouvoir vous en offrir autant.

Rayen, « Le bonheur », tu es mon « fils », mon ami, mon rayon de soleil. Nos moments magiques me font oublier tous mes soucis et m'apportent de l'élan.

Je suis profondément reconnaissante et, même si nous sommes séparés par la distance physique, ce n'est certainement pas celle du cœur.

Malek « ma meilleure amie », ton grand cœur, tes connaissances et ton écoute ont toujours fait de toi mon refuge préféré. Merci pour ce que tu es en te souhaitant un avenir radieux, meilleur que tu ne l'imagines.

A mon cercle familial Ben Hemdène ; A mes beaux-parents, mes beaux-frères et mes belles-sœurs, mes neveux et ma nièce

Vous m'avez accordé amour, attention, confiance, encouragements et compréhension. Vous avez contribué à mon équilibre. J'exprime mon profond respect et ma grande affection. Sachez mes trois petits choux, que vous êtes indispensables à mon bonheur et à ma détente.

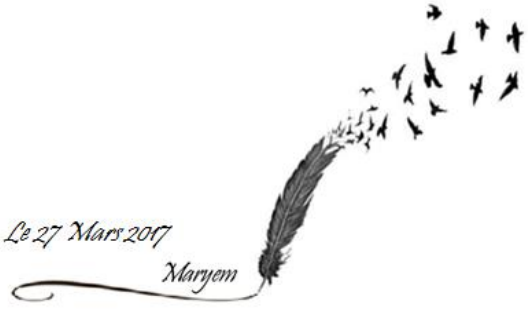
A mes oncles et à mes tantes, cousins et cousines

Votre gentillesse, votre amour et votre confiance ont été ma source d'apaisement. J'espère que vous trouverez dans ce travail l'expression de ma vive reconnaissance.

Aux membres et amis de ma famille

A tous mes amis

A la mémoire de mes grands-parents, ma tante et ma cousine, *qui ont toujours été dans mon esprit et dans mon cœur.*



Remerciements

La thèse, quelle aventure ! A la fois unique, passionnante, épanouissante et, parfois, éprouvante.
Je ne pourrais dire si ces plusieurs années de labeur se sont écoulées vite ou lentement, mais une chose est sûre, elles auront été remplies de moments et de rencontres inoubliables.

Je remercie tout d'abord mes directeurs de thèse : Monsieur **Mohamed Makni**, Professeur à la Faculté des Sciences de Tunis, d'avoir accepté que je puisse continuer mes recherches au sein de son équipe de Génomique des Insectes Ravageurs des Cultures d'intérêt agronomique (GIRC), UR11ES10 et Monsieur **Pierre Capy**, Professeur à l'Université de Paris-Saclay, de m'avoir fait confiance en m'accueillant au sein du laboratoire Evolution, Génomes, Comportement et Ecologie (EGCE), CNRS à Gif-sur-Yvette.
Je voudrais leur exprimer ma gratitude pour toutes les heures qu'ils m'ont consacrées à diriger ma recherche et à me conseiller tout en me laissant libre de mener mes travaux. Je n'ai pas eu de difficulté à les suivre, ils étaient toujours sur la même longueur d'onde, voire complémentaires, avec plusieurs opportunités de discussions, fructueuses et constructives. J'aimerais également leur dire à quel point j'ai apprécié leur grande disponibilité et leur respect sans faille des délais serrés de relecture que je leur ai adressés. Enfin, j'ai été extrêmement sensible à leurs qualités humaines d'écoute, de compréhension et de soutien tout au long de ce travail doctoral. Ils sont des exemples à suivre et sont devenus comme des pères au fil des années. J'espère être toujours à la hauteur de leurs estimations.

Mes remerciements vont également à mes rapporteurs, Madame **Cristina Vieira-Heddi**, Professeur à l'Université Claude-Bernard de Lyon 1 et Monsieur **Ali-Faouzi Gargouri**, Professeur au Centre de Biotechnologie de Sfax, pour tout l'intérêt qu'ils ont porté à mes travaux et pour m'avoir fait l'honneur de les évaluer et de les juger.

Je suis particulièrement reconnaissante envers Madame **Mireille Bétermier**, Directrice de recherche à l'Institut de Biologie Intégrative de la Cellule (I2BC, Paris-Saclay), pour son accueil chaleureux à chaque fois que je l'avais sollicitée, pour ses multiples conseils et encouragements, ainsi que pour avoir accepté d'être membre du jury.

Je tiens à remercier vivement Madame **Amel Ben Ammar-Gaaied**, Professeur à la Faculté des Sciences de Tunis, d'avoir aimablement accepté de présider mon jury et de juger ce travail. Ses remarques pertinentes et ses qualités humaines m'ont toujours été d'un grand support.

Mention spéciale à mon Parrain, **Jacques-Deric Rouault**, chercheur retraité à l'EGCE, de m'avoir permis de faire mes premiers pas au sein du laboratoire, et pour l'esprit et les approches mathématiques-philosophiques qu'il m'a inculqués. Très humblement, je voudrais le remercier pour tous les conseils et les encouragements répétés ainsi que pour tous les moments culturels et humoristiques qui m'ont fait évader.

J'adresse mes vifs remerciements à Madame **Hanem Makni**, Professeur et Directrice de l'Institut Supérieur de l'Animation pour la Jeunesse et la Culture de Bir El Bey et à Madame **Maha Mezghani-Khemakhem**, Maître de conférences à la Faculté des Sciences de Tunis, d'avoir suivi l'évolution de ce travail, pour leurs conseils judicieux et leurs recommandations ainsi que pour leur soutien. Je leur suis particulièrement redevable.

Je voudrais exprimer toute ma gratitude à **Aurélie Hua-Van**, Maître de conférences à l'Université Paris-Sud, et **Jonathan Filée**, chargé de recherche au CNRS/EGCE, pour leurs précieux conseils stimulant, à maintes fois, ma réflexion vers de nouvelles pistes, pour leur disponibilité et leur aide précieuse.

J'aimerais aussi remercier, **Julien Bischerour**, chargé de recherche à l'Institut de Biologie Intégrative de la Cellule (I2BC, Paris-Saclay) pour les nombreuses discussions constructives et ses conseils avisés.

Mes remerciements vont également à **Dhia Bouktila**, Maître de conférences à l'Institut Supérieur de Biotechnologie de Béja, pour ses conseils judicieux et ses recommandations.

J'ai pu valser entre deux cadres qui m'ont offert beaucoup de sympathie et de convivialité, je remercie tous les membres de l'unité GIRC ainsi que tous les membres du laboratoire EGCE, en particulier, ceux du pôle de recherche Evolution & Génomes. Les discussions enrichissantes que j'ai pu avoir avec chacun d'entre eux, leur disponibilité et leur aide précieuse m'ont beaucoup apportées.

Ce travail n'aurait pu être accompli sans la tutelle du Ministère de l'Enseignement Supérieur et de la Recherche Scientifique de Tunis, du Centre National de Recherche Scientifique (CNRS), de l'Université de Tunis El Manar et de l'Université de Paris-Sud/Saclay, des membres des écoles doctorales Sciences et Technologies du Vivant et de la Terre (STVT) & Structure et Dynamique Des Systèmes Vivants (SDSV, N°577), ainsi qu'à l'intervention de l'Institut Diversité Ecologie et Evolution du Vivant (IDEEV). Même si, souvent, les étapes administratives étaient périlleuses, il me paraît indispensable de les remercier pour toutes les aides et bourses qui m'ont été octroyées pour avancer, faciliter mes voyages et présenter mes travaux de recherche à l'échelle internationale.

Enfin je suis reconnaissante envers tous mes enseignants de la Faculté des Sciences de Tunis, qui ont contribué à ma formation académique et qui ont joué un rôle d'incitateurs. Ils ont su me communiquer la passion de ce métier et faire naître en moi le goût de la recherche.

Sommaire

Liste des abréviations

ADN : Acide Désoxyribonucléique
ADNc : ADN Complémentaire
ADNg : ADN génomique
ARN : Acide Ribonucléique
BETs : Bromure d'éthidium
BLAST : Basic Local Alignment Search Tool
CO I : Cytochrome Oxydase sous unité I
CTAB : Cetyl trimethylammonium bromide
D : Acide aspartique
dNTP : DésoxyriboNucléosides Tri-Phosphate
DO : Densité Optique
DR : Direct Repeats
E : Acide glutamique
ETs : élément Transposable
h : Heures
HTH : Hélice-Tour-Hélice
IPTG : Isopropyl-β-D-1-thiogalactopyranoside
IS : Séquence d'Insertion
kb : Kilo paires de bases
LB : Luria Bertani
LTR : Long Terminal Repeats
Mb : Mégabase
Mha : Millions d'hectares
min : Minutes
MITE : Miniature inverted–repeat transposable element
ML : Maximum likelihood
MLE : *mariner*-Like Element
Mt : Millions de tonnes
Mya : Millions d'années
NCBI : National Center for Biotechnology Information
NLS : Signal de Localisation Nucléaire
ORF : Open Reading Frame
pb : Paire de bases
PBLE : *piggyBac*-Like Elements
PCR : Polymerase Chain Reaction
PCRi : PCR inverse
PGBD : piggyBac derived genes/proteins
pH : Potentiel Hydrogène
rpm : Rotation par minute
sec : secondes
SOB : Super Optimal Broth
STIR : Subterminal Inverted Repeats
Th : Température d'hybridation
TH : Transfert horizontal
TIR : Terminal Inverted Repeats
TLE : *Tc1*-Like Element
TSD : Target Site Duplication
UPGM-VM : UPGM- Variation Metric
UTR : Untranslated region

Avant-propos

L'orge, le blé et l'avoine sont des *poacées* considérées parmi les principales céréales cultivées dans le monde. Ces céréales sont utilisées pour la consommation humaine et l'alimentation des bétails.

Toutefois, la céréaliculture est menacée par divers facteurs abiotiques, c'est à dire l'ensemble des facteurs physico-chimiques agissant sur la biocénose, et des facteurs biotiques plus particulièrement les insectes ravageurs capables également de transmettre des agents phytopathogènes. Parmi ces derniers, les pucerons des céréales dont les plus répandus sont *Rhopalosiphum padi*, *R. maidis*, *Sitobion avenae* et *Schizaphis graminum*, espèces vectrices de plusieurs phytovirus augmentant les risques de baisse des rendements.

Etant donné les dégâts provoqués par ces espèces de pucerons, différentes mesures (physiques, chimiques, biologiques) ont été entreprises pour les éliminer ou limiter leur pullulation. Cependant ces moyens n'ont pas toujours connu le succès attendu, et dans certains cas, la situation a été aggravée. Ainsi, les agents chimiques (les insecticides en particulier) utilisés peuvent conduire à l'apparition de résistance et avoir un impact sur l'environnement, dans la mesure où leur action ne cible pas à une seule espèce. Par ailleurs, la sélection de variétés résistantes et/ou tolérantes à ces pucerons semble être le moyen le plus efficace sauf que certaines populations de pucerons ont pu contourner cette résistance génétique.

Les récents progrès en génétique ont permis la mise au point de nouveaux moyens de lutte contre les insectes ravageurs. Ainsi, les éléments transposables, séquences autonomes mobiles au sein des génomes, sont considérés d'une part comme des facteurs puissants intervenant dans la plasticité des génomes voire dans le traçage de l'histoire évolutive des organismes et, d'autre part comme des outils exploités en biotechnologie pour effectuer la mutagenèse et la transgénèse. Par conséquent, il est ainsi intéressant de chercher à isoler de telles séquences potentiellement actives afin de les utiliser en tant que vecteur de transfert de gènes chez différentes espèces.

Dans ce contexte, les éléments transposables à ADN sont particulièrement utilisés. Les premiers essais de transformation des lignées germinales ont commencé chez *Drosophila melanogaster* avec l'élément *P* (Rubin et Spradling 1982; Spradling et Rubin 1982) puis avec l'élément *hobo* (Blackman *et al.* 1989). Cependant, ces transposons ne sont pas actifs chez les non *Drosophilidae* (Handler *et al.* 1993). Ceci a donc orienté les recherches vers l'identification de nouveaux éléments transposables ayant un spectre de transformation plus large tel que les éléments appartenant aux deux superfamilles *Tc1-mariner* et *piggyBac*. Les éléments de cette dernière, en particulier, ont été utilisés en tant que vecteurs de transfert de gènes dans le domaine de l'agriculture (Handler 2002) ainsi que dans le domaine de la santé humaine

(Robinson *et al.* 2004). Il en est de même pour *mos1* et *Himar1*, éléments de la famille *mariner*, qui sont exploités comme vecteur de transgénèse chez plusieurs espèces (Wang *et al.* 2000; Keravala *et al.* 2006).

Dans cette optique, mon objectif était de répondre aux questions suivantes :

- Les éléments transposables *mariner* (*mariner-like element* ou *MLE*) et *piggyBac* sont-ils présents dans les génomes des pucerons des céréales ?
- S'ils sont présents, quelles sont leur structure et leur fonctionnalité (*e.g.* existent-ils des transposons potentiellement actifs ?) et comment sont-ils classés par rapport aux autres éléments de la même superfamille ?
- Existent-ils chez les céréales associées ?

Pour cela, des outils bio-informatique ont été utilisés pour balayer les génomes et rechercher des éléments *mariner*. Toutefois, vu l'absence de génome complet séquencé chez les pucerons des céréales objets de notre recherche, trois pucerons proche phylogénétiquement et appartenant à la même tribu des *Macrosiphini* ont été pris comme référence : le premier est le puceron vert du pois *Acyrthosiphon pisum*, dont le génome de 541Mb a été publié en 2010 (Richards *et al.* 2010), le deuxième est le puceron russe du blé *Diuraphis noxia*, dont le génome de 393Mb a été publié en 2015 (Nicholson *et al.* 2015) et le troisième est le puceron vert du pêcher *Myzus persicae*, dont le génome assemblé (398Mb) a été récemment publié dans la base de données « AphidBase ». Ceci a permis l'identification, la caractérisation, la classification et la comparaison des éléments détectés dans ces trois génomes.

Ces données obtenues *in silico* ont été ensuite utilisées pour rechercher ces éléments transposables *in vitro* chez les principales espèces de pucerons des céréales *R. padi*, *R. maidis*, *S. avenae* et *S. graminum*, ainsi que chez les plantes hôtes associées orge, blé, avoine et égilope. Des analyses *in silico* effectuées dans les génomes publiés des céréales complètent ces recherches afin de caractériser ces éléments et de mettre en évidence l'existence d'éventuels transferts horizontaux de ces éléments entre les pucerons et leurs plantes hôtes.

Pour la superfamille des *piggyBac*, la recherche des séquences identifiées chez les eucaryotes et disponibles dans les banques de données (NCBI, RepBase), a permis d'enrichir considérablement le répertoire de cette superfamille. Leur caractérisation fine, l'étude de leur origine, leur distribution, leur structure et leur évolution sera très utile pour la recherche de ces éléments dans plusieurs génomes, entre autre ceux des pucerons des céréales.

Outre l'impact de ces éléments sur la dynamique et la plasticité des génomes, l'ensemble des résultats de ce travail pourra être utilisé pour proposer des pistes afin d'élaborer de nouvelles stratégies de protection des céréales cultivées et le contrôle de leurs ravageurs.

Introduction Bibliographique

La céréaliculture constitue l'une des premières activités agricoles. Bien que les céréales aient été initialement exploitées à des fins médicales et de tissages, elles fournissent actuellement une alimentation régulière et abondante aussi bien pour l'Homme que pour le bétail.

C'est au début du Néolithique, il y a environ 10.000 ans, que l'Homme a commencé à cultiver les céréales en pratiquant la domestication d'espèces sauvages. Le croissant fertile (Moyen-Orient), l'Amérique Centrale et l'Asie du Sud-Est sont les berceaux de la majorité des espèces céréalières consommées actuellement dans le monde entier.

La sélection artificielle des caractères qui facilitent la culture, la récolte et l'utilisation des espèces cultivées a été effectuée en exploitant la diversité naturelle des variétés spontanées. Par exemple, l'engrain (*Triticum boeoticum*) et l'amidonnier (*Triticum dicoccoides*) sont les céréales sauvages à l'origine du blé. De même, l'orge et le maïs ont été obtenus respectivement par domestication de l'orge sauvage (*Hordeum spontaneum*) et de la téosinte (*Zea mays ssp parviglumis*) (Anderson-Gerfaud *et al.* 1991). Cette domestication s'est exercée parmi les espèces sauvages aptes à cette transformation. Toutefois, très peu d'espèces domestiquées sont présentes parmi l'immense famille des céréales qui compte plus de 700 genres. On peut citer essentiellement le blé tendre, le blé dur, le blé poulard, l'épeautre, l'orge, le seigle, l'avoine, le maïs, le millet, le sorgo et le riz. Ces céréales sont consommées sous diverses formes : grains mais aussi semoule, boulgour, farine, flocon, fécules qui correspondent à divers état de la granulométrie de la graine d'origine.

En raison de la faible teneur en eau des grains, qui facilite leur transport et leur stockage, ainsi que la richesse en amidon, les céréales ont constitué une importante ressource alimentaire pour plusieurs civilisations. Ceci a favorisé le développement de populations denses, d'économies de commerce et de structures administratives. Ainsi, les populations Moyen-Orientales puis Européennes se sont construites autour du blé et de l'orge, celles de l'Extrême-Orient autour du millet, du riz et du blé, celles des Amérindiens autour du maïs, tournesol et quinoa, celles d'Afrique autour du millet, du sorgho et du riz (Diamond 2002).

Depuis les débuts de leur domestication, l'importance des céréales a pris une telle ampleur dans le secteur économique que de nombreuses innovations techniques ont été développées (notamment sélection-hybridation et transformation génétique) pour améliorer les rendements voire lutter contre divers facteurs biotiques et abiotiques. Toutefois, la production reste variable et dépend essentiellement de nombreux facteurs climatiques et biotiques.

I. Les céréales : sous-famille des *Pooideae*

1. Systématique

Les *Pooideae* appartiennent au clade des angiospermes (plantes à fleurs). Ce sont des monocotylédones de la famille des *Poaceae*. Ils incluent environ 4200 espèces, réparties en quatre tribus majeures: *Triticeae, Aveneae, Poeae* et *Brachypodieae* (Löve 1984; Kellogg 2015). De nombreuses espèces ont été domestiquées et cultivées. Parmi elles, figurent :

- Des céréales à pailles : le blé tendre *Triticum aestivum*, le blé dur *T. turgidum* ssp. *durum*, l'orge *Hordeum vulgare*, le seigle *Secale cereale* et l'avoine *Avena sativa*.
- De nombreuses espèces utilisées comme fourrage : l'égilope *Aegilops tauschii*, le brachypode à deux épis *Brachypodium distachyon.*

2. Evolution des génomes

De par leur adaptabilité à des environnements différents, cette sous-famille est caractérisée par une forte variabilité génomique qui touche aussi bien le nombre et la taille des chromosomes que le niveau de ploïdie et la composition en gènes (Keller et Feuillet 2000). L'intérêt économique d'un grand nombre d'espèces de *Poaceae* en a fait la famille la plus étudiée du point de vue génomique. Les premières études de génomique comparative ont permis de déterminer les chromosomes ou segments partagés entre ces espèces (synténie) et de montrer une forte conservation de la position et de l'ordre des marqueurs (Gale et Devos 1998; Abrouk *et al.* 2010). Conservation et variabilité suggèrent que les espèces de cette sous-famille ont divergé il y a 35 Millions d'années (Mya) (Moore *et al.* 1995; Prasad *et al.* 2005).

Le séquençage des génomes de plusieurs *Poaceae* a permis une meilleure compréhension de leur évolution génétique. Ces espèces se différencient par :

- Le génome dont ils sont issus : H = *Hordeum*, A = *T. monococcum*, B = *Aegilops* (*bicornis* ou *longissima* ou *searsii*…), D = *T. tauschii*, S = *A. speltoides*.
- Leur degré de ploïdie (exemple orge diploïde : HH ; blé tétraploïde : AABB).
- Le nombre de chromosomes (14, 28, 42) par exemple le blé dur ne contient que les deux génomes AA, BB constitués chacun de sept paires de chromosomes homéologues (A1...A7, B1…B7) soit au total 28 chromosomes.

Spécifiquement chez les *Pooideae,* quatre espèces ont été séquencées et sont publiquement disponibles :

- *B. distachyon* 2n = 10 avec une taille de 272 Mb (Vogel *et al.* 2010),
- *H. vulgare* 2n = HH = 14 à 5Gb (International Barley Genome Sequencing Consortium 2012),
- *T. aestivum* 2n = AABBDD = 42 à 17 Gb (Brenchley *et al.* 2012; International Wheat Genome Sequencing Consortium 2014),
- *A. tauschii* 2n = DD = 10 à 4,3Gb (Jia *et al.* 2013).

La comparaison des séquences des gènes orthologues d'Acetyl-CoA carboxylase (*Acc-1* et *Acc-2* chloroplastiques), présents en une seule copie, a permis la datation des divergences entre les espèces (Huang *et al.* 2002). Plus tard, Chalupska *et al.* (2008) ont estimé la divergence entre les *Panicoideae* (maïs / sorgho) et les *Ehrhartoideae* (riz) à 60 Mya, celle du blé avec le riz à 50 Mya, avec l'avoine à 31 Mya, avec l'orge à 11,6 Mya et avec le seigle à 6,7 Mya (**Figure 1**). Quant aux différentes espèces de *Triticum sp* portant les génomes A, B et D, elles ont divergé plus récemment, il y a 2,5 à 4 Mya.

Ainsi, bien que ces espèces partagent les mêmes événements ancestraux de polyploïdisation (on parle de paléopolyploïdie), la comparaison de la taille des génomes montre une grande diversité. Cette diversité de taille peut être expliquée par deux forces majeures :

- Les événements récents d'allopolyploïdisation (**Figure 2**) : les chromosomes viennent d'espèces différentes mais suffisamment proches pour s'hybrider, on parle de chromosomes homéologues. C'est le cas des génomes de l'amidonnier sauvage *T. turgidum ssp dicoccoides* (2n = AABB = 28), ou des espèces modernes de blé domestiqués tétraploïdes (*T. turgidum ssp durum* et *T. turgidum spp dicoccum* 2n = AABB = 28) et hexaploïdes (*T. aestivum* 2n = AABBDD = 42) (Feldman *et al.* 1995; Blake *et al.* 1999).
- La prolifération différentielle des éléments transposables (ETs) : en comparant les tailles des génomes et leur teneur en ETs. Par exemple, les espèces de *Pooideae* ayant un grand génome (>2 Gb) comme le blé, l'aegilops et l'orge, possèdent un plus grand nombre d'ETs que les petits génomes (<800Mb) tel que *Brachypodium* (Charles *et al.* 2008; Jia *et al.* 2013; Mazaheri *et al.* 2014). De plus, les familles d'ETs identifiées dans les différentes espèces ne sont pas les mêmes (Pritham 2009), suggérant une invasion indépendante des ETs (voir plus loin dans la partie III de ce chapitre : Les éléments transposables chez les Eucaryotes).

Les études portant sur l'évolution des génomes de ces différentes espèces ont contribué à l'amélioration de l'adaptation de ces cultures intimement liées à la civilisation humaine.

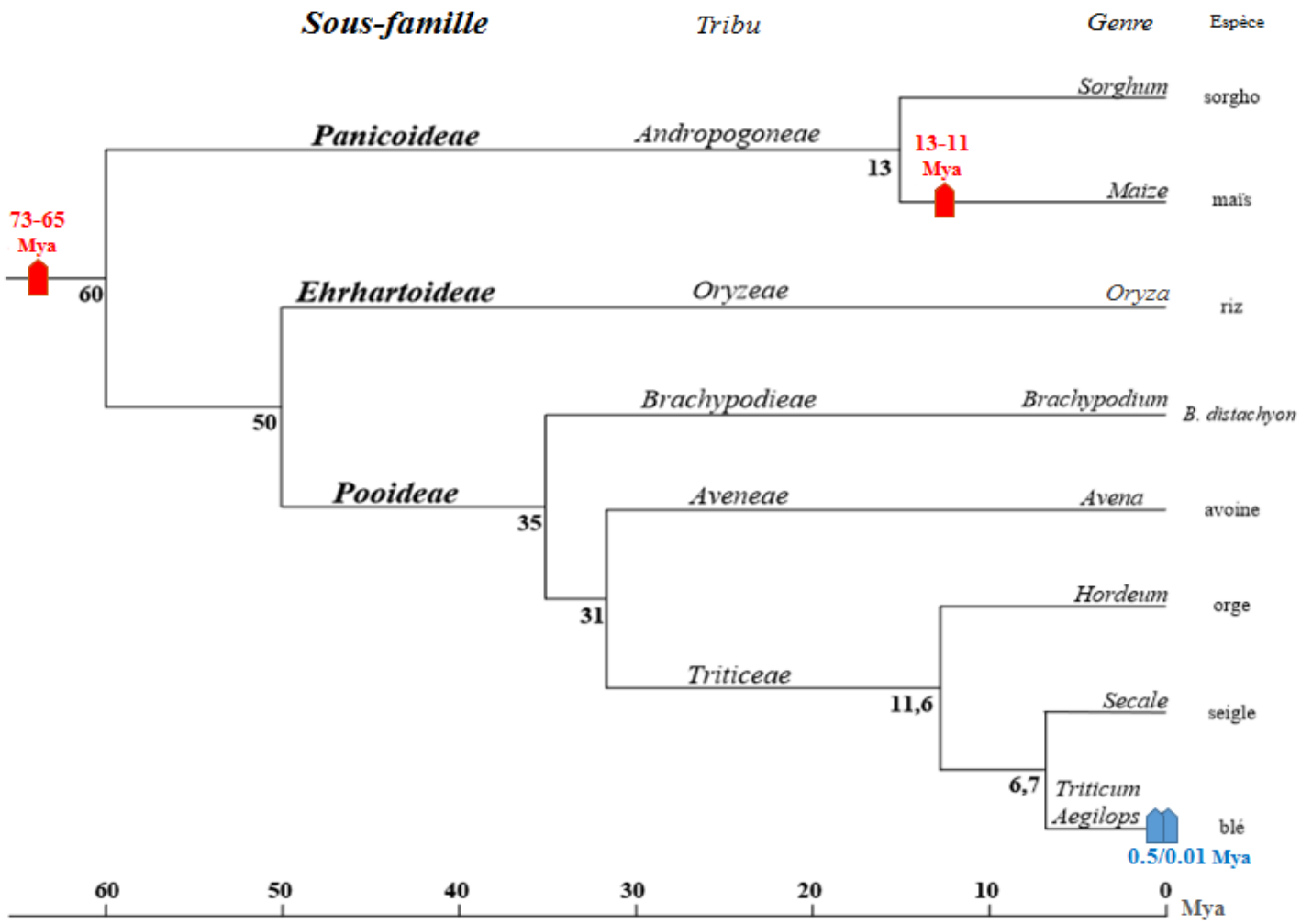


Figure 1. Divergence des espèces de la famille des *Poacées* en Million d'années (d'après Chalupska *et al.* 2008).

Cet arbre a été établi à partir de la comparaison des séquences des gènes orthologues d'acétyl-CoA carboxylase (*Acc-1* et *Acc-2*), présents en une seule copie. Les événements anciens et récents de polyploïdisation du blé sont indiqués en rouge et en bleu respectivement.

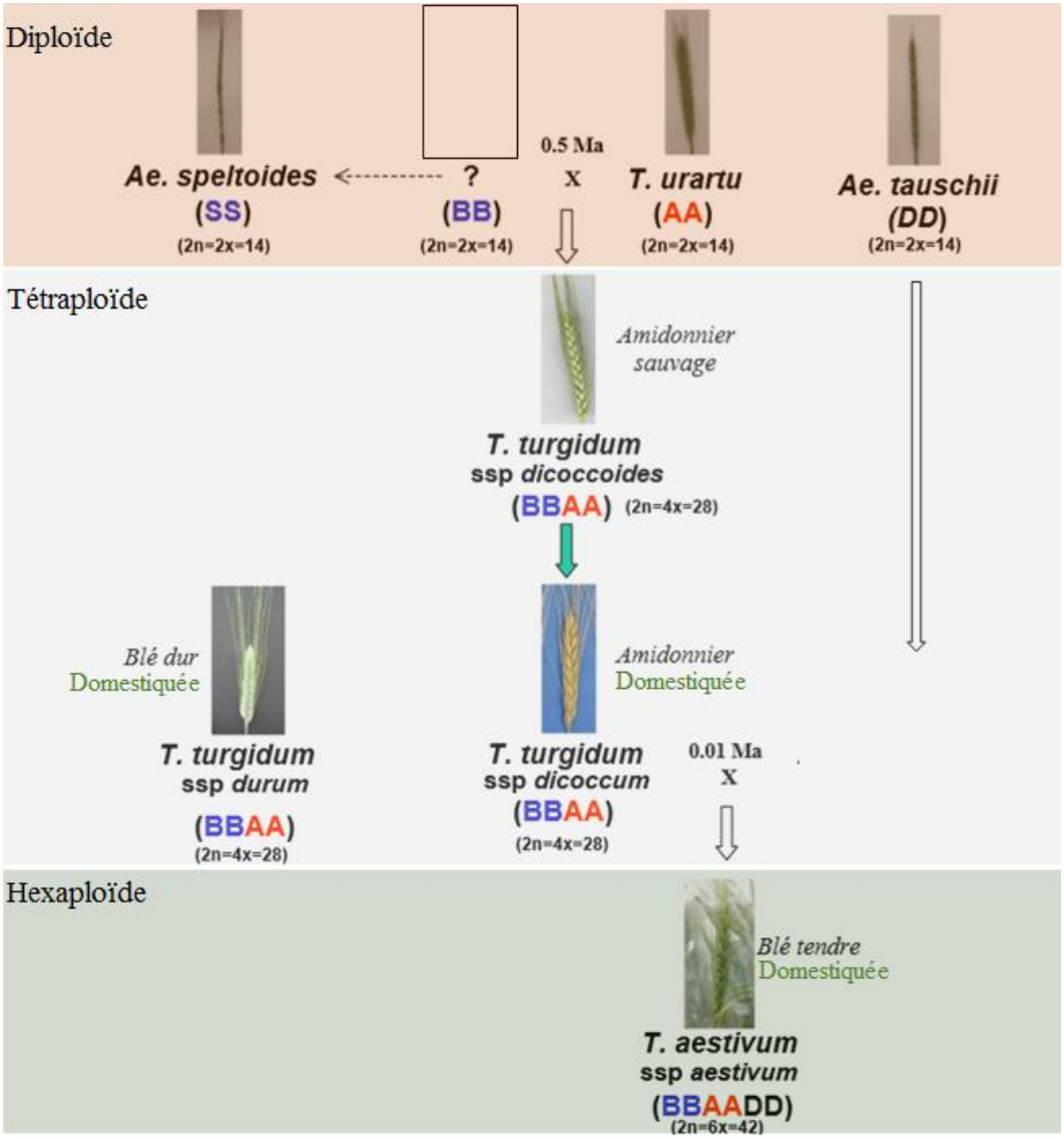


Figure 2. Polyploïdisation et domestication des blés sauvages engendrant la formation des blés cultivés tétraploïdes et hexaploïdes.

Bien que le donneur du génome B reste inconnu, les progéniteurs des génomes A et D sont clairement identifiés. (Feldman *et al.* 1995; Blake *et al.* 1999; Dvorak *et al.* 2006). En effet, un événement d'allopolyploidisation entre l'espèce diploïde *T. urartu* progénitrice du génome A et une espèce progénitrice du génome B qui reste à identifier, proche de *A. speltoides*, a formé, il y'a environ 0.5 Mya, l'amidonnier sauvage tétraploïde (BBAA). La sélection des plantes dont les épis ne perdent pas leurs grains avant la récolte, a abouti à la domestication de l'amidonnier *T. turgidum spp dicoccum* (BBAA) venant de l'amidonnier sauvage. L'hybridation de ce dernier avec l'espèce diploïde sauvage *Aegilops tauschii* (DD) a donné le blé tendre allohexaploide *T. aestivum ssp. aestivum* (BBAADD). Par ailleurs, l'origine du blé dur *T. turgidum ssp. durum* (BBAA) et sa relation avec les premières espèces domestiquées tétraploïdes restent mal élucidées.

3. La céréaliculture

La surface cultivée en céréales dans le monde est estimée à 720 millions d'hectares (Mha), soit 51 % des terres arables, 14.6 % de la surface agricole mondiale et 5.5 % des terres émergées du monde (FAOstat 2014, http://faostat.fao.org/). Le maïs est la céréale la plus produite au monde et représente 40% de la production mondiale, suivi du blé 29%, du riz blanchi 19% et de l'orge 6%. Les trois plus grands pays producteurs de céréales sont la Chine avec 492,9 millions de tonnes (Mt), suivie des Etats-Unis d'Amérique avec 439,4 Mt et de l'Europe avec 329,1 Mt (USDA campagne 2014-2015, www.usda.gov - **Figure 3**).

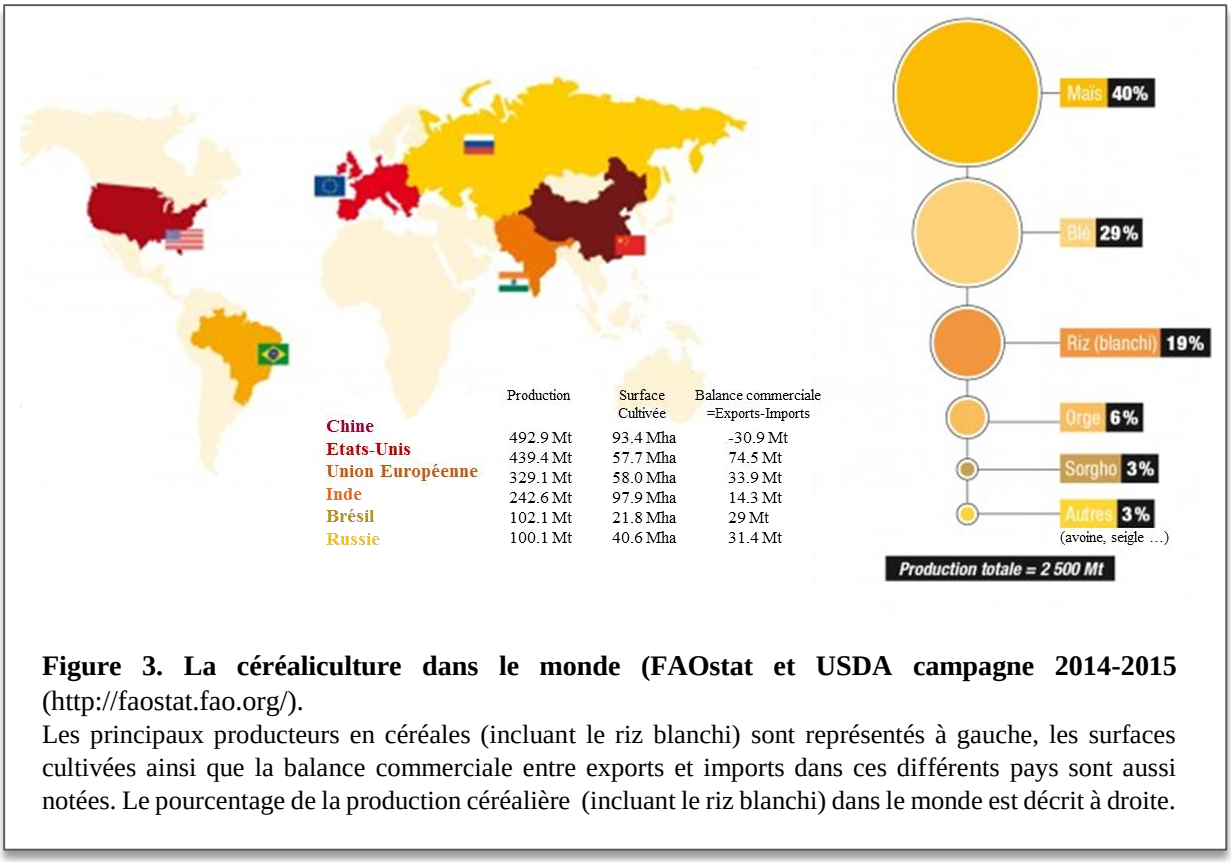


Figure 3. La céréaliculture dans le monde (FAOstat et USDA campagne 2014-2015 (http://faostat.fao.org/).
Les principaux producteurs en céréales (incluant le riz blanchi) sont représentés à gauche, les surfaces cultivées ainsi que la balance commerciale entre exports et imports dans ces différents pays sont aussi notées. Le pourcentage de la production céréalière (incluant le riz blanchi) dans le monde est décrit à droite.

En Tunisie, la production est d'environ 2.35 Mt pour une surface moyenne de 1,4 Mha (FAOstat 2014, http://faostat.fao.org/). Toutefois, bien qu'occupant près du tiers de la surface agricole utile et répartie essentiellement au nord du pays, les céréales ne contribuent en moyenne qu'à hauteur de 13 % à la valeur ajoutée agricole (Office National des Céréales ONC 2014). Les céréales sont représentées essentiellement par le blé dur soit 49% des superficies réservées, l'orge 40%, le blé tendre 10% et les autres triticales ne dépassant pas 1% de la superficie. De plus, l'avoine est une culture bien enracinée dans la tradition des petits exploitants pour l'utiliser comme foin. C'est le fourrage prédominant, il occupe des superficies quasiment stables d'environ 184.000 ha par an, soit 70 % environ des superficies annuelles en fourrages (ONC 2014).

Par ailleurs, les rendements sont fort variables d'une année à l'autre et dépendent d'une part des facteurs abiotiques, telles que les conditions climatiques ou pédologiques (Lacroix 2002; Djili *et al.* 2003; Slama *et al.* 2005) et d'autre part des facteurs biotiques représentés par divers organismes tels que les virus, les bactéries, les champignons, les nématodes ou les insectes particulièrement les pucerons, responsables d'une baisse considérable des rendements.

4. Les pathogènes et les ravageurs des céréales

Plusieurs organismes nuisibles affectent la croissance des céréales et par conséquent, leur rendement, conduisant à des pertes économiques considérables.

4.1. Les agents pathogènes des céréales

4.1.1. Les virus

Les phytovirus, parasites obligatoires, provoquent des maladies qui se traduisent par des perturbations métaboliques entrainant différents symptômes tels que chloroses, nécroses, enroulement des feuilles, rabougrissement des plante. Généralement, la transmission du virus se fait horizontalement par l'intermédiaire d'un vecteur, le plus souvent un insecte. Ce dernier prélève le virus à partir d'une plante infestée en se nourrissant de la sève et l'inocule dans une nouvelle plante saine assurant ainsi sa dissémination spatiale (Lapierre et Signoret 2004). Les virus des céréales les plus fréquemment cités dans la littérature sont :

- B/WSMV = *Barley/ Wheat Stripe Mosaic virus* (Lin et Langenberg 1984).
- B/CYDV = *Barley/Cereal Yellow Dwarf virus* (Oswald et Houston 1951).
- SBCMV = *Soil-Borne Cereal Mosaic virus* (Canova et Quaglia 1960).

4.1.2. Les bactéries

Plusieurs bactérioses entrainent une désorganisation profonde des systèmes radiculaires et aériens. *Xanthomonas campestris pv. transluscens*, responsable des glumes noires des céréales ou brûlures, est soit transmise par la semence, soit par un inoculum provenant d'hôtes alternatifs ou de débris. Cette bactérie se développe de manière épiphyte sur les feuilles et remonte les étages foliaires, véhiculée par les éclaboussures d'eau jusqu'à l'épi (Cunfer et Scolari 1982).

Par ailleurs, une autre bactérie *Pseudomonas syringae pv. atrofaciens* peut causer des symptômes proches de ce dernier. Toutefois, l'infection débute généralement à la base de la glume et de manière plus intense à l'intérieur de la plante. Il existe d'autres brûlures touchant les feuilles et les épis des céréales. Elles sont provoquées par des pathovars différents, à savoir

P. syringae pv. syringae ou encore *P. syringae pv. striafaciens* qui entrainent des stries au niveau des feuilles d'avoine (Pasichnyk 1999; Lapierre et Signoret 2004).

4.1.3. Les champignons

Il existe un nombre considérable de champignons phytopathogènes chez les céréales. Ce sont des parasites qui infestent les plantes sauvages et cultivées, causant des symptômes variés tels que les pourritures, les nécroses, les chancres ou les stries. Les champignons peuvent être transmis verticalement par propagation de graines infestées ou par des résidus de paille (Lacroix 2002; Lapierre et Signoret 2004). Parmi les espèces les plus néfastes, on peut citer :

- *Fusarium sp* responsables de la fonte des semis et de la pourriture du collet, tige, racine et semences, conduisant à la mort des plantules avant leur levée. Certaines espèces de *Fusarium* sont connues pour leur capacité de synthétiser des mycotoxines.
- *Zymoseptoria tritici* (anamorphe de *Septoria tritici*) se manifeste par des tâches foliaires ou septoriennes essentiellement sur le blé.
- *Puccinia sp* responsables de la rouille des feuilles ou des tiges.
- *Phaeosphaeria nodorum* ainsi que *P. avenaria* qui engendrent des tâches sur les feuilles, les fruits et des chancres de la tige.

4.2. Les ravageurs des céréales

4.2.1. Les nématodes

Ces parasites sont des vers ronds non segmentés, susceptibles d'interférer avec la croissance des plantes et difficiles à contrôler. Les attaques de nématodes sont observables sur les parties aériennes comme sur les parties souterraines. En surface, l'infection se traduit par un jaunissement des feuilles, un tallage réduit voire un arrêt de la croissance (Kerry et Crump 1998). Pour les parties souterraines, les racines sont flétries et leur couleur est altérée (Nicol et Rivoal 2008). Les principales espèces de nématodes parasites sont *Meloidogyne*, *Heterodera* et *Pratylenchus* (Kerry et Crump 1998; Nicol et Rivoal 2008).

4.2.2. Les insectes

De nombreuses espèces d'insectes sont des déprédateurs des céréales et s'attaquent aux divers stades de croissance de la plante. Ils représentent une contrainte majeure suite aux dégâts directs et indirects qu'ils occasionnent. Les principales espèces d'insectes ravageurs des céréales appartiennent à plusieurs ordres (Miller et Pike 2002) :

- l'ordre des diptères tel que les cécidomyies du genre *Mayetiola* ou les mineuses du genre *Agromyza* ;

- l'ordre des coléoptères tel que le charançon du genre *Sitophilus* ;
- l'ordre des lépidoptères tel que la teigne des céréales du genre *Sitotroga*, la pyrale du maïs du *genre Ostrinia* ou la pyrale de la farine *Ephestia*, et les foreurs des tiges de céréales et de maïs des genres *Busseola* et *Sesamia* ;
- l'ordre des orthoptères tel que le criquet pèlerin du genre *Schistocerca* ou le criquet migrateur du genre *Locusta* ;
- l'ordre des hémiptères tel que les pucerons.

Durant cette étude, nous nous sommes focalisés sur les quatre espèces de pucerons des céréales les plus répandues, à savoir *Rhopalosiphum padi*, *Rhopalosiphum maidis*, *Schizaphis graminum* et *Sitobion avenae* (Blackman et Eastop 2000).

II. Les pucerons des céréales

1. Systématique

Les espèces de pucerons des céréales appartiennent à l'Embranchement des Arthropodes, Classe des Insectes, sous-classe des Ptérygotes (insectes possédant des ailes au stade adulte), ordre des Hémiptères (caractérisés par deux paires d'ailes, dont l'une est transformée en hémiélytre, et d'un rostre articulé protégeant un stylet piqueur), famille des Aphididae et sous-famille des Aphidinae (Blackman et Eastop 2000). Les principales espèces de pucerons des céréales sont réparties en deux tribus (**Figure 4**) :

- Tribu des Aphidini regroupant *Rhopalosiphum padi*, *Rhopalosiphum. maidis* et *Schizaphis graminum*.
- Tribu des Macrosiphini représenté par *Sitobion avenae*.

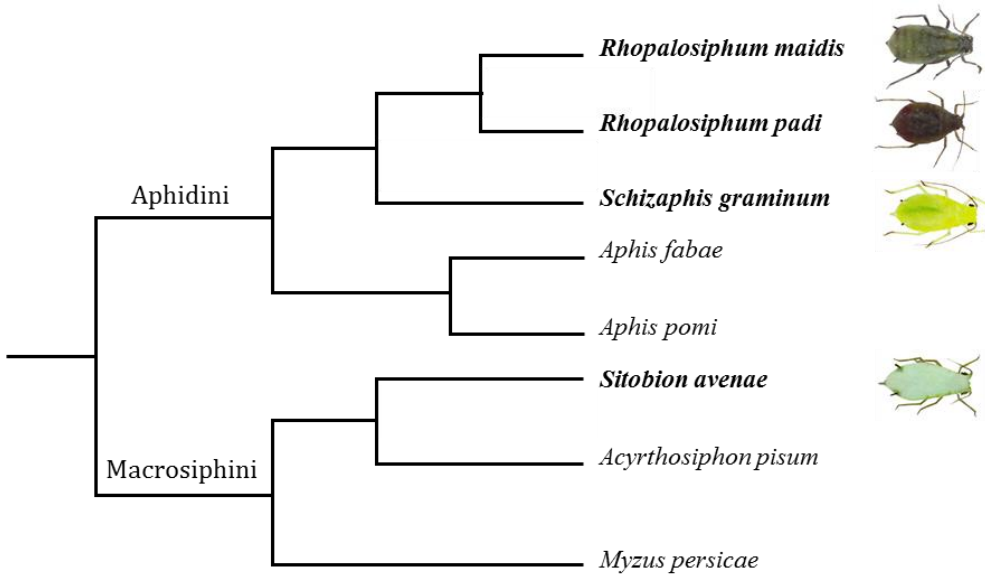


Figure 4. Représentation de la phylogénie des pucerons appartenant aux deux tribus des *Aphididae* (adapté à partir de Foottit *et al.* 2008)
Quelques exemples de membres des tribus des *Aphidini* et *Macrosiphini* incluant les quatre espèces étudiées dans ce travail. La phylogénie est basée sur la séquence du gène mitochondrial de la cytochrome oxydase I (COI).

2. Caractéristiques chromosomiques

A ce jour, les génomes entiers de ces quatre espèces n'ont pas encore été séquencés. Toutefois, les analyses cytogénétiques ont permis de déterminer le nombre de chromosomes de chacune d'elle (Blackman et Eastop 2000), à savoir : 2n = 8 pour *R. padi* et *S. graminum*, 2n = 18 pour *S. avenae*.

Concernant *R. maidis*, le caryotype du puceron varie en fonction de la plante hôte, suggérant l'existence de races d'hôte de caryotypes différents. Ainsi, les populations qui s'attaquent à l'orge sont 2n = 10, alors que celles qui vivent sur le sorgho et le maïs sont à 2n=8 chromosomes (Brown et Blackman 1988; Blackman et Eastop 2000). Cette variation peut être due à la présence d'origines multiples de la population sexuée ou à des mutations au sein des lignées parthénogénétiques (Van Emden et Harrington 2007), voire à des événements survenus au cours de la spéciation sur un hôte ou même à des effets de dérive génétique.

3. Cycle biologique

Chez les pucerons, il existe deux types de cycle de vie (Williams et Dixon 2007):

- Le premier est appelé holocyclique. De la fin de l'hiver à la fin de l'été, les colonies de pucerons ne sont composées que de femelles parthénogénétiques. Il s'agit d'un mode de reproduction sans fécondation, c'est-à-dire que les nouvelles femelles sont des clones de leur mère. A la fin de l'été et en conditions défavorables, des femelles sexupares apparaissent et donnent naissance à des mâles et des femelles fécondables. Durant l'automne, la fécondation s'effectue. Les femelles fécondables pondent début de l'hiver. Les œufs éclosent en fin d'hiver et des femelles fondatrices en émergent. Les femelles fondatrices sont les premières femelles parthénogénétiques qui sont à l'origine des colonies printanières (**Figure 5**).
- Le deuxième cycle, plus simplifié, est appelé anholocyclique puisque la phase de la reproduction sexuée est inexistante (**Figure 5**). Les pucerons ne se reproduisent que par parthénogénèse tout au long de l'année.

En général, pour la plupart des espèces de pucerons, il y a une fraction de la population qui est holocyclique et une autre fraction anholocyclique. En outre, les pucerons sont capables de produire des adultes ailés et des aptères (Williams et Dixon 2007).

Par ailleurs, le cycle biologique des pucerons peut s'effectuer sur une même espèce hôte (monoécique) ou sur deux espèces différentes (dioécique). Ainsi, la ponte des œufs s'effectue sur un « hôte primaire » généralement une espèce arbustive ou arborescente, puis le reste du cycle se déroule sur un « hôte secondaire » (Blackman et Eastop 2000).

Plus spécifiquement, chez *R. padi*, certaines lignées ont un cycle holocyclique dioecique ayant comme hôte primaire, le merisier à grappe, et d'autres se développent de manière anholocyclique sur le blé, l'orge, l'avoine, le sorgho et le maïs.

Les populations de *R. maidis* ont un cycle strictement anholocyclique sur le blé, l'orge, l'avoine, le sorgho, la canne à sucre et le maïs.

En revanche, *S. avenae* et de *S. graminum* sont généralement holocyclique monoecique sur graminées. Leurs hôtes sont le blé, l'orge, l'avoine, le seigle, le sorgho, le riz et le maïs (Blackman et Eastop 2000).

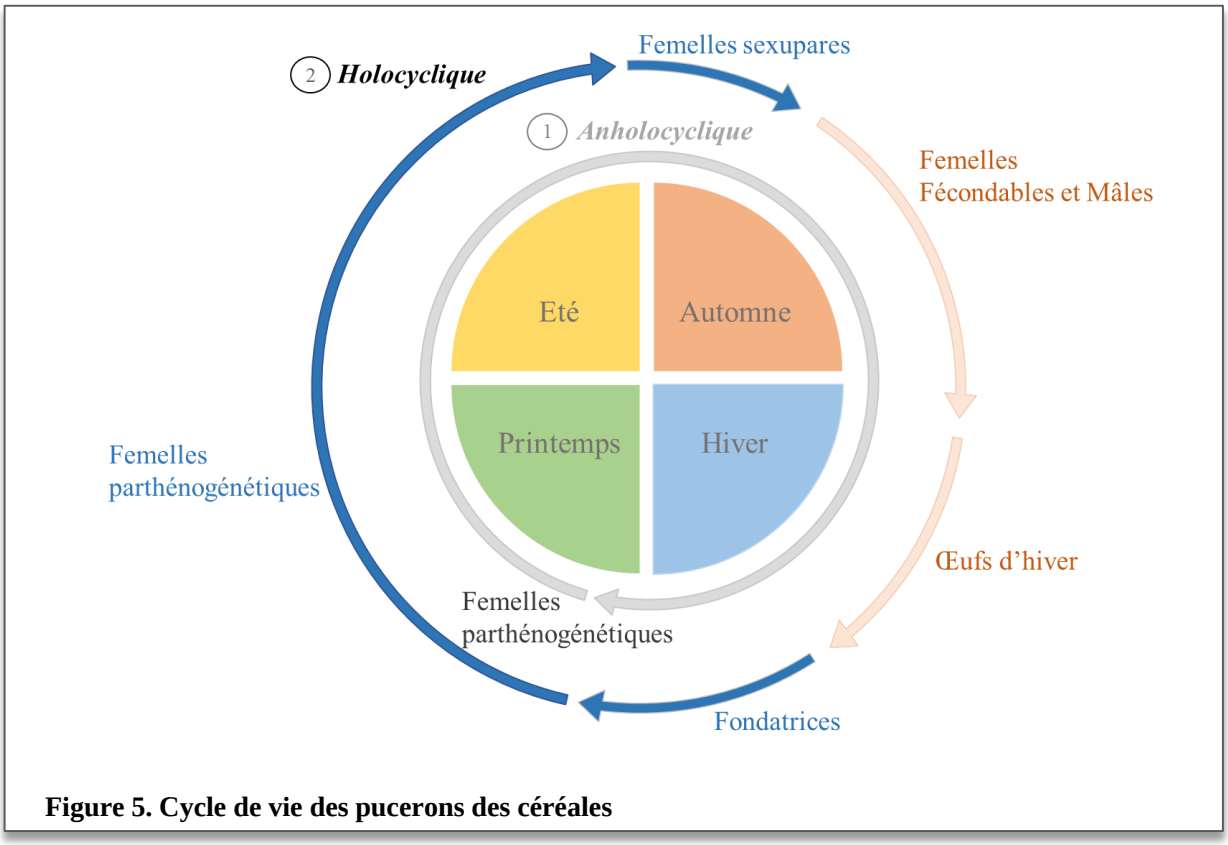


Figure 5. Cycle de vie des pucerons des céréales

4. Dégâts des pucerons

La succion de la sève élaborée par les colonies de larves et d'adultes de pucerons entraine des dégâts directs se traduisant par une diminution de la vigueur de la plante (affaiblissement et flétrissement général) et par des réactions physiologiques provoquées par la sécrétion salivaire menant à la chlorose et la nécrose (Ryan *et al.* 1990; Miller et Pike 2002). Les dégâts indirects sont causés par :

- le développement de maladies à champignon comme la fumagine. En effet, le miellat collant très riche en sucre, est excrété par les cornicules des pucerons favorisant le développement de diverses espèces de champignons ascomycètes, ectophytes et saprophytes (*i.e. Capnodium oleaginum*, *Fumago salicina*). Une couche noire se forme sur les feuilles et réduit la photosynthèse (Huang *et al.* 1981).

- la transmission de maladies virales. Le virus le plus préoccupant est celui de la jaunisse nanisante de l'orge et des céréales B/CYDV présentant plusieurs sérotypes qui infestent une large gamme de céréales (Oswald et Houston 1951; Miller et Pike 2002).

5. Stratégies de lutte

Afin de limiter la pullulation et minimiser l'impact des pucerons des céréales, plusieurs méthodes de lutte contre ces ravageurs ont été développées.

Les pratiques culturales telles que la suppression de plantes réservoirs ou l'utilisation de paillage réfléchissant pour empêcher l'arrivée des pucerons, sont intéressantes dans la mesure où elles n'impliquent pas des couts supplémentaires, toutefois elles ne sont pas suffisantes pour maitriser ces ravageurs (Brault *et al.* 2001).

La lutte chimique par l'emploi d'insecticides est le principal moyen de lutte utilisé à grande échelle contre les pucerons. La pulvérisation d'insecticides organophosphorés et des pyréthroïdes sur les parcelles pose à long terme des problèmes dus au développement de pucerons résistants et à la dégradation de la biocénose (Charbonnier *et al.* 2016).

Par ailleurs, certaines substances naturelles agissent en tant que bioinsecticide telles que les endotoxines *Cry,* insecticides produits par la bactérie *Bacillus thuringiensis* (Bt), qui ont toutefois montré une faible efficacité contre les pucerons (Chougule *et al.* 2013). En revanche, l'efficacité aphicide de certaines neurotoxines de scorpions et d'araignées a été démontrée dans la gestion des pucerons (Chougule et Bonning 2012).

La lutte biologique, quant à elle, consiste à utiliser des prédateurs naturels notamment : des coccinelles (Iperti 1999), des chrysopes (Duelli 2001) ou des diptères (Volkl *et al.* 2007). Les parasitoïdes constituent une autre alternative, il s'agit de plusieurs espèces de hyménoptères qui pondent et se développent dans les larves des pucerons (Traugott *et al.* 2008; Vollhardt *et al.* 2008).

Pour d'autres espèces, la technique de l'insecte incompatible (TII) est utilisée. Elle consiste à introduire dans une population une bactérie endosymbiotique du genre *Wolbachia* qui va entrainer une incompatibilité cytoplasmique (IC). Ce phénomène cause la mort précoce des œufs issus d'un croisement entre un mâle infecté et une femelle non infectée. En revanche, la femelle porteuse va transmettre la bactérie à sa descendance, quel que soit le mâle et permet la propagation rapide de la bactérie au sein de la population hôte. Cette technique a été utilisée

chez *Culex pipiens* (Laven 1967), *Ceratitis capitata* (Sarakatsanou *et al.* 2011), *Aedes albopictus* (Zhang *et al.* 2015).

La technique de l'insecte stérile a également été développée chez plusieurs espèces d'insectes. Elle consiste à l'irradiation aux rayons ionisants (*i.e.* gamma) uniquement des mâles élevés au laboratoire, préservant leur vigueur et leur performance sexuelle. Le lacher de ces mâles devenus stériles est éffectué en quantité supérieure à celle des mâles sauvages (Klassen et Curtis 2005) permettant ainsi de controler les populations de ravageurs. Une autre alternative à cette technique a été décrite, notamment chez *Anopheles gambia*, où pour assurer la reproduction, le mâle produit à partir d'une transglutaminase AGAP009099 et son substrat enrichi en glutamine *Plugin*, une protéine coagulée MAG qui bloque le liquide séminal à l'intérieur de la cavité d'accouplement de la femelle. Il a été démontré que l'utilisation d'ARN double brins ciblant le gène AGAP induit une réduction significative de son expression et par conséquent la stérilité (Rogers *et al.* 2009).

D'autres travaux récents portant sur la modification génétique des diptères (*e.g.* mâles stériles de *Bactrocera oleae* modifiés génétiquement par un vecteur recombinant comprenant les TIR des *piggyBac*), ont révélé que les éléments transposables sont d'excellents vecteurs de transfert de gènes en vue de développer des méthodes de contrôle plus efficaces (Koukidou *et al.* 2006; Ant *et al.* 2012).

III. Les éléments transposables chez les Eucaryotes

1. Histoire et définition

Le génome a longtemps été considéré comme étant relativement stable (aux mutations ponctuelles près) et évoluant lentement. Cette vision a été bouleversée dans les années 1940, par les travaux de Barbara McClintock portant sur des instabilités phénotypiques chez le maïs. Ces analyses ont montré que « des éléments de contrôle » *Ac* (Activateur) et *Ds* (Dissociation) sont capables de réguler l'expression des gènes responsables de la pigmentation des grains de l'épi (McClintock 1950; McClintock 1951). La découverte de ces séquences, a donc remplacé le paradigme statique par une théorie dynamique et fluide du génome dans lequel des réarrangements génétiques, parfois profonds, peuvent avoir lieu. Il a fallu attendre les années 1970, pour que l'intérêt porté à ces éléments se développe. A cette époque, Shapiro (1969) découvre que les éléments *IS* (séquences d'insertion) chez *Escherichia coli* sont impliqués dans l'acquisition d'une résistance à un antibiotique. Par ailleurs, chez *Drosophila melanogaster*, les éléments *P* et *I* ont été décrits comme étant responsables de la dysgénésie hybride (Kidwell *et al.* 1977, Bucheton *et al.* 1984).

Ces séquences appelées aujourd'hui éléments transposables (ETs), sont capables de se répliquer et de se déplacer d'une position chromosomique à une autre dans un génome. Ces éléments bougent au sein d'une cellule, ce qui les différencie des virus. Ceci dit, de nombreux cas de transferts horizontaux entre espèces sont suspectés (Daniels et al. 1990; Loreto *et al.* 2008). La plupart des ETs possède des séquences promotrices et des séquences codantes nécessaires à leur transposition, même s'ils ont besoin de la machinerie de transcription et de traduction du génome hôte (comme les virus). Par ailleurs, les ETs sont de trois types (**Figure 6**) (Feschotte et Mouches 2000; Craig *et al.* 2002) :

- Les autonomes, qui contiennent un ou plusieurs gènes codant pour des protéines fonctionnelles impliquées dans leur mobilité et la régulation de leur expression.
- Les non-autonomes qui possèdent des cadres de lecture ouverts (ORFs) délétés ou mutés codant pour des protéines non fonctionnelles mais qui peuvent être mobilisés en *trans* par la machinerie d'un ETs autonome.
- Les éléments tronqués (mort) incapables de bouger.

Si sur le court terme, ces éléments peuvent avoir des effets délétères sur la structuration et le fonctionnement des génomes, sur le long terme, ils peuvent être source de variabilité génétique et peuvent également être domestiqués par l'hôte pour de nouvelles fonctions (exaptation). Ainsi, les ETs qu'on qualifiait de gènes égoïstes ou de parasites moléculaires sont

considérés comme de véritables moteurs évolutifs d'adaptation et de plasticité génomique (Orgel et Crick 1980; Dawkins 1990; McDonald 1995; Britten 1996; Capy *et al.* 2000; Schmidt et Anderson 2006; Böhne *et al.* 2008).

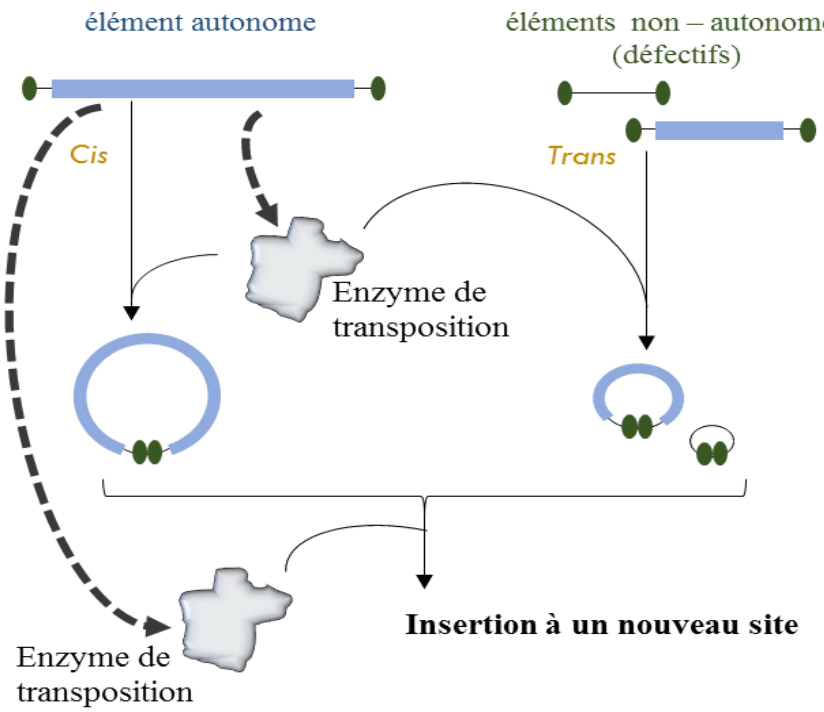


Figure 6. Régulation en *cis* pour les éléments autonomes et en *trans* pour les éléments non-autonomes de Classe II (d'après Anxolabéhère *et al.* 2007).
Les éléments autonomes, codant une ou plusieurs enzymes, possèdent généralement des séquences terminales qui permettent la régulation en *cis* de leur transposition. Les éléments non - autonomes utilisent *en trans* les enzymes codées par les ETs autonomes.

2. Classification

Les éléments transposables ont été divisés, en fonction de leur mécanisme de transposition, en deux Classes : la Classe I ou rétrotransposons et la Classe II ou transposons à ADN (**Figure 7**) (Finnegan 1989,1992; Jurka *et al.* 2005; Wicker *et al.* 2007, Kapitonov et Jurka 2008) :

- Les Rétrotransposons ou Classe I transposent par l'intermédiaire d'un ARN transcrit par l'ARN polymérase II de l'hôte. Celui-ci est rétro-transcrit par la transcriptase réverse codée par le rétrotransposon, ensuite l'ADNc généré est inséré dans un nouveau locus du génome au niveau d'un site d'insertion TSD (terminal site duplication) spécifique par une intégrase. On parle ainsi d'une amplification réplicative selon le modèle copier-coller (**Figure 8**).
- Les transposons ou Classe II se déplacent *via* un intermédiaire ADN et sont subdivisés, selon leur mode de transposition conservatif ou réplicatif, en deux sous-classes (**Figure 8**).

Par ailleurs, en fonction de leur structure, de leur cycle de transposition, de leur domaine protéique ainsi que de leur site d'insertion TSD dans le génome, chaque Classe est subdivisée en différents ordres, superfamilles, familles qui peuvent coexister dans le même génome (Capy *et al.* 1997; Wicker *et al.* 2007).

2.1. Les éléments de la Classe I

Les éléments de la Classe I, ou rétrotransposons, peuvent être divisés en deux groupes : Les rétrotransposons à LTR qui possèdent de longues répétitions terminales (Long terminal repeat) et des rétrotransposons sans LTR.

2.1.1 Les rétrotransposons à LTR

Ils regroupent 5 grandes familles, *Copia*, *Gypsy*, *Bel-Pao, les Rétrovirus* et les rétrovirus endogènes (*ERV*). Leur taille varie de 5 à 10 kb (Bennetzen *et al.* 2005; Wicker *et al.* 2007; Charles *et al.* 2008). Ces ETs sont flanqués par deux répétions terminales identiques, longues et directes (LTR) dont la taille varie entre 100 pb à plusieurs kb.

Ces éléments autonomes contiennent principalement deux à trois ORF (**Figure 7**) :

- un gène *gag* codant pour trois protéines structurales de la capside virale chez les rétrovirus à savoir la protéine de la matrice et la protéine de la nucléocapside (Warmus et Brown 1989) ;
- un gène *pol* codant pour une polymérase nécessaire à la transposition constituée de quatre domaines : une protéase (PR), une transcriptase réverse (RT), une ribonucléase H (RH) et une intégrase (INT).
- un gène *env*, présent uniquement chez les superfamilles des rétrovirus et des *ERV*, codant pour la protéine d'enveloppe.

Il faut signaler que les éléments de l'ordre des *DIRS* (Dictyostelium Intermediate Repeat Sequences) dépourvus de LTR sont classés, de par la proximité phylogénétique de leurs séquences, avec les rétrotransposons à LTR (Poulter et Goodwin 2005). Ils présentent une tyrosine recombinase à la place de l'intégrase (Goodwin et Poulter 2001). Ces éléments sont subdivisés en trois superfamilles en fonction de leur structure terminale (**Figure 7**) : les *DIRS1*-like présentant des extrémités répétées inversées, les *Ngaro* et les *VIPER* présentant des répétitions directes dupliquées en 3' mais qui différent par leur hôtes.

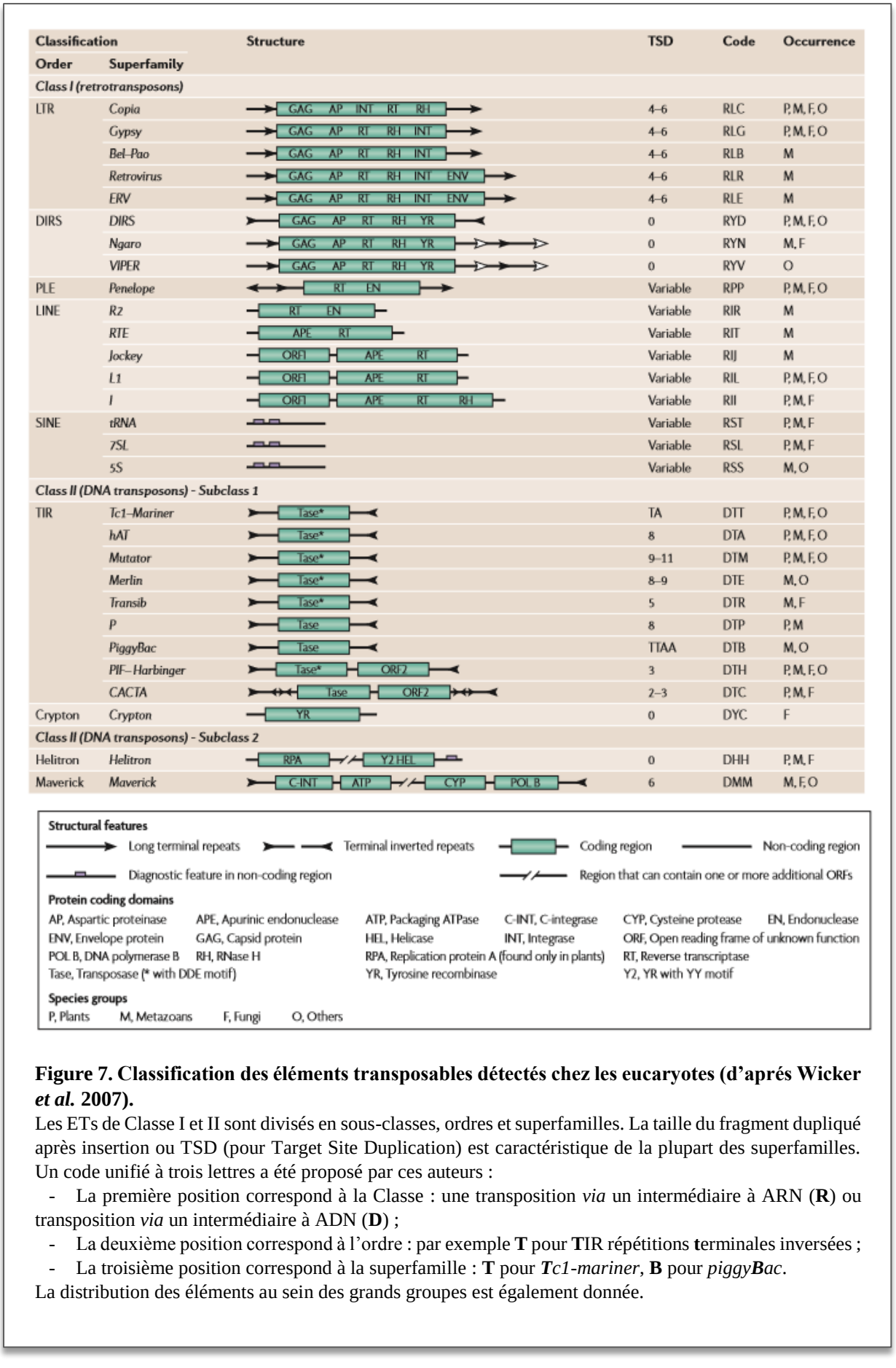


Figure 7. Classification des éléments transposables détectés chez les eucaryotes (d'aprés Wicker *et al.* 2007).

Les ETs de Classe I et II sont divisés en sous-classes, ordres et superfamilles. La taille du fragment dupliqué après insertion ou TSD (pour Target Site Duplication) est caractéristique de la plupart des superfamilles. Un code unifié à trois lettres a été proposé par ces auteurs :

- La première position correspond à la Classe : une transposition *via* un intermédiaire à ARN (**R**) ou transposition *via* un intermédiaire à ADN (**D**) ;
- La deuxième position correspond à l'ordre : par exemple **T** pour **T**IR répétitions **t**erminales inversées ;
- La troisième position correspond à la superfamille : **T** pour ***T****c1-mariner*, **B** pour *piggy****B****ac*.

La distribution des éléments au sein des grands groupes est également donnée.

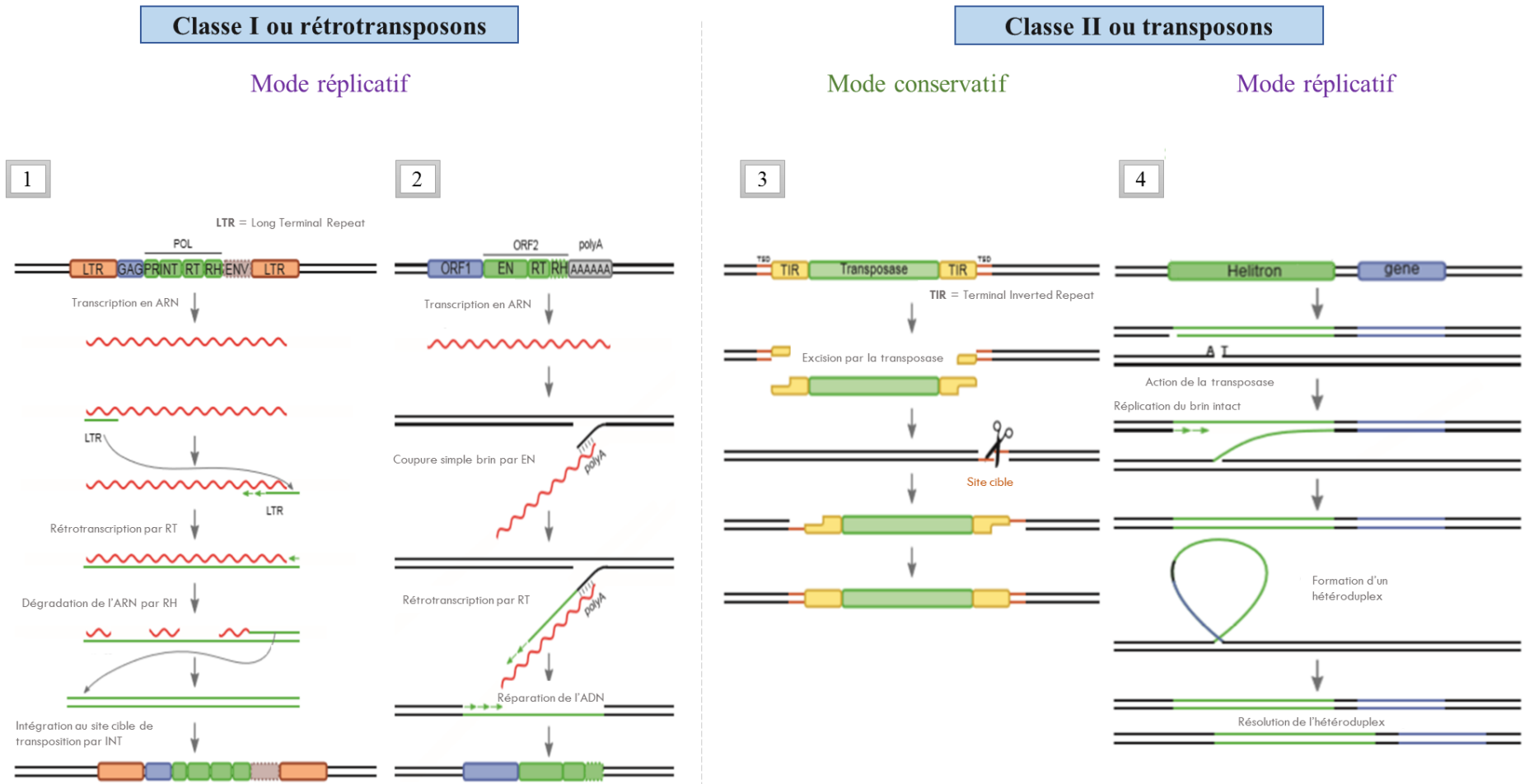


Figure 8. Structure et cycle de transposition (d'aprés Wicker *et al.* 2007; Modolo 2014).

Quatre types de transposition sont résumés dans ce schéma. Les deux premiers opèrent selon un mode réplicatif. Ils correspondent à une transposition de type « copier-coller » et concerne les éléments de la Classe I. Le premier schéma (1) résume la transposition d'un rétrotransposon à LTR (la reverse transcription se fait dans la capside) et le deuxième (2) celle d'un rétrotransposon sans LTR pour lequel la reverse transcription se fait dans le noyau au moment de l'intégration. Les deux schémas suivants (3 et 4) concernent des éléments de la Classe II. Le schéma 3 représente un modèle de transposition de type « couper-coller » conservatif, relatif aux éléments de la sous-classe I. Au cours de ce cycle, le site cible, présenté en marron, subit une coupure à extrémités adhésives qui va conduire à sa duplication et la formation d'un TSD de part et d'autre de la copie de l'élément. Le schéma 4 correspond au mode de transposition des éléments de type *Hélitron* (sous-classe II) qui transposent selon un modèle « couper-coller » réplicatif. Un seul brin du transposon est clivé et transposé vers un nouveau site et la brèche laissée est réparée par des mécanismes de réparation de l'ADN en se basant sur la séquence du transposon original comme matrice. Au final un total de quatre copies est obtenu, soit deux par chromatide.

L'ADN génomique est représenté par une ligne droite noire, l'ETs en vert et l'ARN est par une ligne ondulée rouge. Les différents domaines protéiques sont représentés par des rectangles de couleurs : EN pour endonucléase, ENV pour enveloppe, INT pour intégrase, PR pour protéase, RH pour RNAseH, RT pour transcriptase inverse.

2.1.2 Les rétrotransposons sans LTR

Ces éléments se répartissent en deux grands groupes :

- les éléments autonomes qui incluent les *LINE* (Long Interspersed Nuclear Elements) et les *PLE* (*Penelope*-Like-Elements).

Les *LINE* ont une taille de 5 à 8 kb. Ils codent pour une transcriptase réverse et une endonucléase. Les différentes familles (*R2, RTE, Jockey, L1* et *I*) possèdent une queue poly-A à l'extrémité 3'et sont dépourvus d'intégrase mais possèdent une endonucléase, leur conférant un cycle de transposition différent de celui des rétrotransposons à LTR. Par ailleurs, ces éléments sont fréquemment tronqués en 5' en raison d'intégrations imparfaites. Les éléments de la famille *Penelope* (*PLE*) présentent, quant à eux, des répétitions terminales inversées ou directes et possèdent un ORF unique de 2.5 kb qui code pour une transcriptase réverse et une endonucléase. De plus, ils peuvent comporter des introns (Evgen'ev et Arkhipova 2005).

- les éléments non-autonomes incluant les *SINE* (Short Interspersed Nuclear Elements).

Les *SINE* ont une taille de 80 à 500 bp et ne possèdent ni ORF ni LTR. Néanmoins, ils comportent un promoteur d'ARN polymérase III en 5' assurant leur transcription et une queue poly A qui est reconnue par la machinerie de transposition des *LINE* et notamment la RT ce qui les rend dépendant des *LINEs* (Unsal et Morgan 1995; Okada *et al.* 1997; Kramerov et Vassetzky 2005). L'élément *Alu*, trouvé dans le génome humain à plus de un million de copies, est responsable de plusieurs maladies (Deragon et Capy 2000 ; Roy-Engel *et al.* 2001; Dewannieux *et al.* 2003) est considéré comme étant le plus représentatif de cette superfamille.

2.2. Les éléments de la Classe II

Les éléments de la Classe II ou transposons à ADN sont subdivisés en deux sous-classes selon le mode de transposition conservatif ou réplicatif (**Figure 7**) (Feschotte et Pritham 2007; Wicker *et al.* 2007) :

- Les ETs de la sous-classe I sont coupés directement de l'ADN hôte puis insérés dans un nouveau site TSD, selon un modèle de transposition conservatif de type « couper-coller » assuré par un gène qui code pour la transposase (**Figure 8**).
- Les ETs de la sous-classe II transposent *via* un modèle réplicatif de type « copier-coller » ayant pour conséquence l'excision et le déplacement d'un seul brin d'ADN (**Figure 8**).

Par ailleurs, un groupe d'éléments non-autonomes *MITE* (Miniature Inverted Repeat Transposable Element) sont des versions délétées des transposons de Classe II et ne sont mobilisables qu'en *trans* par les ETs autonomes de cette Classe. Ils ont été décrits chez les plantes, les champignons, les amphibiens, les poissons et l'Homme (Bureau et Wessler 1994; Yeadon et Catcheside 1995; Izsvàk *et al.* 1999; Dufresne *et al.* 2007). Plusieurs mécanismes

de réparation de l'ADN pourraient donner lieu à des délétions associées à des microhomologies au niveau des sites de rupture (Puchta 2005; McVey et Lee 2008). Le mécanisme *abortive gap repair* (AGR), proposé par Rubin et Levy (1997), est basé sur l'avortement de la synthèse de l'ADN, après excision d'un élément, à partir d'un brin complémentaire avant la fin de la réparation. Ceci conduit à l'émergence de diverses copies non autonomes ou à des copies délétées chimèriques rares avec une courte insertion interne (Brunet *et al.* 2002). D'autres mécanismes comme les systèmes non homologous end joining (NHEJ), microhomology – mediated end joining (MMEJ) et single strand annealing (SSA) peuvent également être impliqués. Ils sont capables d'induire des délétions de longueur variable à l'intérieur de divers éléments (Brunet *et al.* 2002; Negoua *et al.* 2013).

2.2.1 La sous-classe I : ordres des *TIR* et des *Crypton*

L'ordre des TIR, le plus fréquent, est caractérisé par des répétitions terminales inversées (TIR) qui régulent en *cis* la transposition et d'un ORF qui code une transposase. Il comprend 9 superfamilles : les *Tc1-mariner* (Plasterk *et al.* 1999), les *hAT* (Calvi *et al.* 1991), les *Mutator* (Lisch 2002), les *P* (Bingham *et al.* 1982), les *Merlin* (Feschotte 2004), les *Transib* (Kapitonov et Jurka 2003), les *piggyBac* (Fraser *et al.* 1983), les *PIF-Harbinger* (Jurka et Kapitonov 2001) et les *CACTA* (Kunze et Weil 2002). Une description plus détaillée des superfamilles *Tc1-mariner* et *piggyBac*, objet de notre recherche, sera présentée plus loin. Chaque superfamille est définie par sa structure, la longueur des TIR et les motifs du site catalytique de la transposase DDD/E responsable de l'excision et de l'intégration de l'élément (Plasterk *et al.* 1999; Feschotte et Pritham 2007; Wicker *et al.* 2007).

L'ordre des *Crypton* est représenté par une seule superfamille dépourvue de TIR et possède un ORF codant une Tyrosine recombinase à la place de la transposase, lui assurant un mécanisme de déplacement impliquant la recombinaison entre une molécule circulaire et l'ADN cible (Goodwin *et al.* 2003).

2.2.2 La sous-classe II : ordre des *Helitron* et des *Maverick*

Ces deux ordres sont représentés chacun par une seule superfamille.

La superfamille *Helitron* de taille de 5.5 à 17 kb est dépourvue de TIR et contient un ORF codant pour une tyrosine recombinase qui lui confère un modèle de réplication selon un mécanisme des cercles roulants (Kapitonov et Jurka 2001; Feschotte et Wessler 2001).

Les éléments *Maverick* (ou *Politrons*) sont, quant à eux, de longues séquences de 10 à 20 kb flanquées de TIR de 150 à 700 pb et codent jusqu'à 11 protéines, dont une polymérase, une

intégrase de type rétrovirale, une protéase de type adénovirale et une ATPase, avec une variation de l'ordre des gènes selon les éléments (Kapitonov et Jurka 2006).

3. Caractéristiques, dynamique et impact des éléments transposables

En plus de leur capacité à se déplacer dans le génome, les ETs possèdent des caractéristiques qui peuvent leur conférer un rôle dans l'évolution des génomes.

3.1. Caractéristiques des ETs

Les éléments transposables sont ubiquitaires. Ils sont capables de se multiplier dans un génome mais également d'en coloniser d'autres *via* deux modes de transferts. Le premier mode correspond à un transfert vertical mendélien, ces éléments peuvent ainsi envahir le génome à partir d'un ancêtre, *via* la lignée germinale. Le deuxième mode, plus difficile à mettre en évidence, correspond au transfert horizontal (TH) d'ETs entre espèces reproductivement isolées et pouvant nécessiter l'intervention d'un organisme vecteur (navette). Généralement, les études illustrant les TH sont basées sur le pourcentage d'identité entre les séquences trouvées chez les espèces donatrices et réceptrices sans pour autant connaitre le vecteur (Pace *et al.* 2008; Loreto *et al.* 2008, Dupeyron *et al.* 2014; Tang *et al.* 2015). Ce mode de transfert a été en premier décrit avec l'élément *P* chez *D. melanogaster*, transféré horizontalement il y a 60 ans, à partir *D. willistoni* (Brookfield *et al.* 1984; Anxolabéhère *et al.* 1988; Daniels *et al.* 1990; Clark *et al.* 1994). Plusieurs arguments en faveur de ce mode de transfert ont été ensuite publiés. On peut citer ceux d'El Baidouri *et al.* (2014) qui ont montré la présence de rétrotransposons à LTR avec une distribution aléatoire chez 40 espèces de plantes phylogénétiquement éloignées appartenant aux Monocotylédones (*i.e. Poaceae*, *Arecaceae*, *Musaceae)* et aux Dicotylédones (*i.e. Fabaceae*, *Vitaceae*, *Solanaceae*…), et ceux de Gilbert *et al.* (2010) qui ont constaté que quatre familles distinctes de transposons chez *Rhodnius prolixus*, vecteur de la maladie de Chagas chez l'homme, ont été retrouvées chez deux hôtes de mammifères et que l'une de ces familles est présente chez des limnées, vecteurs de trématodes infectant divers vertébrés, et chez les mammifères de l'ancien monde.

Malgré leur présence dans les génomes, la répartition des ETs est variable d'une espèce à une autre indépendamment de la complexité (*e.g.* le nombre de fonctions cellulaires) de celui-ci (**Figure 9**) (Hua-Van *et al.* 2005; Feschotte et Pritham 2007). En effet, les ETs peuvent constituer la majeure partie d'un génome tels que ceux des poacées avec 90% chez le blé, 85% chez le maïs et 60% chez le riz (Goff *et al.* 2002; Schnabel *et al.* 2009; Biémont 2010). A l'inverse, chez l'abeille *Apis mellifera* (Biémont 2010) et la levure *Saccharomyces cerevisiae* (Spellman *et al.* 1998) les ETs ne représentent respectivement que 1% et 5% des génomes.

La proportion des ETs peut être quasiment identique pour un même groupe phylogénétique comme les primates (Homme, chimpanzé et macaque) avec 45%. Cependant, chez d'autres espèces proches, comme par exemple *D. melanogaster* et *D. simulans*, la proportion des ETs est variable, avec 15% et 5% respectivement (Clark *et al.* 2007).

Par ailleurs, au sein d'un même génome, la proportion des différentes familles d'ETs peut être très variable et certaines superfamilles ou Classe peuvent être plus abondantes (Feschotte et Pritham 2007; Pritham 2009), sans qu'aucune règle puisse être dégagée. Par exemple, les éléments de Classe I sont prépondérants dans les génomes des végétaux (*e.g* orge, blé) et dans ceux des mammifères (*e.g. Homo sapiens*) alors que les génomes des invertébrés (*e.g. Acyrthosiphon pisum*) présentent une majorité d'ETs de Classe II (**Figure 10**).

3.2. Dynamique des ETs dans les génomes

La dynamique ou cycle des ETs est généralement divisée en trois phases (**Figure 11**). La première correspond soit à l'invasion de la population suite à l'arrivée d'un nouvel élément actif par transfert horizontal dans le génome d'une espèce, soit à la réactivation d'une copie précédemment inactive (Kidwell 1992; Sanchez-Gracia *et al.* 2005). Dans une telle situation, un élément doit être très efficace de façon à se multiplier rapidement dans le génome hôte, sinon, il risque d'être éliminé par simple dérive génétique et/ou sélection (Le Rouzic et Capy 2005).

L'étape suivante est une phase durant laquelle des mécanismes de régulation ainsi que l'apparition de mutations et de délétions limitent l'invasion de l'élément. Ces mécanismes peuvent être de plusieurs types :

- régulation par la répression de la transcription comme observé pour les séquences d'insertions bactériennes *IS* (Zerbib *et al.* 1990).
- autorégulation par la surproduction de la transposase (*overproduction inhibition OPI*) entrainant la formation d'oligomères inactifs ou moins actifs qui diminuent l'efficacité du processus de transposition, ou par l'inhibition compétitive entre les éléments actifs et inactifs pour l'accès aux répétitions terminales inversées (TIRs) ce qui bloquerait l'excision des copies actives. Les copies délétées agissent comme des inhibiteurs négatifs dominants de la transposition (*dominant negative complementation DNC*). Ces deux modèles ont été décrits chez *D. melanogaster* avec l'élément *mos1* de la famille *mariner* (Lohe et Hartl 1996; Lohe *et al.* 1997).
- régulation épigénétique par la méthylation d'ADN (Martienssen et Baron 1994; Slotkin et Martienssen 2007) ou par les ARN anti-sens formant des dimères avec les ARNm des ETs (Sijen et Plasterk 2003; Petit *et al.* 2007; Dowling *et al.* 2017).

Enfin, la phase de sénescence durant laquelle une dérive aléatoire peut entrainer une forte diminution du nombre de copies jusqu'à l'élimination complète de la famille. Ceci peut être dû soit à l'inactivation verticale par accumulation de mutations conduisant à la perte totale d'activité des ETs ainsi que l'impossibilité d'être mobilisé en *trans* (Lohe *et al.* 1995; Lohe *et al.* 1997; Le Rouzic et Capy 2005), soit à des recombinaisons ectopiques (Biémont *et al.* 1997; Kalendar *et al.* 2011) ou encore à la perte stochastique de copies par dérive génétique (Capy *et al.* 1992a; Lohe *et al.* 1995).

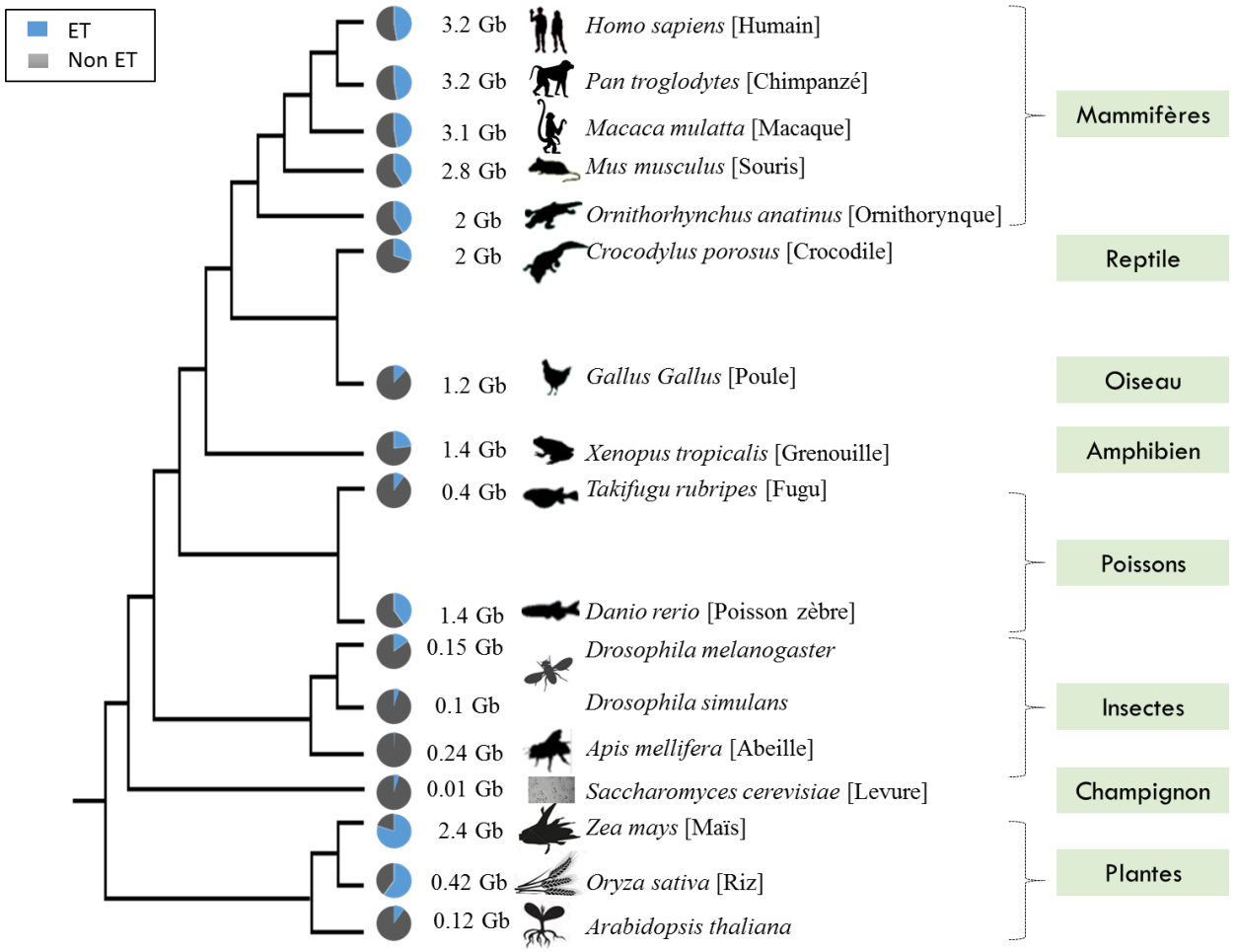


Figure 9. Représentation de la variation de la quantité des ETs dans les génomes Eucaryotes.
Cette figure résume la composition en ETs et la taille des génomes de différentes espèces. La distribution des ETs dans ces génomes semble stochastique. On note chez certaines espèces, une abondance d'ETs associée à une augmentation de taille du génome comme chez le maïs et le riz. Toutefois, on remarque aussi la faible proportion d'ETs (comme chez l'abeille), souvent associée à des phénomènes de contraction du génome et à la perte de grande quantité d'ADN.
Données obtenues à partir du séquençage de chacun de ces génomes : *Apis mellifera* [Honeybee Genome Sequencing Consortium, 2006], *Arabidopsis thaliana* [Kaul *et al.* 2000], *Danio rerio* [Haffter *et al.* 1996], *Crocodylus porosus* [St John *et al.* 2012], *Drosophila simulans* et *D. melanogaster* [Clark *et al.* 2007], *Gallus gallus* [Brandström et Ellegren, 2007], *Homo sapiens* [Lander *et al.* 2001], *Macaca mulatta* [Gibbs *et al.* 2007], *Mus musculus* [Chinwalla *et al.* 2002], *Ornithorhynchus anatinus* [Warren *et al.* 2008], *Oryza sativa* [Goff *et al.* 2002], *Pan troglodytes* [Chimpanzee Sequencing and Analysis Consortium, 2005], *Saccharomyces cerevisae* [Spellman *et al.* 1998], *Takifugu rubripes* [Kurowaka *et al.* 2005], *Xenopus tropicalis* [Hellsten *et al.* 2010], *Zea mays* [Schnable *et al.* 2009].

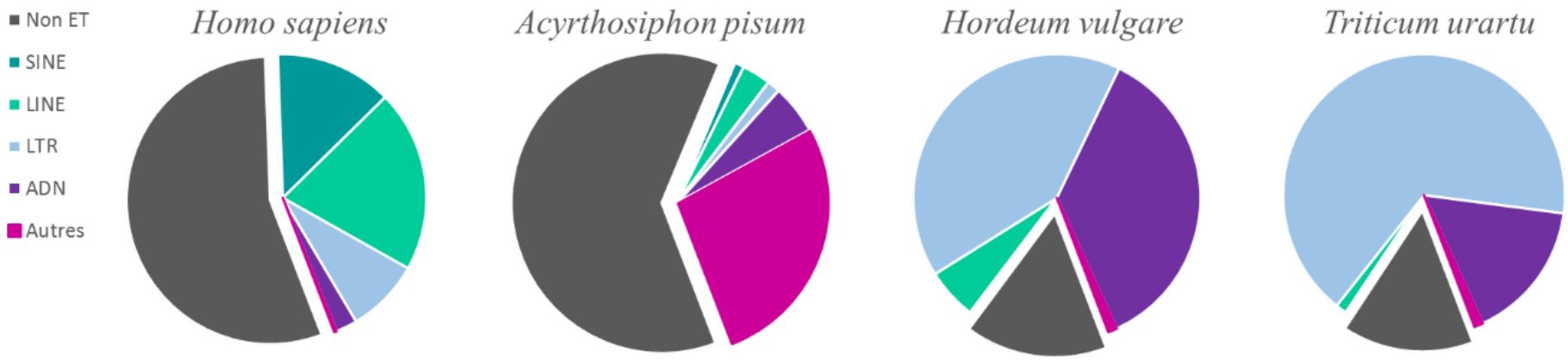


Figure 10. Proportion des différents types d'éléments transposables dans les génomes de l'Homme, du puceron vert du pois et de triticales.
La figure montre les fractions de génomes occupées par les grandes familles d'ETs. Les variations de proportion des différentes familles d'ETs sont souvent liées à une amplification massive. Chez l'homme, les ETs sont dominés par des rétrotransposons sans LTR (*LINE* et *SINE*) suivi de rétrotransposons à LTR. La tendance est inversée chez les triticales (orge et blé) avec une dominance de rétrotransposons à LTR (*e.g. Gypsy*, *Copia*). Toutefois, dans le génome de l'orge et du puceron vert du pois, les éléments de Classe II sont bien plus abondants que chez *H. sapiens*.
Le groupe noté « autres » inclut des régions répétitives qui correspondent à des consensus d'ETs spécifiques à l'hôte mais qui n'ont pas pu être classées par le pipeline REPETs vu l'absence de caractéristiques structurelles et de similarités avec les ETs connus.
Données obtenues à partir du séquençage de chacun de ces génomes : *Acyrthosiphon pisum* [International Aphid Genomics Consortium 2010], *Homo sapiens* [Lander *et al.* 2001], *Hordeum vulgare* [Mazaheri *et al.* 2014], *Triticum urartu* [Daron *et al.* 2014].

Dans un même génome l'ensemble des ETs ne sont pas tous dans la même phase de leur cycle. Aussi, il peut y avoir de nombreuses différences d'une famille à une autre, certains ETs sont dans une phase d'invasion alors que d'autres sont dans leur phase de sénéscence (Le Rouzic *et al.* 2007).

Bien que la plupart des copies vont être perdues au cours du temps, certaines d'entre elles peuvent s'insérer dans des régions où elles induisent une augmentation de la fitness des individus qui les portent (insertions adaptatives). D'autres peuvent être domestiquées ou « exaptées » et survivent dans le génome hôte où elles sont « récupérées » pour une nouvelle fonction. Généralement, ces séquences sont en copie unique, elles sont conservées et persistent dans la population grâce aux effets de la sélection (Le Rouzic *et al.* 2007; Jurka *et al.* 2012). Par ailleurs, elles sont responsables d'un balayage sélectif dans la région de leur insertion. Par exemple, des travaux antérieurs ont montré que la résistance à l'insecticide DDT chez *D. melanogaster* et *D. simulans* est corrélée à l'insertion du rétrotransposon *Accord* dans la région 5 ' du gène *Cyp6g1* du cytochrome P450. Une réduction significative de la variabilité s'étendant au moins 20 kb en aval du gène de résistance a été observée (Catania *et al.* 2004; Schlenke et Begun 2004).

3.3. Impacts des éléments transposables

Les ETs jouent un rôle important dans l'évolution du génome de l'hôte et dans la biodiversité à travers trois mécanismes majeurs :

- les recombinaisons ectopiques entre régions homologues d'ETs : ces recombinaisons entre des éléments situés sur un même chromosome peuvent engendrer des duplications, des délétions ou des inversions de séquences dans le génome de l'hôte. Ceci peut également conduire à des différences significatives de taille des génomes entre espèces d'un même genre (Capy *et al.* 1992b; SanMiguel et Bennetzen 1998; Vicient *et al.* 1999; Devos *et al.* 2002; Tenaillon *et al.* 2010). Ces recombinaisons peuvent également avoir lieu entre des éléments situés sur des chromosomes différents et ont pour conséquence des translocations chromosomiques ou des réarrangements plus complexes (Gray 2000; Han *et al.* 2005). C'est le cas des éléments *Alu* trouvés chez les primates et qui seraient responsables de la translocation entre les chromosomes 11 et 12 chez l'Homme (Hill *et al.* 2000). Il peut y avoir également des recombinaisons intra-élément (essentiellement des rétrotransposons à LTR) à l'origine de la suppression des séquences situées entre les deux LTR et conduisant à la formation de solo-LTR (SanMiguel *et al.* 1996; Tenaillon *et al.* 2010).

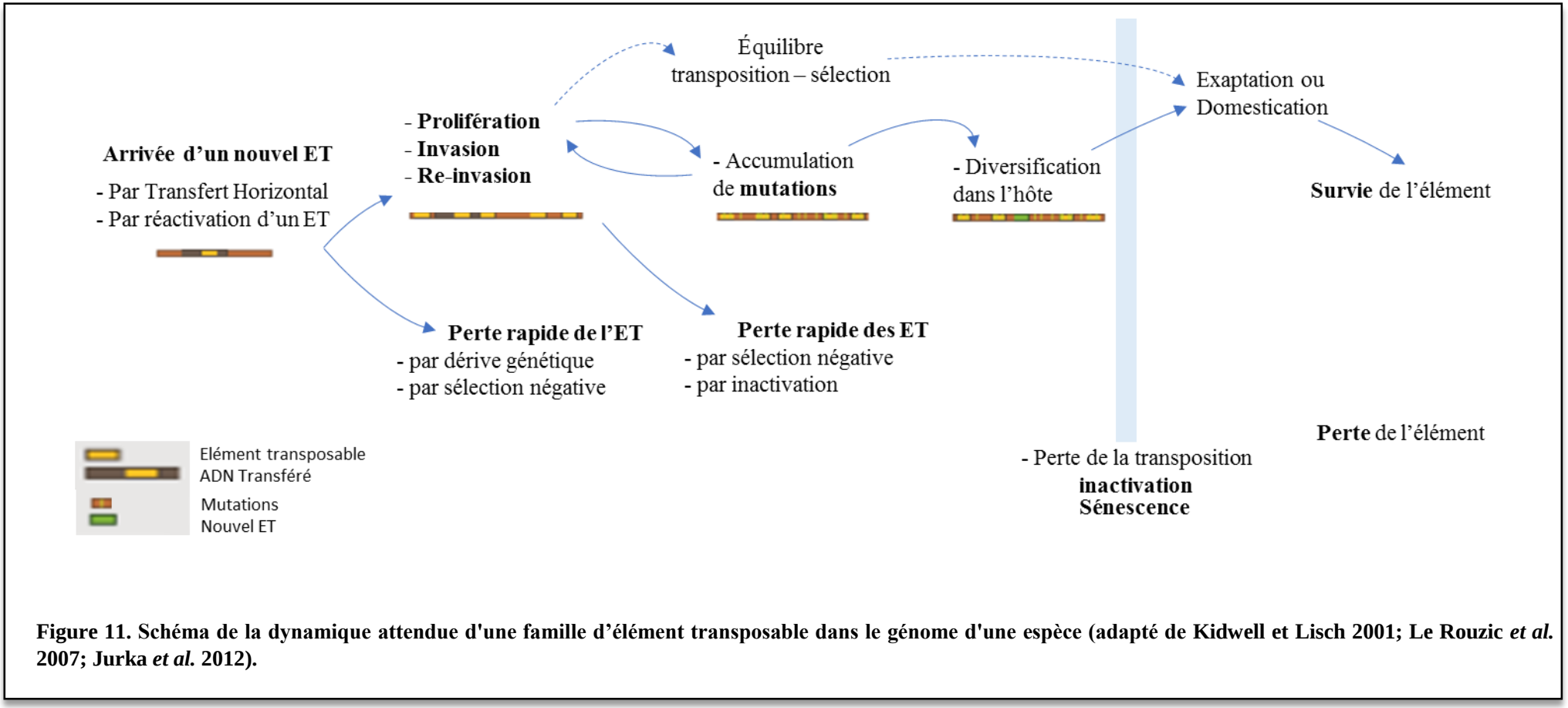


Figure 11. Schéma de la dynamique attendue d'une famille d'élément transposable dans le génome d'une espèce (adapté de Kidwell et Lisch 2001; Le Rouzic *et al.* 2007; Jurka *et al.* 2012).

- La transduction : les ETs peuvent promouvoir le mécanisme de transduction qui consiste à déplacer certaines séquences géniques vers un autre endroit du génome. Ce mécanisme a été mis en évidence sur des lignées cellulaires humaines pour des éléments *L1* qui sont des *LINE* (Moran *et al.* 1999). En effet, la région transduite, pouvant contenir des promoteurs, enhancers ou des exons, situeé en 3' des *L1* permet de créer de nouveaux gènes ou de modifier la régulation de gènes préexistants (Burki et Kaessman 2004). Dans d'autres cas, les ETs sont responsables de changements phénotypiques pouvant être néfastes tels que les insertions indépendantes de rétrotransposons tronqués L1 dans le gène codant le facteur VIII de la coagulation sanguine induisant l'hémophilie de type A (Dombroski *et al.* 1991).Il faut signaler que leur fixation dans le proto-oncogène *c-myc* ou dans le gène suppresseur de tumeur *APC* provoquant respectivement le cancer du sein et du colon (Rodić et Burns 2013). De plus, ces éléments peuvent participer indirectement à la régulation de l'expression des gènes *via* les *micro*ARN (*mi*RNA) qui dérivent eux-mêmes parfois de séquences d'ETs (Piriyaponga *et al.* 2007).

- La domestication : durant l'évolution, les ETs peuvent être « détournés », « réutilisés » et devenir des éléments constitutifs du génome. Ce processus a été baptisé domestication ou « exaptation ». Les éléments domestiqués se trouvent en une seule copie, dans des loci orthologues chez différentes espèces et sont généralement associés à une fonction (Britten 1996; Kidwell et Lisch 1997; Miller *et al.* 2000; Kidwell 2002; Volff 2006; Sinzelle *et al.* 2009; Joly-Lopez *et al.* 2016). Nous décrirons dans les résultats de cette thèse, des événements de domestication survenus dans la superfamille des *piggyBac*. Le cas le plus typique chez les mammifères est celui de la recombinaison V(D)J des immunoglobulines, due aux gènes RAG1 et RAG2, dérivant de l'élément *transib,* qui initient cette recombinaison (Biémont et Vieira 2006). La domestication peut être la conséquence d'un phénomène de fusion entre un transposon délété avec une séquence d'ADN qui code pour un domaine ou une protéine fonctionnelle. C'est l'exemple de la protéine domestiquée *SETsMAR* chez les primates, qui provient de l'insertion du transposon *Hsmar1* appartenant à la famille des *mariner* dans un domaine SETs (activité histone méthyl transférase) formant ainsi une protéine chimérique qui persiste depuis 50 millions d'années (Cordaux *et al.* 2006).

3.4. Utilisation des éléments transposables en biotechnologie

Les éléments transposables, considérés comme source de variabilité génétique ayant un impact sur le pouvoir adaptatif des espèces et sur leur évolution, présentent plusieurs propriétés qui font d'eux de bons candidats pour générer des outils de mutagenèse insertionnelle ou de transgenèse.

Plusieurs éléments tels que *P*, *minos*, *piggyBac*, *Sleeping beauty* ont été utilisés en mutagenèse insertionnelle dans différentes espèces d'insectes (*D. melanogaster*, *C. capitata*, *Tribolium castaneum*), des cellules humaines et de souris, ainsi que chez l'ascidie (*Ciona intestinalis*). Ces vecteurs permettent d'analyser l'expression de certains gènes et leurs fonctions (Spradling 1995; Klinakis *et al.* 2000; Zagoraiou *et al.* 2001; Bonin et Mann 2004; Awazu *et al.* 2007; Ivics et al. 2009; Rad *et al.* 2010; Yusa 2015).

Concernant la transgenèse, l'élément *P* a été le premier élément utilisé comme outil pour les études génétiques et génomiques chez la drosophile (Ryder et Russel 2003). D'autres éléments sont capables de transposer dans des cellules différentes de leur cellule d'origine. En effet, certains éléments de la superfamille des *Tc1-mariner* ont été exploités. C'est l'exemple de l'élément *Sleeping Beauty* (Ivics *et al.* 1997) qui a été utilisé comme vecteur non viral en thérapie génique humaine sur les cellules somatiques, (Essner *et al.* 2005; Miskey *et al.* 2005; Ivics et Izsvak 2006), pour la génération de souris (Clark *et al.* 2004; Ivics *et al.* 2004; Dupuy *et al.* 2005; Miskey *et al.* 2005), et de poissons transgéniques (Davidson *et al.* 2003; Grabher *et al.* 2003). D'autres vecteurs appartenant à cette superfamille ont été aussi utilisés comme vecteurs en transgénèse tels que *Himar1* dans les cultures cellulaire de l'human *(*Keravala *et al.* 2006), *mos1* dans les cultures cellulaires de *Bombyx mori* (Wang *et al.* 2000) et *Frog Prince* dans les cultures cellulaires d'amphibiens, de poissons et de mammifères (Miskey *et al.* 2003).

Il est important de préciser que l'efficacité de ces ETs est différente selon l'espèce étudiée (Izsvak *et al.* 2000; Wu *et al.* 2006). A ce titre, il est important de noter que l'élément transposable, le plus largement utilisé pour la transformation des lignées germinales d'un large éventail d'espèces d'insectes, y compris les coléoptères, diptères, hyménoptères, lépidoptères et orthoptères est l'élément *piggyBac*. Cet élément a également servi à la transformation d'autres organismes Eucaryotes incluant les plantes, les levures, les protozoaires et les vertébrés (Yusa 2015).

L'objectif de la transgenèse est l'insertion d'un gène dans la lignée germinale de l'hôte, de façon à ce qu'elle soit transmise à sa descendance. Généralement, elle consiste en l'insertion d'un couple de vecteurs dérivés du transposon (Handler 2002), à savoir :

- un vecteur recombinant incluant un gène d'intérêt associé à un marqueur de sélection tel qu'un gène codant pour une protéine fluorescente (GFP ou RFP) sous le contrôle d'un promoteur spécifique flanqué par les séquences terminales répétées ;
- un vecteur assistant non autonome codant une transposase active.

Cette stratégie est maintenant largement utilisée pour le développement de souches génétiquement modifiées appropriées pour la lutte contre les ravageurs.

IV. Les superfamilles *Tc1-mariner* et *piggyBac*

1. La superfamille *Tc1-mariner*

Cette superfamille comprend des éléments apparentés à l'élément *Tc1* identifié chez le nématode *Caenorhabditis elegans* (Emmons *et al.* 1983) nommés *TLE* (*Tc1*-like elements) et les éléments apparentés à l'élément *mariner* « *Dromar* » ou « *mos1* » identifié chez *Drosophila mauritiana* (Jacobson *et al.* 1986) nommés *MLE* (*mariner*-like elements). Etant donné l'existence de fortes similitudes structurales avec les transposons bactériens IS630 (*Insertion Sequence 630*) (Matsutani *et al.* 1987;), tous ces éléments ont été rassemblés pour constituer la superfamille *ITm* pour *IS630-Tc1-mariner* (Henikoff 1992; Shao et Tu 2001).

Largement distribuée au sein des espèces, cette superfamille se caractérise par deux répétitions terminales inversées (TIR) flanquées par la duplication du site cible «TA». Les gènes de transposase présentent des similitudes de séquences, principalement à proximité des acides aminés impliqués dans l'activité de l'enzyme formant une triade catalytique. Cette triade peut être formée de deux acides aspartiques (D) et d'un résidu de glutamate (E) typique des *TLE* ou de trois acides aspartiques (DDD) pour les *MLE* (Doak *et al.* 1994; Plasterk *et al.* 1999). Plus tard, sur la base du nombre de résidus (x) séparant les 2ème et 3ème composants de la triade catalytique du motif DDxD/E, les *Tc1-mariner* ont été subdivisés en plusieurs familles, soit : DD34D (*MLE* animaux), DD34E (*TLE* ou *Gambol*), DD37D (*maT*), DD37E, DD39D (*MLE* plantes), d'autres éléments DDxE n'ont pas été bien définis (Feschotte et Mouches 2000; Gomulski et al 2001; Shao et Tu 2001; Feschotte *et al.* 2002). La classification de cette superfamille est en perpétuelle réactualisation (Bui *et al.* 2008; Rouault *et al.* 2009).

Dans le paragraphe suivant, seules les caractéristiques de la famille des *mariner* (incluant les *MLE* de plantes), objet de cette thèse, sont développées.

1.1. Caractéristiques générales des *MLE*

Le premier élément *MLE* a été identifié chez *D. mauritiana* en étudiant une mutation instable du gène *white* suite à l'insertion d'un transposon, appelé *peach*, dans le promoteur. Cet élément est incapable de se déplacer d'une manière autonome mais il est mobilisé en *trans* par l'élément actif similaire *mos1* (pour mosaïque), qui ne diffère de la copie *peach* que par 11 nucléotides (Medhora *et al.* 1988; Maruyama et Hartl 1991; Medhora *et al.* 1991). Depuis, les *MLE* d'origine animale ont été répartis en plusieurs sous-familles présentant un pourcentage de similitude qui varie de 40 à 56 %, et désignées en fonction des espèces dans lesquelles les

premiers membres ont été découverts (Robertson et MacLeod, 1993). Les cinq sous-familles majeures sont : *mauritiana* pour *D. mauritiana*, *cecropia* pour *Hyalophora cecropia*, *elegans/briggsae* pour *Caenorhabditis elegans* et *C. briggsae*, *irritans* pour *Haematobia irritans*, *mellifera/capitata* pour *Apis mellifera* et *Ceratitis capitata* (Jacobson *et al.* 1986; Lidholm *et al.* 1991; Robertson 1993; Robertson et MacLeod 1993; Witherspoon et Robertson 2003).

Au cours du temps, la plupart des éléments ont subi des mutations au niveau de la transposase entrainant ainsi la perte de la capacité de déplacement. A ce jour, les seuls éléments naturels actifs et autonomes identifiés sont *mos1* (*mauritiana*), *Famar1* (*mellifera*) isolé chez le perce-oreille *Forficula auricularia* (Barry *et al.* 2004) et Mboumar-9 (*mauritiana*) découvert chez la fourmi *Messor bouvieri* (Palomeque *et al.* 2006; Munoz-Lopez *et al.* 2008). Un autre *MLE* a été artificiellement reconstruit à partir du consensus d'un ensemble de séquences non autonomes. Ceci a permis de générer une copie active vraisemblablement proche de la copie ancestrale dont sont dérivées les copies non autonomes actuelles. Il s'agit de l'élement *Himar1* (*irritans*) (Miskey *et al.* 2007; Rholl *et al.* 2008). Ces éléments actifs naturels et artificiels sont actuellement exploités comme des outils de transfert de gènes d'intérêt (Wang *et al.* 2000; Sassetti *et al.* 2003; Keravala *et al.* 2006; Jegot 2007).

Les études concernant les *MLE* de plantes sont peu nombreuses. Le niveau de similitude avec des éléments de la famille *mariner* est généralement inférieur à 50% (Bigot *et al.* 2005). Par ailleurs, ces éléments possèdent une triade catalytique de type DD39D. Certains éléments sont potentiellement actifs tels que *Soymar1* (Soybean mariner element 1) isolé chez le soja, *Osmar1* (Oryza sativa mariner element 1), *Osmar2* et *Osmar2b* chez le riz ou *Psmar1* (Pisum sativum mariner element 1) chez le pois (Jarvik et Lark 1998; Tarchini *et al.* 2000; Feschotte *et al.* 2002; Macas *et al.* 2005). Cependant, aucune transposition de ces *MLE* de plantes n'a été mise en évidence (Feschotte et al. 2002; Macas *et al.* 2005). En outre, les éléments du règne végétal ont été désignés par Feschotte *et al.* (2002) « *mariner*» ou *MLE* de plantes. Toutefois, l'appartenance de ces éléments à la famille *mariner* fait débat à cause des motifs de la triade catalytique et de leur taille (Shao et Tu 2001; Brillet *et al.* 2007).

1.2. Structure des *MLE*

Ces éléments, caractérisés chez un grand nombre d'espèces animales (insectes, nématodes, crustacés, mammifères, poissons), présentent une taille d'environ de 1300 pb et une triade catalytique DD34D (Robertson 1997; Plasterk *et al.* 1999; Robertson 2002; Casse *et al.* 2006).

Chez les végétaux, leur taille est généralement plus importante et peut atteindre 7000 pb et leur triade catalytique est constituée par le motif DD39D (Feschotte *et al.* 2002; Leroy *et al.* 2003).

Ces éléments ont des séquences terminales répétées inversées (TIR) bordant deux séquences non traduites UTR (Untranslated Terminal Repeat) et un gène ayant une seule phase de lecture ouverte (ORF) codant la transposase, enzyme capable d'assurer à elle seule toutes les étapes de la transposition. Il faut signaler que les *MLE* de plantes peuvent contenir des introns (Robertson 2002; Feschotte *et al.* 2002; Feschotte *et al.* 2003; Zhou *et al.* 2011). Enfin, ces *MLE* s'insèrent dans un site cible réduit à deux nucléotides TA (**Figure 12**).

Les TIR ont une taille variable de 20 à 40 pb (Jarvick et Lark 1998; Feschotte *et al.* 2003; Jacobs *et al.* 2004; Macas *et al.* 2005; Brillet *et al.* 2007) à l'exception de *Mcmar1* isolé chez le nématode *Meloidogyne chitwoodi* où ils peuvent atteindre 355 pb (Leroy *et al.* 2003). Ces TIR possèdent un motif cardinal conservé en début de séquence 5'YYAGRT 3' (Y pour C ou T, R pour A ou G) qui constitue le signal de clivage de l'ADN, suivi de motifs palindromiques qui sont des sites de liaison à la transposase, et de motifs miroirs qui seraient impliqués dans la stabilisation du complexe Transposase-TIR (Bigot *et al.* 2005).

Les UTR, régions non traduites, comportent généralement des séquences régulatrices de la transcription et de la traduction de la transposase soit la TATABox et la GC Box en 5'UTR et le signal de polyadénylation en 3'UTR (Palomeque *et al.* 2006; Leroy *et al.* 2003).

L'ORF code une transposase de taille d'environ 345 acides aminés (aa) chez les animaux (entre 412 et 520 aa chez les plantes). Cette enzyme est composée de deux domaines : le domaine de liaison à l'ADN et le domaine catalytique (**Figure 12**) :

- Le domaine de liaison à l'ADN, situé dans la partie N-terminale, contient deux motifs hélice-tour-hélice (HTH) impliqués dans la dimérisation de la transposase et sa liaison aux TIR (Augé-Gouillou *et al.* 2005a ; Brillet *et al.* 2007). Le premier motif, localisé en position 25 à 54, est séparé du deuxième motif en position 88 à 108, par une séquence conservée GRPR (Plasterk *et al.* 1999; Bigot *et al.* 2005; Richardson *et al.* 2006). De plus, chez les MLE des animaux, deux autres motifs ont été mis en évidence dans cette région, à savoir, un signal de localisation nucléaire NLS permettant le transport de la transposase dans le noyau, ainsi qu'un motif signature WVPHEL qui interviendrait dans les interactions protéines/protéines (Augé-Gouillou *et al.* 2005b).

- Le domaine catalytique est composé de trois motifs contenant la triade catalytique formée par les trois résidus Aspartate. Ces motifs hautement conservés sont TG**D**E, H**D**NA, YSPDLAPX**D**, pour la famille DD34D et HI**D**EKWF, IQQ**D**NARTH, NSPDL/MN-L**D**LGFF, pour la famille DD39D (Robertson 1993; Feschotte *et al.* 2003).

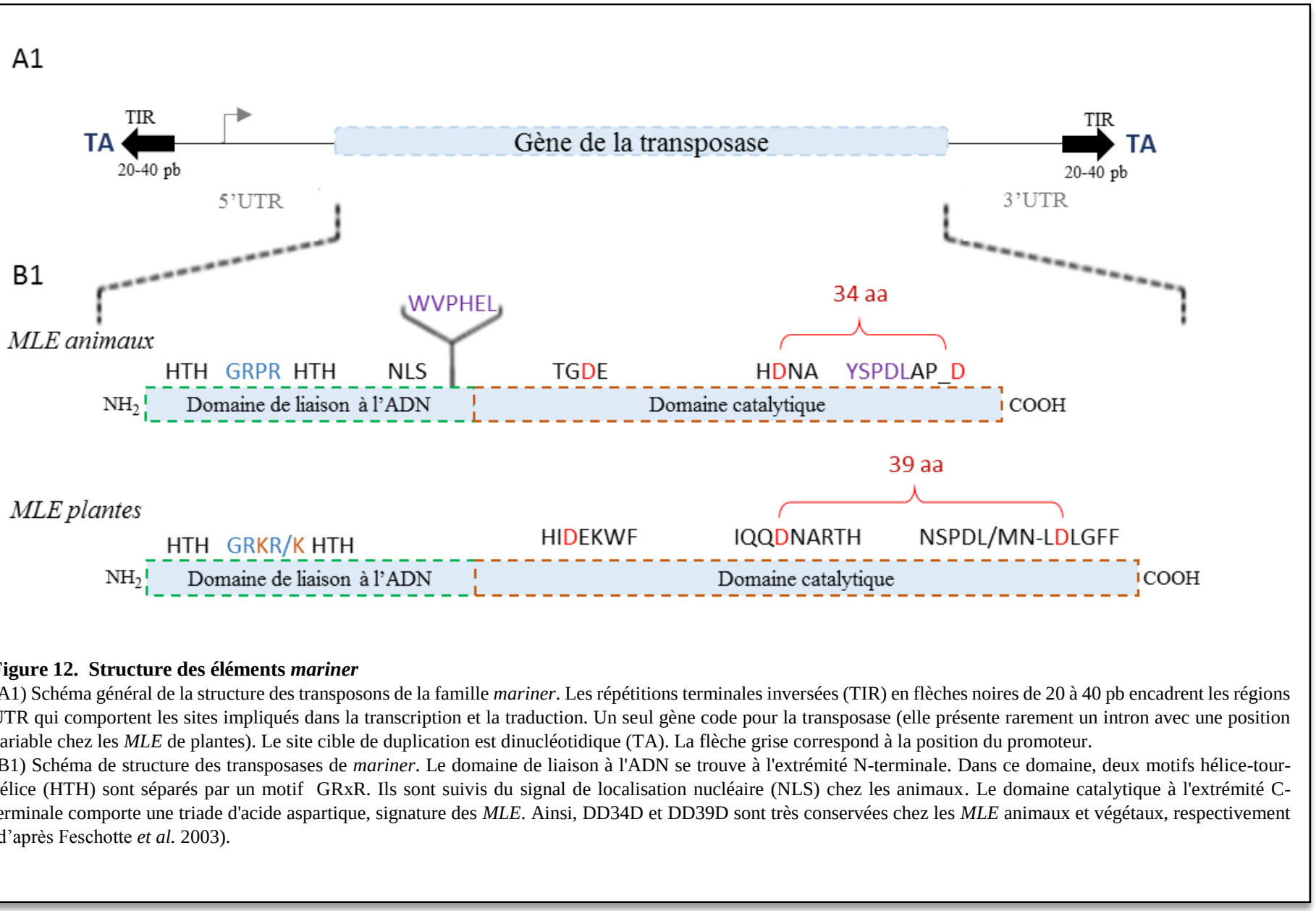


Figure 12. Structure des éléments *mariner*

(A1) Schéma général de la structure des transposons de la famille *mariner*. Les répétitions terminales inversées (TIR) en flèches noires de 20 à 40 pb encadrent les régions UTR qui comportent les sites impliqués dans la transcription et la traduction. Un seul gène code pour la transposase (elle présente rarement un intron avec une position variable chez les *MLE* de plantes). Le site cible de duplication est dinucléotidique (TA). La flèche grise correspond à la position du promoteur.

(B1) Schéma de structure des transposases de *mariner*. Le domaine de liaison à l'ADN se trouve à l'extrémité N-terminale. Dans ce domaine, deux motifs hélice-tour-hélice (HTH) sont séparés par un motif GRxR. Ils sont suivis du signal de localisation nucléaire (NLS) chez les animaux. Le domaine catalytique à l'extrémité C-terminale comporte une triade d'acide aspartique, signature des *MLE*. Ainsi, DD34D et DD39D sont très conservées chez les *MLE* animaux et végétaux, respectivement (d'après Feschotte *et al.* 2003).

1.3. Mécanisme de transposition des *MLE*

Plusieurs travaux portant sur l'étude du mécanisme de transposition de *Himar1* (*irritans*) et *mos1* (*mauritiana*) ont révélé quatre étapes majeures (Lampe *et al.* 1996; Augé-Gouillou *et al.* 2005a; Richardson *et al.* 2006) résumée dans la **Figure 13**.

Tout d'abord, la transposase reconnait les TIR et s'y fixe *via* le ou les motifs HTH situé(s) dans le domaine N-terminal. Deux modèles ont été proposés pour expliquer la fixation de *mos1*. Le premier modèle (Augé-Gouillou *et al.* 2005b) implique un seul motif HTH (résidus 88-108) et un dimère de transposase se fixe sur chaque TIR formant un complexe *cis*-SEC2 (single-end complex). Le second modèle (Richardson *et al.* 2006), quant à lui, implique les deux motifs HTH ; le premier HTH (résidus 25-54) se fixe au niveau des nucléotides 20-25 du TIR et le second HTH (résidus 88-108) au niveau des nucléotides 3-8 du TIR.

Dans les deux modèles, après rapprochement des transposases fixées sur les deux TIR, un complexe nucléoprotéique dit complexe synaptique est formé. Il s'en suit le clivage ou l'excision de chaque brin du transposon à deux ou trois bases à l'intérieur de l'extrémité 5' du TIR (Augé-Gouillou *et al.* 2005a). Un premier clivage génère une extrémité 5'-phosphate au niveau du transposon et une extrémité 3'OH au niveau de l'ADN flanquant (Dawson et Finnegan 2003; Augé-Gouillou *et al.* 2005b), un changement de configuration conduit au positionnement du second brin dans le site actif de la transposase qui effectue un second clivage générant cette fois une extrémité 3'OH au niveau du transposon et une extrémité 5'phosphate au niveau de l'ADN flanquant (Richardson *et al.* 2006). Enfin, l'insertion du transposon s'effectue dans un site cible contenant un dinucléotide TA, suivie de la réparation des brins de l'ADN excisés.

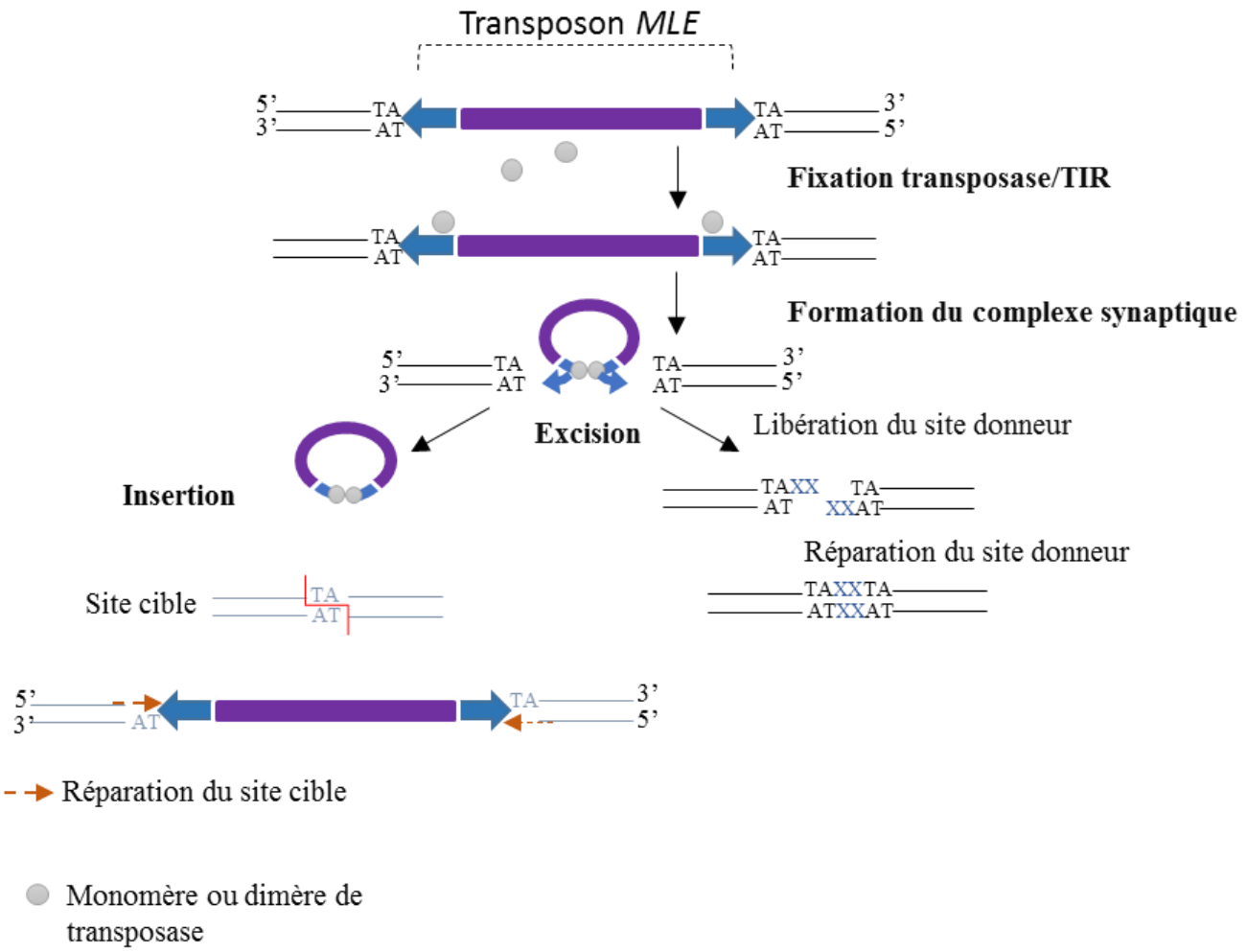


Figure 13. Mécanisme de transposition « couper-coller » des *mariner* (adapté de Tellier *et al.* 2015)
La séquence du site donneur est en noir, celle du site receveur en bleu, la coupure de ce dernier, catalysée par la transposase, est symbolisée en rouge. La formation du complexe synaptique à partir de la transposase fixée sur les TIR, entraine une excision spécifique conduisant d'une part à la libération et « cicatrisation » du site donneur et d'autre part à l'éventuelle réinsertion du transposon, dans un nouveau site cible receveur. La coupure cohésive de la séquence d'insertion conduit à la duplication du site cible.

2. La superfamille *piggyBac*

L'élément *piggyBac* (anciennement nommé IFP2 pour *Insertion Few Polyhedral occlusion bodies*) a été isolé, pour la première fois, à partir d'un baculovirus mutant dans une culture de cellules du lépidoptère *Trichoplusia ni,* fausse-arpenteuse du chou (Fraser *et al.* 1983, 1985). Depuis la découverte de cet élément autonome actif (Cary *et al.* 1989), la distribution taxonomique, restreinte au départ aux différents ordres des insectes, a été significativement élargie pour couvrir plusieurs groupes d'Eucaryotes à l'exception des plantes (Penton *et al.* 2002; Wang *et al.* 2006; Xu *et al.* 2006; Hikosaka *et al.* 2007; Sun *et al.* 2008; Wang *et al.* 2008; Wu *et al.* 2008; Carpes *et al.* 2009; Wang *et al.* 2009; Daimon *et al.* 2010; Luo *et al.* 2011; Wu *et al.* 2011; Luo *et al.* 2014; Wu et Wang 2014).

Le séquençage complet des génomes a également permis de révéler l'existence de gènes homologues à celui de la transposase *piggyBac.* Par exemple, le génome humain contient cinq éléments *piggyBac* domestiqués appelés *piggyBac derived protein/gene/element* et désignés PGBD1 à PGBD5 (Sarkar *et al.* 2003). PGBD1 et PGBD2 étaient probablement présents chez l'ancêtre commun des mammifères, tandis que PGBD3 et PGBD4 l'étaient chez l'ancêtre des primates, alors que PGBD5 est largement distribué chez les vertébrés, y compris la lamproie et le lancelet, suggérant une domestication ancienne concomitante ou antérieure à l'apparition des vertébrés il y a environ 525 Mya (Sarkar *et al.* 2003; Pavelitz *et al.* 2013). De tels évènements de domestication ont été également identifiés chez les ciliés et le xénope (amphibien). En effet, des éléments domestiqués désignés *piggyMac* (PGM) et *TPB2* ont été identifiés chez *Paramecium tetraurelia* (Baudry *et al.* 2009) et *Tetrahymena thermophila* (Cheng *et al.* 2010). Ces gènes interviennent au cours du développement du macronoyau, en induisant l'élimination de séquences internes essentielles pour la reconstruction de gènes fonctionnels (Baudry *et al.* 2009; Cheng *et al.* 2010). En outre, chez *Xenopus*, la transposase domestiquée nommée *KOBUTA,* conservée depuis 100 Mya, semble être impliquée dans une activité de liaison ou de recombinaison à l'ADN, pouvant inactiver les éléments autonomes par hétérodimérisation (Hikosaka *et al.* 2007).

Ainsi, deux principaux groupes, appartenant à cette superfamille, peuvent être distingués : les éléments *PBLE* (*piggyBac-like* element) qui contiennent TIR, UTR et ORF codant une transposase qui sont des éléments transposables, et les éléments *piggyBac* domestiqués *PGBD,* non mobilisables, contenant un gène homologue à celui codant la transposase des *PBLE* et qui constituent des gènes de l'hôte à part entière.

2.1. Structure des PBLE

Ces éléments autonomes s'insèrent dans un site cible de quatre nucléotides TTAA, qui est dupliqué lors de l'insertion. L'élément de référence *TniPBLE*, découvert chez *T. ni*, présente une taille d'environ 2475 pb et comporte des TIR de 13 pb commençant par le motif CCCTTT. Comme pour la superfamille des *Tc1-mariner*, les régions UTR comportent des séquences régulatrices de la transcription et de la traduction de la transposase soit une TATA Box et une GC Box dans l'UTR-5' et un signal de polyadénylation dans l'UTR-3'.

Des répétitions inversées sub-terminales (STIR) asymétriques situées à 3 pb de l'extrémité 5'TIR et à 31 pb de l'extrémité 3'TIR peuvent être également présentes (Cary *et al.* 1989; Fraser *et al.* 1996; Lobo *et al.* 1999). De plus, des répétitions directes (DR) situées au niveau des deux UTR, peuvent être détectées, tel est le cas de l'élément *SfrPBLE-wu* isolé chez le lépidoptère *Spodoptera frugiperda* (Wu et Wang 2014). Si les TIRs sont présentes dans toutes les copies complètes, les STIRs et les DRs sont facultatives (les détails de ces structures seront présentés dans la partie résultat) Ces répétitions terminales (TIR) et sub-terminales (STIR et DR) seraient reconnues spécifiquement par la transposase (Mitra *et al.* 2008) (**Figure 14**).

La transposase de 594 aa, codée à partir d'un cadre de lecture unique, est composée de deux domaines fonctionnels (**Figure 14**) : un domaine catalytique et un domaine de fixation à l'ADN :

- Le domaine catalytique, est caractérisé par la présence de trois résidus Aspartate D^{268}, D^{346} et D^{447} hautement conservés, qui assurent les réactions catalytiques de clivage de l'ADN et d'intégration (Sarkar *et al.* 2003; Keith *et al.* 2008; Mitra *et al.* 2008). Dans les cultures cellulaires, un quatrième motif D^{450} pourrait éventuellement être impliqué dans l'excision des *piggyBac.* Toutefois, le rôle de ce résidu n'est pas bien élucidé étant donné que son remplacement par un glutamate (E – même groupe d'aa) peut être toléré (Keith *et al.* 2008).
- Le domaine de fixation à l'ADN, situé dans la région C-terminale, est riche en cystéine formant un motif en doigts de zinc. Sept cystéines sont présents dans la transposase de *T. ni*. L'espacement entre ces résidus est relativement bien conservé.

Par ailleurs, Keith *et al.* (2008) ont suggéré que la région *ZnF* pourrait être impliquée dans l'interaction protéine-protéine requise pour une dimérisation putative de la transposase.

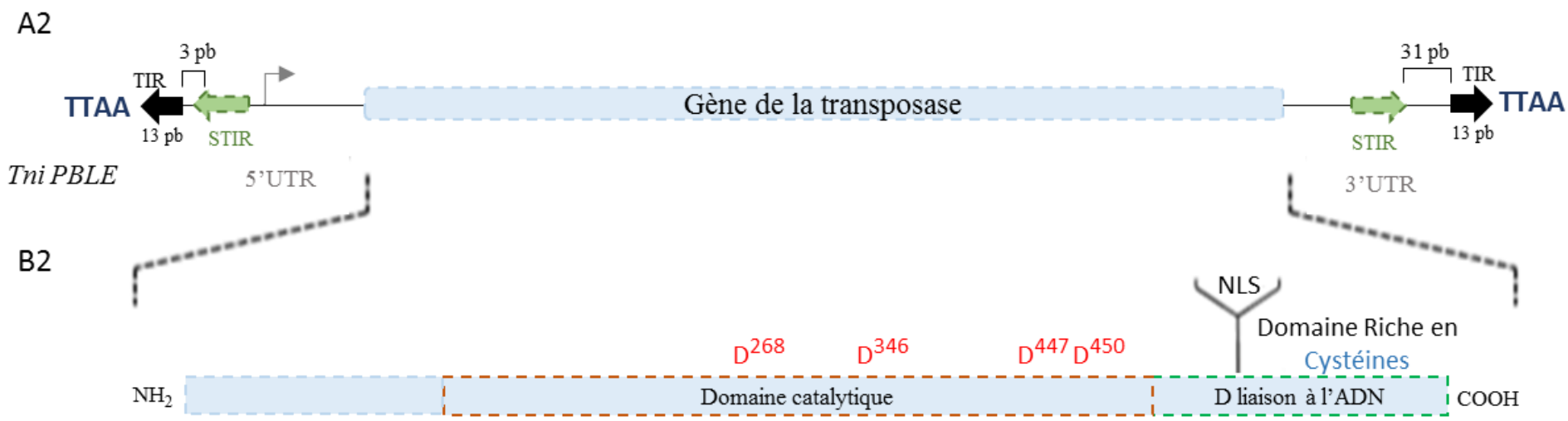


Figure 14. Structure de l'élément *piggyBac-like element* (PBLE).
(A2) Schéma général de la structure des transposons appartenant à la famille des *piggyBac* (*e.g. TniPBLE* identifié chez *Trichoplusia ni*). Les répétitions terminales inversées (TIR) encadrent les régions UTR (Untranslated region) qui comportent des répétitions sub-terminales inversées (STIR) et des sites impliqués dans la transcription et la traduction. Un seul gène code la transposase. Le site cible est TTAA. Les flèches en noir et en vert représentent, respectivement, les TIR et STIR. La flèche grise, dans la région UTR, correspond à la position du promoteur.
(B2) Schéma de structure de la transposases de *TniPBLE*. La transposase comprend le domaine catalytique avec trois acides Aspartiques. Le domaine de liaison situé dans la région C-terminale inclut une région riche en cystéines ainsi que le motif de signal de localisation nucléaire NLS. (D'après Keith *et al.* 2008).

2.2. Structure des séquences domestiquées PGBD

Toutes les séquences domestiquées se trouvent en une seule copie dans les génomes. Chez certaines d'entre elles, la séquence codante dérivée de la transposase est fusionnée avec un gène de structure ou un domaine particulier (Sarkar *et al.* 2003).

- PGBD1 est le résultat d'une fusion ancestrale entre une séquence codant une protéine dérivée d'un élément *piggyBac* de taille de 519 aa et une séquence codant un domaine riche en leucine (LER ou SCAN) de 290 aa, dans la région N-terminale.
- PGBD2, correspondant à la fusion de plusieurs exons, code une protéine de taille de 592 aa dont la fonction reste inconnue.
- PGBD3, qui code une protéine de 593 aa, est une séquence insérée dans le cinquième intron du gène responsable du syndrome du groupe B de Cockayne (CSB). L'épissage alternatif de cette région conduit à la formation d'une protéine de fusion CSB-PGBD3 (Newman *et al.* 2008).
- PGBD4 code une protéine de taille de 585 aa, de fonction inconnue.
- PGBD5, le groupe le plus ancien, contient plusieurs introns et code une protéine de taille variant de 414 à 595 aa (Pavelitz *et al.* 2013). Ce gène semble impliqué dans des fonctions neuronales.
- PGM et TPB2 contiennent également plusieurs introns (2 et 12) et codent respectivement pour des protéines de tailles 1065 et 1220 aa (Baudry *et al.* 2009; Cheng *et al.* 2010), impliqués dans les réarrangements programmés des macronoyaux chez les ciliés.
- Enfin, *KOBUTA* code une protéine de taille de 610 aa (Hikosaka *et al.* 2007) de fonction inconnue.

Par ailleurs, il est important de souligner que dans la majorité de ces séquences domestiquées, le motif DDD n'est pas conservé. De plus, chez PGBD1 et PGBD5, le domaine de fixation à l'ADN de la région C-terminale est absent.

Etant donné les différences de structures observées entre les PGBD, il est probable qu'ils résultent de différents événements de domestication.

2.3. Mécanisme de transposition des PBLE

La mobilisation des *PBLE* se fait en quatre étapes (Mitra *et al.* 2008) (**Figure 15**) :

La coupure de l'élément commence avec une entaille à la jonction de l'extrémité 3' du transposon avec l'ADN flanquant donneur.

Le 3'OH libre au niveau du transposon agit comme un nucléophile et coupe en 5' la jonction entre TTAA et l'ADN flanquant donneur, générant après repliement une structure en épingle à cheveux à l'extrémité du transposon par complémentarité entre AA et TT, et libérant l'ADN donneur. Ce dernier comporte des extensions complémentaires simple brin 5'TTAA qui vont s'apparier et se souderpar le système de réparation.

La structure en épingle à cheveux à l'extrémité du transposon est ensuite excisée par la transposase pour se déplier et libérer en 5' l'extrémité cohésive TTAA des deux côtés du transposon.

Le transposon est enfin intégré de façon covalente dans un autre site cible. Cette intégration a pour conséquence la duplication du site cible.

Il est important de noter que l'excision de cet élément est précise, et ne laisse pas de duplication du site cible, contrairement à l'excision des éléments de type *mariner*.

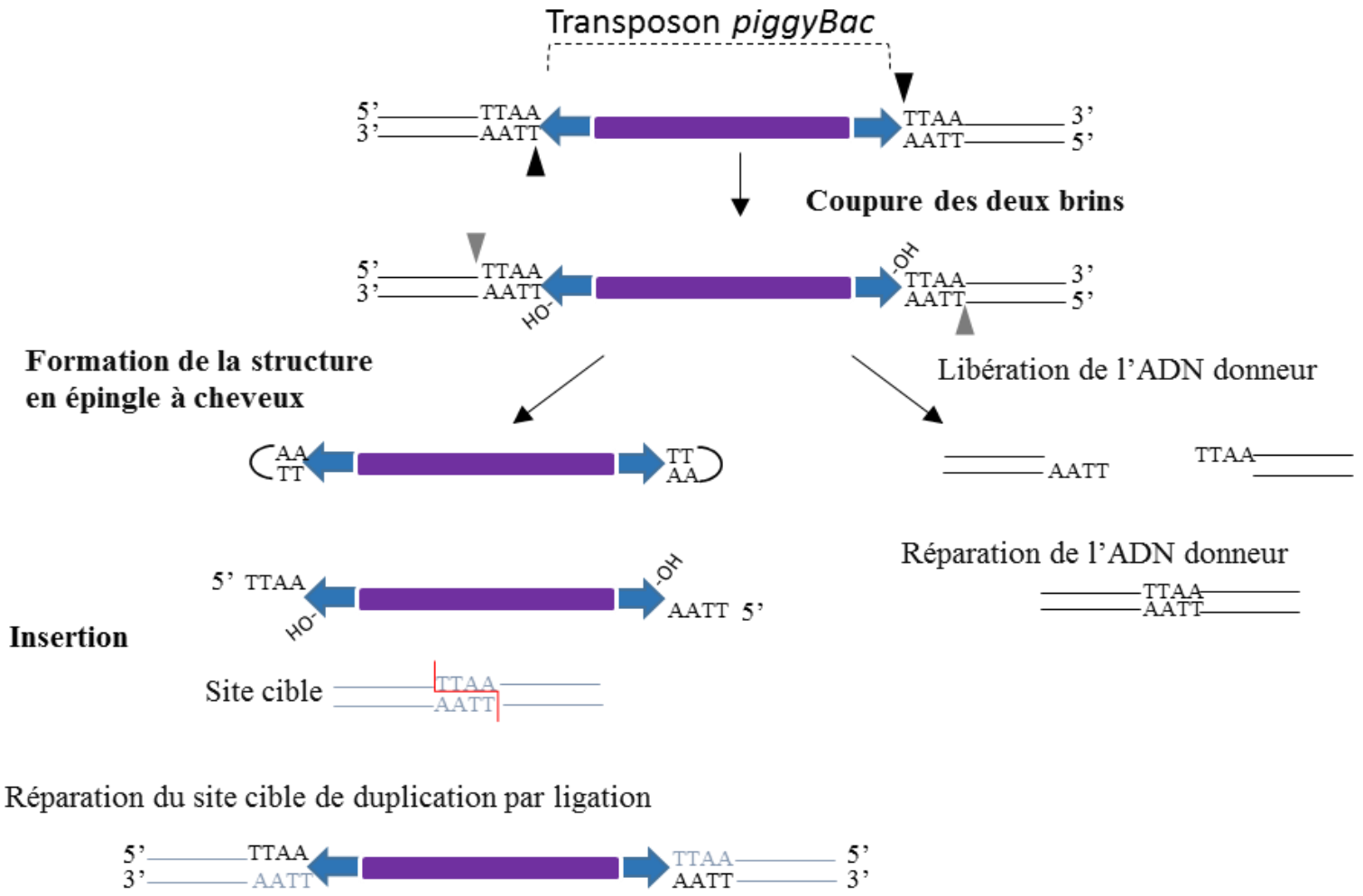


Figure 15. Mécanisme de transposition « couper-coller » des *piggyBac* (adapté de Mitra *et al.* 2008)
La séquence du site donneur est en noir, celle du site receveur en bleu, la coupure de ce dernier, catalysée par la transposase, est symbolisée en rouge. Une double excision est effectuée pour libérer d'une part le site donneur et d'autre part, le transposon avec une extrémité 3' OH et une extrémité 5' TTAA. Il se forme alors une structure en épingle à cheveux qui sera à son tour excisée puis le transposon est réintégré dans un nouveau site cible qui est dupliqué.

Délimitation du sujet

Les récents développements de la génomique ont permis de mettre en évidence le rôle considérable des ETs dans la plasticité structurale et fonctionnelle des génomes. Ils sont ainsi des acteurs majeurs de l'adaptabilité et de l'évolution des espèces à moyen et à long termes. Par ailleurs, leurs caractéristiques insertionnelles sont de plus en plus utilisées en biotechnologie pour développer notamment des vecteurs de transfert de gènes.

Les superfamilles *Tc1-mariner* et *piggyBac*, identifiées chez divers Eucaryotes, notamment les insectes, appartiennent à la Classe II qui transpose selon un mécanisme conservatif de type « couper-coller » et présentent une structure simple avec un gène unique qui code une transposase. La plupart de ces éléments sont inactifs en raison de l'acquisition de mutations (ponctuelles et/ou indels) ou de leur inactivation par « gene silencing ». Cependant, certains transposons naturels sont actifs comme *mos1*, *Mboumar9* ou *piggyBac-Like element* ou *PBLE*. D'autres sont des éléments synthétiques, reconstruits à partir des séquences consensus déduites d'un ensemble d'éléments non-autonomes comme *Sleeping beauty*, *Frog Prince* ou *Himar*.

Dans cette thèse, notre objectif est d'étudier les ETs de la superfamille *Tc1-mariner-IS630* (spécifiquement la famille *mariner*) et *piggyBac* dans les génomes des pucerons des céréales (*Rhopalosiphum padi, R. maidis, Sitobion avenae, Schizaphis graminum*) et de leurs plantes-hôtes (céréales : orge, blé, avoine, égilope). Par ailleurs, en raison de la forte interdépendance des espèces considérées, nous avons cherché à savoir si des transferts horizontaux avaient pu avoir lieu entre le couple plante-hôte/ravageur.

Dans le premier chapitre, nous avons cherché *in silico* les ETs appartenant à la famille *mariner* et à la famille *rosa* initialement définie comme une variante de *mariner* chez l'espèce modèle *Acyrthosiphon pisum* (puceron vert du pois), chez *Diuraphis noxia* (puceron russe du blé) et chez *Myzus persicae* (puceron vert du pêcher). Ces trois espèces de pucerons appartiennent à la même tribu des Macrosiphini et leur génome est entièrement séquencé.

Dans le deuxième chapitre, nous nous sommes intéressés à la caractérisation *in vitro* des éléments de la sous-famille *irritans* aussi bien chez les pucerons des céréales *Rhopalosiphum padi, R. maidis, Sitobion avenae, Schizaphis graminum* que des plantes-hôtes (céréales : orge, blé, avoine, égilope). Une analyse *in silico* de ces éléments dans les génomes séquencés des céréales a été également entreprise.

Le dernier chapitre a été consacré à l'étude *in silico* de la superfamille *piggyBac* dans un large éventail de génomes Eucaryotes afin de caractériser leur polymorphisme, leur distribution et leur évolution intra et inter-familles.

Les résultats sont présentés sous la forme d'articles en anglais structurés en introduction, matériel et méthodes spécifiques, résultats et discussion. Chaque article est encadré par une introduction et une conclusion en français. Enfin, une partie reprenant l'ensemble des résultats pour les replacer dans un contexte plus général, les discuter et proposer des perspectives viendra clore ce manuscrit.

Matériel et Méthodes

Ce chapitre est consacré à la description du matériel animal (pucerons) et végétal (céréales sous-famille des *Pooideae*) ainsi que des différentes méthodes utilisées, depuis l'extraction de l'ADN, jusqu'au traitement des séquences en vue de leur exploitation dans les analyses génomiques et phylogénétiques.

I. Matériel biologique

Des pucerons de céréales ont été prélevés au centre régional des recherches en grandes cultures de Béjà (CRRGC) au cours des campagnes de 2012 à 2014. L'identification des espèces *Rhopalosiphum padi*, *R. maidis*, *Sitobion avenae*, *Schizaphis graminum* est basée sur les critères morphologiques décrits par Blackman et Eastop (2000). Des feuilles de plusieurs variétés d'orge (Rihane, Manel, Martin, Bowman, Roho, HD29), de blé (BLG, Salambo) et celles d'avoine ont été également collectées. Des graines d'aegilops fournies par le CIMMYT (The International Maïze and Wheat Improvement Centre, Mexico) ont été utilisées. Enfin, des échantillons du puceron vert du pois *Acyrthosiphon pisum*, nous ont été aimablement fournis par l'Institut National de Recherche Agronomique de Tunis (INRAT).

II. Méthodes

1. Extraction de l'ADN

L'ADN total est extrait selon deux méthodes différentes :

1.1. Méthode de Doyle et Doyle (1987)

Ce protocole ancien et long, basé sur l'utilisation du CTAB (hexadecyltrimethyl-ammonium bromide), est efficace et permet d'obtenir jusqu'à 200 µg d'ADN par insecte. Le broyage des pucerons est effectué à l'aide d'un piston conique dans un tube Eppendorf contenant 50µl de tampon d'extraction (2% CTAB ; 1,4M NaCl ; 0,2% 2-mercaptoéthanol; 20mM EDTA ; 100mM Tris HCl ; pH 8) préalablement chauffé à 65°C. Le broyat est additionné de 150µl du tampon d'extraction et incubé à 65°C pendant 1h. Après traitement par un volume de chloroforme-alcool isoamylique (24 : 1), une centrifugation pendant 10 min à 8000 rpm à 4°C permet de récupérer le surnageant dans un tube Eppendorf contenant un volume égal d'isopropanol. Après une nuit à -20°C, le culot contenant les acides nucléiques est récupéré suite à une centrifugation de 15 min à 13500 rpm et à 4°C. Le culot est ensuite lavé avec 250µl d'éthanol 70% et séché au Speed-Vac avant d'être repris dans 20µl de H_2O bi-distillée stérile.

Pour les végétaux, l'extraction de l'ADN a été effectuée suivant le même protocole. Le broyage a été fait dans un mortier contenant de l'azote liquide, puis dans un volume de 150µl de tampon d'extraction, avant d'ajouter de 450µl du tampon d'extraction.

1.2. Méthode basée sur l'utilisation des kits d'extraction

Cette méthode suit les spécifications des fournisseurs des kits d'extraction rapides Wizard® Genomic DNA Extraction Kit (Promega®) ou NucleoSpin® Tissue DNA Kit (Macherey-Nagel) qui permettent d'obtenir entre 10 à 50 µg d'ADN. Le broyage du matériel biologique s'effectue dans le tampon d'extraction fourni dans le kit.

Quelle que soit la méthode d'extraction utilisée, la pureté de l'ADN est ensuite vérifiée par la mesure de la DO à 260nm et 280nm dans un spectrophotomètre (Nanodrop) ou/et un fluoromètre (Qubit 3.0). Le rapport DO260/280 doit être compris entre 1,8 et 2.

2. Amplification et purification de l'ADN

La PCR « Polymerase Chain Reaction » (Mullis 1990) comporte une phase de dénaturation initiale de 5 min à 94°C, environ 30 cycles d'amplification et une phase d'élongation finale de 10 min à 72°C. Chaque cycle est composé de trois étapes: une étape de dénaturation de l'ADN matrice à 95°C, une étape d'hybridation des amorces à une température variable selon la séquence de l'amorce utilisée et une étape d'élongation à 72°C.

Le milieu réactionnel comprend de 50ng à 100ng d'ADN, 5µl tampon Taq 5X (Promega®), $MgCl_2$ 2.5mM, dNTP 0.2mM, amorces 1µM, GoTaq Polymérase à raison de 1U (GoTaq™ DNA Polymerase, Promega®) dans un volume final de 25µl. Les amorces utilisées seront décrites dans les différents chapitres.

Par ailleurs, une autre variante de la PCR, la PCR-inverse (iPCR), a été utilisée pour identifier les régions flanquantes d'un élément dans un génome de plantes ou d'insectes (Ochman *et al.* 1988). Celle-ci présente les mêmes étapes d'amplification, toutefois, elle est précédée d'une étape de digestion de l'ADN génomique (500ng) par 10U d'une endonucléase (SmaI, HaeII ou EcoRII; Promega®) durant 5h à 37°C. Les fragments obtenus sont alors circularisés dans 150 µL du tampon de ligation avec 100U de T4 DNA ligase (Promega®) à 37°C pendant 5h. Après l'inactivation de cette enzyme à 65°C pendant 10min, 10µl de l'ADN génomique ligué est utilisé pour réaliser la PCR inverse, avec des amorces spécifiques.

Après migration sur gel d'agarose des produits PCR/iPCR obtenus, la purification des fragments d'ADN est effectuée en utilisant le kit Wizard Genomic DNA Purification (Promega®). La solution d'ADN est ensuite stockée à -20°C.

3. Clonage et purification des plasmides

3.1. Préparation des bactéries compétentes

Pour effectuer le clonage, une pré-culture de bactéries *Escherichia coli DH5α* (lacZ-) sensibles à l'ampicilline est réalisée pendant une nuit à 37°C, dans 3ml de milieu SOB : Super Optimal Broth (Bacto tryptone 2% ; extrait de levure 0.5% ; NaCl 10mM ; KCl 2,5mM ; $MgCl_2$ 10mM ; $MgSO_4$ 10mM). Puis, 1ml de la pré-culture est repiqué à nouveau dans 100ml de milieu SOB et le mélange est soumis à une agitation modérée à 37°C jusqu'à l'obtention d'une densité optique DO à 600nm comprise entre 0,45 et 0,5.

La culture est ensuite transférée dans des tubes Falcon et placée dans la glace pendant 15 min. Après 20 min de centrifugation à 3000 rpm à 4°C, le culot obtenu est re-suspendu dans 20ml d'une solution mère froide (1ml de $CaCl_2$ 4M, 1ml de $MnCl_2$ 1M, 1ml de $MgCl_2$ 0,5M, 500µl d'Acétate de potassium 1M et 5ml de glycérol dans un volume final de 50ml), puis mis dans la glace pendant 30 min.

Le mélange est centrifugé pendant 30 min à 3000 rpm, le culot est re-suspendu dans 3ml de solution froide. La suspension de bactéries compétentes est incubée pendant 30 min dans la glace, aliquotée puis conservée à -80°C.

3.2. Préparation du vecteur recombinant par ligation

Les produits PCR purifiés sont clonés dans le vecteur plasmidique pGEM T Easy Vector (Promega®). Ce vecteur contient, en plus de l'origine de réplication, un gène de résistance à l'ampicilline et un fragment du gène LacZ codant pour une β-Galactosidase active comportant une séquence MCS (Multiple Cloning Site), ayant plusieurs sites uniques de restriction, encadrée par les promoteurs T7 et SP6. L'insertion d'un fragment d'ADN est assurée par l'enzyme T4 DNA ligase dans un mélange réactionnel composé de 50ng du vecteur pGEM-T Easy, 10U de T4 DNA ligase, 5µl du tampon de ligation (2X) et d'environ 50ng de produit PCR dans un volume final de 10µl. Le mélange est incubé pendant une nuit à 4°C.

3.3. Transformation et sélection des bactéries

10 µl de produits de ligation sont ajoutés à 100µl de bactéries *DH5α.* Après une incubation pendant 30 min dans la glace suivie d'un choc thermique à 42°C pendant 45sec puis une ré-incubation dans la glace pendant 2 min, 900µl de milieu LB liquide sont ajoutés. Après avoir incubé sous agitation le mélange pendant 1h à 37°C, une centrifugation pendant 3 min à 5000 rpm est effectuée et le culot est re-suspendu dans 100µl de LB liquide. 70µl de bactéries sont ensuite étalés sur un milieu LB solide contenant 100µg/ml d'ampicilline, 40µl X-gal comme

substrat (2%) et 40µl IPTG à 20mg/ml comme inducteur. Les boites sont ensuite incubées à 37°C toute la nuit. Les colonies recombinantes de couleur blanche sont ensuite récupérées, mises en culture dans du milieu LB liquide à raison de 3ml en présence d'ampicilline (10µg/µl) et incubées pendant une nuit à 37°C sous agitation. L'ADN plasmidique est extrait en utilisant le kit «Wizard® Plus Minipreps DNA purification system » (Promega®). Afin de vérifier la présence d'un insert, une PCR par les amorces T7 et SP6 (Tableau 1) est réalisée dans un volume réactionnel de 25µl contenant 5µl de tampon PCR (5X), 1µl d'ADN plasmidique, 1µl de chacune des amorces T7 et SP6 (10µM), 0,15µl de l'enzyme Taq Polymérase, 2µl de $MgCl_2$ 25mM, 1,5 µl de dNTP à 10mM. Après confirmation par électrophorèse sur un gel d'agarose à 1%, l'insert est séquencé en utilisant les amorces T7 et SP6 dans un séquenceur automatique ABI Prism 3100.

Tableau 1 : Amorces spécifiques utilisées pour vérifier la présence de l'insert et pour le séquençage

Amorces	Séquences 5'➞ 3'	T°	Tailles attendues (pb)
T7	TAATACGACTCACTATAGGG	53	Taille de l'insert + ~170pb
SP6	ATTTAGGTGACACTATAG		

4. Filtrage et vérification des séquences

Les séquences ainsi obtenues sont utilisées comme requête dans BLAST (http://www.ncbi.nlm.nih.gov/blast) sur NCBI-NR (base de données non redondante) afin de vérifier qu'elles correspondent bien aux types de séquences attendues (Altschul *et al.* 1990). Plusieurs variantes de ce programme sont proposées :

- BLASTN est une recherche par similarité à partir de séquences nucléotidiques dans une base de données nucléotidiques.
- BLASTP est une recherche par similarité à partir de séquences protéiques dans une base de données protéiques.
- BLASTX est une recherche par similarité à partir de séquences nucléotidiques traduites sur une base de données protéiques.
- TBLASTN est une recherche par similarité à partir de séquences protéiques sur une base données nucléotidiques traduites.

Plusieurs informations sont également fournies : le pourcentage d'identité, la position et l'orientation de la séquence extraite de la base de données par rapport à la séquence requête, la longueur de recouvrement, l'E-value décrivant le bruit de fond aléatoire.

5. Recherche in silico des éléments transposables dans les génomes

La recherche dans les génomes séquencés a été réalisée selon l'approche basée sur l'homologie de séquences par rapport à une séquence requête (query) en utilisant le programme BLAST (http://www.ncbi.nlm.nih.gov/blast).

6. Alignement et traitement des séquences

L'alignement des séquences est une étape fondamentale afin de déterminer le pourcentage de similitude, d'estimer la contrainte sélective (K_A/K_S), de classer les éléments et de construire des arbres phylogénétiques. Cet alignement est effectué par le logiciel Aliview 1.18 qui utilise le programme MUSCLE. Les matrices d'identité sont obtenues *via* MEGA6 (Tamura *et al.* 2013). Les motifs conservés sont visualisés par le logiciel GENEDOC (Nicholas *et al.* 1997) ou Weblogo (Crooks *et al.* 2004). Par ailleurs, les séquences nucléotidiques obtenues sont traduites grâce au logiciel Expasy TranslateTool portail disponible sur le web (http://web.expasy.org/translate/), puis optimisées manuellement. Par ailleurs, la mise en évidence des motifs HTH est effectuée par GYM2.0 (Gao *et al.* 1999; Narasimhan *et al.* 2002). Les signaux de localisation nucléaire (NLS) sont prédits par PSORTII (Bannai *et al.* 2002).

7. Estimation de la contrainte sélective

Le rapport du nombre de substitutions non-synonymes K_A sur le nombre de substitutions synonymes K_S informe sur le type de pression de sélection qui s'est exercé au cours du temps : sélection purifiante (K_A/K_S <1), sélection neutre (K_A/K_S =1), sélection diversifiante (K_A/K_S >1) (Hurst 2002). Ce type d'analyse est réalisé sur des paires de séquences, dans le cas d'une matrice de n séquences, chaque séquence est comparée à chacune des n-1 autres. Ce rapport est calculé par le logiciel MEGA6 (Tamura *et al.* 2013).

8. Classification par la méthode agrégative UPGM-VM (Unweighted Pair Group Method with Variation Metric)

Cette méthode de classification s'inspire de la méthode classique UPGMA (Unweighted Pair Group Method with arithmetic Mean; Sneath et Sokal 1973). Les algorithmes utilisés sont basés sur des stratégies adaptatives permettant d'améliorer l'ajustement par une combinaison d'un grand nombre de modèles tout en éliminant le poids des insertions/délétions ou indels (Rouault *et al.* 2009). Elle permet d'intégrer des séquences nucléotidiques de tailles très différentes. Les deux principales différences avec la méthode UPGMA sont d'une part l'absence de consensus (la distance entre deux groupes est définie par la moyenne des distances entre tous les individus des deux groupes) et d'autre part, la variation de la métrique au cours

du processus de calcul qui commence par une distance globale (identité stricte des séquences) pour devenir progressivement locale (ignorant les *indels*).

Les programmes de cette méthode (**Figure 16**) sont écrits en langage ADA utilisable sous Windows et adaptable sous Linux. Ces programmes ne prennent en compte que les noms des fichiers à traiter « Input » et le format des fichiers « Output » souhaité. Enfin, le logiciel Scilab permet la visualisation de la classification sous forme de rosace (Rouault *et al.* 2009).

9. Construction des phylogénies moléculaires

Elle a pour but de reconstruire, par comparaison nucléotidique ou d'acides aminés, l'histoire évolutive et de déterminer les relations de parenté entre des séquences. Pour tirer le meilleur parti des méthodes phylogénétiques, il est important de choisir le modèle d'évolution le plus approprié pour le jeu de données. Dans ce contexte, le meilleur modèle pour les séquences protéiques est établi par le logiciel ProtTest (http://darwin.uvigo.es/software/prottest2_server.html).

Plusieurs méthodes peuvent être exploitées telles que la méthode des proches voisins ou Neighbor-Joining NJ (Saitou et Nei 1987), la méthode du maximum de parcimonie ou Maximum Parcimony (Moore *et al.* 1973) et la méthode du maximum de vraisemblance ou Maximum Likelihood ML (Hutchinson 1929). Les particularités de chacune d'elle, sont les suivantes :

- **Neighbor-Joining (NJ) :** C'est une méthode phénétique. Elle se base sur l'utilisation d'une matrice des distances globales. Le calcul des distances nécessite la clustérisation des séquences, prenant en compte le nombre de caractères qui diffèrent entre les séquences prises deux à deux. La méthode NJ utilise la matrice triangulaire de distances pour construire l'arbre phylogénétique qui se fait en deux étapes : (i) la première consiste à choisir les séquences les plus proches et à les lier par un nœud représentant leur ancêtre commun hypothétique, (ii) la seconde consiste à recalculer les distances entre le nœud obtenu et toutes les autres séquences pour convertir l'ancêtre commun en un nœud terminal.
- **Maximum Parcimony (MP) :** Elle consiste à minimiser le nombre total d'évènements évolutifs (mutations/substitutions) et à passer d'une séquence à une autre dans une topologie d'arbre en utilisant le chemin évolutif le plus court (minimum de transformation). Elle construit l'arbre le plus parcimonieux, optimal, parmi l'ensemble de tous les arbres possibles.

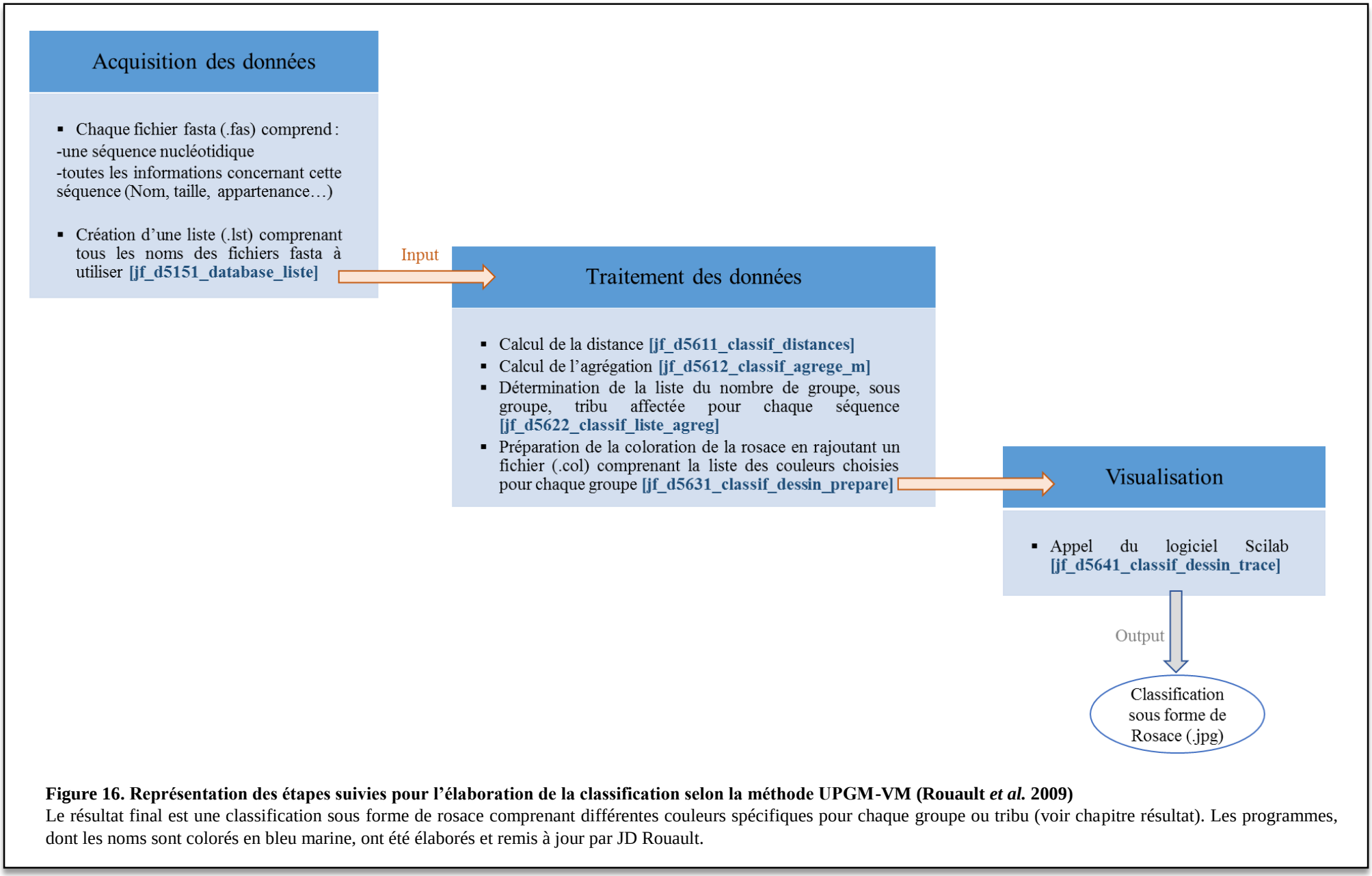


Figure 16. Représentation des étapes suivies pour l'élaboration de la classification selon la méthode UPGM-VM (Rouault *et al.* 2009)
Le résultat final est une classification sous forme de rosace comprenant différentes couleurs spécifiques pour chaque groupe ou tribu (voir chapitre résultat). Les programmes, dont les noms sont colorés en bleu marine, ont été élaborés et remis à jour par JD Rouault.

- **Maximum Likelihood (ML) :** Cette méthode établit des relations phylogénétiques à partir d'un modèle explicite de changements évolutifs, en recherchant l'arbre qui maximalise la vraisemblance de l'échantillon. Suite au calcul de la vraisemblance des différents arbres, leurs scores sont comparés deux à deux. La durée du calcul augmente donc exponentiellement avec le nombre d'échantillons, en raison du plus grand nombre d'arbres possibles. Cette méthode est fortement recommandée pour les séquences phylogénétiquement éloignées.

Par ailleurs, le support des branches est évalué par la méthode de bootstrap. Il s'agit d'un test de ré-échantillonnage au hasard et avec remise des caractères qui permet d'évaluer la robustesse de la reconstruction en fonction des caractères. L'opération est répétée généralement de 500 à 1000 fois et la robustesse d'un nœud est estimée par le pourcentage d'observation du nœud dans l'ensemble des 500 à 1000 arbres re-construits.

La visualisation d'un arbre se fait directement sur MEGA6. Il est polarisé des feuilles vers la racine. Les nœuds terminaux sont connectés par des branches qui se rejoignent au niveau des nœuds internes. Ces derniers représentent les évènements de spéciation déterminés à partir du jeu de données. L'ensemble des feuilles en partant d'un nœud interne est appelé «clade». Un clade monophylétique inclut tous les descendants d'un nœud, un clade paraphylétique désigne un clade incluant une partie seulement de ses descendants d'une espèce ancestrale. Le clade polyphylétique est quant à lui défini par une ressemblance qui n'a pas été héritée d'un ancêtre commun, mais qui est le fruit d'une convergence évolutive. Généralement, un groupe externe (outgroup) est nécessaire pour pouvoir enraciner l'arbre.

Résultats

Chapitre I

Diversité et évolution des *mariner-like elements* dans les génomes des aphides

Diversity and evolution of *mariner-like* elements in Aphid genomes

Chapitre I

Diversité et évolution des *mariner-like elements* dans les génomes des aphides

L'exploitation des génomes séquencés constitue une approche rapide et efficace pour identifier et caractériser les transposons. A ce jour, trois génomes de pucerons sont disponibles dans les banques de données (NCBI, AphidBase). Le génome du puceron vert du pois, *Acyrthosiphon pisum* a été nouvellement annoté (Acyr_2.0, new reference Annotation Release 102, http://www.ncbi. nlm.nih.gov, Richards et *al.* 2010). Il est constitué de 23925 scaffolds couvrant 541Mb. Le génome du puceron russe du blé *Diuraphis noxia* a été également récemment séquencé et annoté. Un total de 393Mb est couvert par 5641 scaffolds (Dnoxia_1.0 reference annotation release 101, http://www.ncbi.nlm.ih.gov, Nicholson et *al.* 2015). De plus, les données préliminaires du nouvel assemblage du génome complet du puceron vert du pêcher *Myzus persicae* ne sont disponibles que dans AphidBase et présentent une taille de 398Mb, incluant 34598 scaffolds (*Myzus persicae* Clone G006 assembly V2, http://tools.genouest.org/tools/myzus/).

Ces trois espèces parasitent plusieurs plantes hôtes cultivées ou sauvages :

- *A. pisum* s'attaque aux plantes de la famille des fabacées ;
- *M. persicae* est un ravageur du pêcher et mais peut également s'attaquer aux solanacées ;
- *D. noxia* infeste principalement le blé mais aussi d'autres céréales cultivées et sauvages.

Ces hémiptères appartiennent à la tribu des Macrosiphini, ordre des *Aphidinae* et sont phylogénétiquement très proches des pucerons des céréales, objets de notre recherche (*Sitobion avenae* appartient à la même tribu des Macrosiphini alors que *Rhopalosiphum padi, R. maidis* et *Schizaphis graminum* appatiennent à la tribu des Aphidini).

La disponibilité de ces trois génomes offre l'opportunité de rechercher les éléments de la Famille *mariner*, largement distribués chez les insectes et caractérisés par le motif catalytique DD34D, ainsi que ceux du « clade » *rosa* qui a été identifié initialement comme une variante des éléments *mariner* mais ayant un motif catalytique spécifique DD41D (Gomulski et *al.* 2001).

Cette recherche *in silico*, basée sur l'homologie des séquences, porte sur :

- la diversité et la distribution intra et inter-spécifiques de ces éléments ;
- la dynamique et l'histoire évolutive de ces éléments.

Les résultats de ce chapitre ont fait l'objet d'un article soumis dans la revue BMC genomics.

Diversity and evolution of *mariner*-like elements in Aphid genomes

Maryem Bouallègue[1, 2], Jonathan Filée[1], Maha Mezghani-Khemakhem[2], Jacques-Deric Rouault[1], Aurélie Hua-Van[1], Mohamed Makni[2] and Pierre Capy[1*]

[1] Laboratoire Evolution, Génomes, Comportement, Ecologie CNRS, Univ. Paris-Sud, IRD, Université Paris-Saclay, 1 avenue de la Terrasse, 91198 Gif-sur-Yvette Cedex, France;

[2] Université de Tunis El Manar, Faculté des Sciences de Tunis, UR11ES10 Génomique des Insectes Ravageurs de Cultures, 1002, Tunis, Tunisie.

Email addresses: maryem.bouallegue@egce.cnrs-gif.fr, jonathan.filee@egce.cnrs-gif.fr, mahakm@planet.tn, jacquesderic@wanadoo.fr, aurelie.hua-van@egce.cnrs-gif.fr, Md.Makni@fst.rnu.tn, pierre.capy@egce.cnrs-gif.fr.

*Correspondence: Pierre Capy, Centre National de la Recherche Scientifique (UMR 9191), Gif-sur-Yvette, France, Telephone number: 33 1 69 82 37 09, Fax number: 33 1 69 82 37 36, E-mail: pierre.capy@egce.cnrsgif.fr.

Abstract

Background

Although transposons have been identified in almost all organisms, genome-wide information on *mariner* elements in Aphididae remains unknown. Genomes of *Acyrthosiphon pisum*, *Diuraphis noxia* and *Myzus persicae* belonging to the *Macrosiphini* tribe, actually available in databases, have been investigated.

Results

A total of 22 lineages were identified. Classification and phylogenetic analysis indicated that they were subdivided into three monophyletic groups, each of them containing at least one putative complete sequence, and several non-autonomous sublineages corresponding to Miniature Inverted-Repeat Transposable Elements (MITE), probably generated by internal deletions. A high proportion of truncated and dead copies was also detected. The three clusters can be defined from their catalytic site: (i) *mariner* DD34D, including three subgroups of the *irritans* subfamily (*Macrosiphinimar*, *Batmar-like elements* and *Dnomar-like elements*); (ii) *rosa* DD41D, found in *A. pisum* and *D. noxia*; (iii) a new clade which differs from *rosa* through long TIRs and thus designated *LTIR-like* elements. Based on its catalytic domain, this new clade is subdivided into DD40D and DD41D subgroups. Compared to other *Tc1/mariner* superfamily sequences, *rosa* DD41D and *LTIR* DD40-41D seem more related to *maT* DD37D family.

Conclusion

Overall, our results reveal three clades belonging to the *irritans* subfamily, *rosa* and new *LTIR-like* elements. Data on structure and specific distribution of these transposable elements in the *Macrosiphini* tribe contribute to the understanding of their evolutionary history and to that of their hosts.

Key words: Aphids, comparative genomics, *Tc1-mariner*, Transposable elements, MITEs, molecular evolution.

Background

Genomes contain diverse repetitive DNA sequences of transposable elements (TEs), contributing to their plasticity, adaptability and evolution [1-3]. *Class II* TEs use a "cut and paste" mechanism. They are either autonomous transposons encoding their own transposase or non-autonomous transposons including truncated copies (*i.e.* copies with only one or no extremity) or copies with internal deletions, but with two intact extremities. Although not encoding for a functional transposase, these shorter copies or miniature inverted repeat transposable elements (MITEs) can be *trans*-mobilized and may reach high copy number with a size homogeneity that distinguishes them from other non-autonomous elements [4].

The *Tc1/mariner* superfamily is ubiquitous and forms the largest group of eukaryotic *Class II* TEs [5]. Its members share several common characteristics and synapomorphies. In particular, the target insertion site is TA, the ORF of autonomous copies encodes a transposase of 282 to 350 amino-acid residues [6]; the transposase contains two helix–turn-helix (HTH) motifs in DNA binding domains and a catalytic triad DDE/D motif [5,7].

Despite these similarities, two major differences can separate families of *Tc1/mariner*: (i) their complete length from 1 to 5 kb due to their TIR (*i.e.* the *mariner*-like element *MLE* 13–34 bp long, the *Tc1*-like element *TLE* ranging from 20 to 600 bp), (ii) the DDE/D signature motif in their catalytic domains which corresponds to DD34D for *mariner*, DD34E for *Tc1*, DD37D for *maT*, DD37E, DD39D, and DD41D for *rosa* [8-10].

The *mariner* family, initially described in *Drosophila mauritiana* [11], is one of the best known elements belonging to this superfamily. This element is characterized by a patchy and large distribution among metazoans [12-14], which can be explained, in part, by horizontal transfer (HT), corresponding to its ability to transpose between genomes [15-17]. Due to the great diversity of this family, these elements are classified into several subfamilies based on phylogenetic studies. Five major distinct subfamilies including *irritans, mauritiana, cecropia,*

mellifera/capitata, and *elegans/briggsae* were reported [12]. However, sixteen minor subfamilies also exist with a more limited distribution [18-20]. Otherwise, the *rosa* monophyletic group, first identified in *Ceratitis rosa* and other Tephritid flies, is closely related to the *mariner* subfamilies [9,16]. Its main characteristic is a transposase with a DD41D motif, and the nucleotide identity between *MLE* subfamilies is about 40 to 56% [12,21].

While *MLE* is characterized by a high proportion of inactive copies due to independent accumulation of substitution and indels, known as vertical inactivation [22], three elements, namely *mos1,* found in the fruit fly *Drosophila mauritiana* (*mauritiana* subfamily)*, Famar1* discovered in the common earwig *Forficula auricularia* (*mellifera* subfamily) and *Mboumar9* isolated from the ant *Messor bouvieri* (*mauritiana* subfamily) are still naturally active, and thus able to be mobilized [12,23-27]. Furthermore, the *Himar1* element from the horn fly *Haematobia irritans* (*irritans* subfamily) has been reconstructed by *in vitro* mutagenesis to restore a potential activity [28,29]. Due to their wide distribution and ability to successfully invade new genomes by horizontal transmission, naturally and artificially active *mariner* transposons are used as powerful molecular tools in transgenesis and insertional mutagenesis, *inter alia* leading to genetic control strategies of pests [29-32].

In plant aphid species, only a few studies have described the presence of *mariner* elements. For instance, (i) internal partial sequences of *irritans* and *mellifera* subfamilies were identified *in vitro* by a Polymerase Chain Reaction (PCR) amplification in the soybean aphid *Aphis glycines* [33], (ii) deleted sequences of *mauritiana* subfamily were characterized in seven fruit tree aphid species [34], (iii) in the first version of pea aphid *Acyrthosiphon pisum* genome [35], only three complete sequences, namely *Mariner-Ap_1, 2* and *3,* were published in RepBase [36]. However these sequences shared catalytic motif DD34E and should be more related to *Tc1*-elements.

Nowadays, three aphid's genomes are available in public databases. Indeed, the recent sequencing of the Russian wheat aphid *Diuraphis noxia* genome (Dnoxia_1.0 reference annotation release 101, http://www.ncbi.nlm.ih.gov) [37], the green peach aphid *Myzus persicae* genome (AphidBase, http://tools.genouest.org/tools/myzus/), and the new annotation of *A. pisum* genome (Acyr_2.0, new reference Annotation Release 102, http://www.ncbi.nlm.nih.gov) offer an opportunity to investigate the diversity of the *mariner* family within and between aphid species, along with the evolutionary history and dynamics of these elements.

These species belong to the Macrosiphini tribe of the Aphididae family and diverged approximately 42.5 Mya [38]. They are found on different host plants: *A. pisum* on Fabaceae, *D. noxia* on cereals and *M. persicae* on peach trees or Solanaceae.

In this paper, we explored these three genomes in order to identify *mariner*-related transposons and their non-autonomous derivatives using a library-based method. Eleven TE clusters from *A. pisum*, seven from *D. noxia* and four from *M. persicae* have been detected. Classification and phylogenetic analysis suggested (i) that these lineages are divided into three groups: the *irritans* subfamily DD34D, *rosa* DD41D and a new group DD40/41D close to *rosa* and characterized by a long TIR, (ii) an evidence of vertical transfer with stochastic losses and several putative HT events. All these data provide new informations about the evolutionary history of these transposable elements in aphids.

Methods

Supporting data

The genome of *Acyrthosiphon pisum* and *Diuraphis noxia* are available at NCBI (http://www.ncbi. nlm.ih.gov). The first contains 541Mb covering 23925 scaffolds and the second includes 393Mb covering 5641 scaffolds [35,37]. The genome of *Myzus persicae*, presenting 398Mb and spanning 34598 scaffolds, is published in aphidbase (The International Aphid Genomics Consortium http://tools.genouest.org/tools/myzus/).

Data mining

A panel of 18 transposases sequences belonging to the five major *mariner* subfamilies DD34D and to the *rosa* DD41D group (Additional file 1) were used as queries in tBLASTN searches on the three aphid genomes, with default parameters. In order to determine the full sequence of each copy, the best hits were extracted with 5 kb flanking sequences and were manually investigated for TIR searches. Each new complete sequence was then used to retrieve more elements. Truncated copies located at the end of scaffolds and sequences less than 250bp were further discarded. The sequences closer to DDxE catalytic motif were excluded after a BLASTX search against transposases from this family. Finally, 115 sequences from *A. pisum*, 45 from *D. noxia* and 23 from *M. persicae* were obtained and used in this work.

Sequence analyses

The nucleotide sequence analyses, including alignment, were done with the Aliview 1.18 [39]. USEARCH6.0 [40] was performed to cluster repetitive sequences using a threshold of 75% identity. Shorter copies flanked by two TIRs and with evidence of transposition (at least 2 copies) were considered as MITEs [4,41]. Consensus sequences were derived using the relative majority rule.

The putative amino acid sequences were deduced by ExpasyTool (http://web.expasy.org/translate/) and then manually optimized. The nuclear localization

sequence (NLS) and the helix turn helix (HTH) domain were searched using PSORTII [42] and GYM2.0 [43,44], respectively (Additional file 2).

Mining of available eukaryote genomes

The complete nucleotide sequences previously identified were used in BLASTn searches against the nr (non-redundant nucleotide) and WGS (whole genome sequence) databases available on the NCBI. Sequences with more than 60% of nucleotide identity over more than 65% of the length of the query were extracted. These thresholds have been chosen to avoid recovering small fragments and sequences phylogenetically far from the subfamilies here considered. Cases of potential horizontal transfers between aphids and other taxa are considered when elements present more than 90% of identity covering more than 90% of the query sequences as proposed by several authors [17,20].

Classification and phylogenetic analysis

The classification is based on the Unweighted Pair Group Method with Variation of Metric UPGM-VM [19], an ascending hierarchical classification analogous to the UPGMA method, with two main differences: (i) there is no arithmetical mean, the nucleotide sequences are aligned by pairs, (ii) the metric varies with the ascending classification and gap is considered as a fifth nucleotide. This variation allows a complete sequence to be gathered in the same group with the corresponding truncated and/or deleted sequences such as MITEs. Thus, the 183 elements extracted from aphid genomes were added to a set of 96 already known complete sequences from the *Tc1-mariner-IS630* superfamily published in GenBank and to 50 sequences found in eukaryote genomes (Additional file 3). MITE classification is based on identity of TIRs, internal sequences of complete transposable elements and on the breakpoints of deletions. For phylogenetic analysis, the amino acid sequences were aligned with Aliview1.18 [39] and the best-fitting ML model (AIC, matrix WAG+F+I+G) was selected using Protest 2.4 server [45]. Then, the phylogenetic analysis of transposases was computed using MEGA6 [46] with 1000 bootstrap replicates.

Results

Distribution and diversity of mariner and rosa elements within the Macrosiphini tribe

Search of sequences belonging to the main *mariner* subfamilies DD34D and to the *rosa* DD41D group was based on a homology approach (tBLASTN) using a set of 18 known transposases as queries (Additional file 1). We found a total of 115 copies from *A. pisum* clustered in eleven lineages, 45 from *D. noxia* clustered in seven lineages and 23 copies from *M. persicae*

distributed in four lineages. A lineage corresponds to a group of sequences that is more than 75% similar and to clear phylogenetic clades (see below).

While 183 copies were extracted, 23 complete and potential autonomous sequences, representing 12.57% of all copies, have been identified in aphid genomes. A low copy number, ranging from one to six, per lineage and per species is observed. More precisely, only ten sequences distributed into nine lineages are found in *A. pisum* genome. All these sequences are named *Apismar.* For *D. noxia*, seven complete copies (*Dnomar*) are grouped into five lineages and only six copies from *M. persicae* (*Mpmar*) are gathered in the same group.

For most of these clusters (14 out of 15), the terminal inverted repeats (TIRs) necessary for transposition have been identified, as well as the TA target site duplication (TSD). The *Apismar4.2* does not display a TSD. Interestingly, the whole nucleotide sequences appear heterogeneous in length. Some clusters with a short TIR (15-32bp) have a full length of approximately 1.3 kb (*i.e. Apismar1.2, Apismar4.1*), while others (*i.e. Apismar5.1*, *Apismar5.2*) showed sizes longer than 2 kb due to long TIR sequences about 460bp (Table 1).

Classification of the 183 aphid sequences, based on the 146 nucleotide sequences of the *Tc1/mariner* superfamily, was performed using a UPGM-VM method. This allows all sequences to be dealt with whatever their length, including the distantly related *Tc1* and *Tc3* sequences of animals, plants, fungi and bacteria like *IS630* (Figure 1, Additional file 3).

Results reveal that 75 copies (18 complete elements and 57 deleted/truncated sequences) belong to the *irritans* subfamily. They can be subdivided into three tribes: the first is specific to aphids, namely *Macrosiphinimar* (*Apismar1*, *Dnomar1* and *Mpmar1*). The second is close to known *Batmar-like-elements* found in the bat *Rhinolophus ferrumequinum* genome. This group includes complete (*Apismar2* and *Dnomar2*) and shorter sequences (deleted or truncated) from the three aphids species. The last tribe, namely *Dnomar-like* element, contains a complete copy from *D. noxia* (*Dnomar3*) and deleted/truncated sequences from *D. noxia* and *M. persicae*.

Furthermore, two other groups can be identified: *rosa* DD41D and a new one close to the latter (Figure 1, Additional file 3). *rosa* DD41D is represented by 44 copies restricted to *A. pisum* (*Apismar4*) and *D. noxia* genomes. They are clustered with *Crmar2* found in the Diptera Mediterranean fruit fly *Ceratitis rosa.* The second group, characterized by a long TIR, named *LTIR-like* elements, mainly comprises sequences from the pea aphid (*Apismar5.1*, *Apismar5.2*) and may correspond to a new subfamily.

In the same genome, at least four lineages can coexist. However, large differences are observed among species (Figure 1). Indeed, in *M. persicae*, a potential autonomous element (*Mpmar1*) from *Macrosiphinimar,* related to short sequences, is identified. No *rosa* elements

are detected and only deleted/truncated copies belonging to *LTIR-like*, *Dnomar-like* and *Batmar-like* elements are detected. In *D. noxia*, five *irritans* lineages are found. They include potential autonomous elements (*Dnomar*) and a few deleted/truncated copies of the same lineage. Two lineages are composed by short sequences belonging to *rosa* and *LTIR* clades. Furthermore, the genome of *A. pisum* is free of *Dnomar-like* elements. The other lineages are mainly represented by deleted/truncated copies and only a few complete sequences (*Apismar1-5*) can be detected. Hence, the large diversity of these elements among species may reflect the independent evolutionary history of these lineages or specific properties of the genome.

TIRs show a higher degree of identity in the *irritans* subfamily, suggesting a possible recent common ancestor, while they seem to be less conserved in *rosa* and *LTIR* elements (Additional file 4). In addition, TIRs do not present palindromic motifs, but only mirror repeats can be detected in *Apismar2.1* and *Dnomar2.1* belonging to *Batmar- like* elements (Table 1).

Otherwise, the screening of NCBI-nr and WGS databases (Eukaryotes) with the complete elements identified in aphid's genomes reveals only one sequence having a level of similarity above 90%, with cover queries up to 90%. In fact, it concerns a complete element belonging to the *irritans* subfamily found in the genome of the Coleoptera *Agrilus planipennis,* which is closely related to *Dnomar2.2* from *D. noxia* with 92% of similarity (Figure 1, Additional file 3).

Protein and phylogenetic analyses

The protein sequences of the 15 full clusters are characterized by an ORF encoding about 339 to 370 aa (Figure 2, Table 1 and Additional file 2). They are aligned with 56 other copies of the *Tc1-mariner* superfamily belonging to non-aphids species. The topology of the ML phylogenetic tree is roughly similar to the classification based on nucleotide sequences (Figure 1, Additional file 3). Indeed, the five tribes, previously described, are supported by high bootstrap values (98-100%). The percentage of identity between these clades varies from 28 to 59% (Additional file 5).

Only six complete sequences (*Apismar1.1*, *Mpmar1*, *Apismar2.1*, *Apismar4.1*, *Apismar4.3* and *Apismar5.1*) present an intact ORF with no frameshift or codon stop, suggesting that they might be active (Table 1, Additional file 2). The sequences related to the conserved motifs, especially WVPHEL and YSPDLA, as well as the catalytic site DD34D considered as the *mariner* signature [47,48], are detected in most of the sequences belonging to the *irritans* subfamily: *Macrosiphinimar*, *Batmar-like* elements and *Dnomar-like* elements (Figure 2, Additional file 2). The less conserved motif is WVPHEL, localized between the HTH motif and the first D. The catalytic site is relatively well conserved (7 out of 10) with a length

polymorphism between the three residues. Two sequences are deprived of HTH and one of NLS. These three copies are probably inactive.

In the *rosa* clade, close to *Crmar2-like* elements, the catalytic domain is DD41D rather than the canonical DD34D (Figure 2, Additional file 2). While the NLS motif is lacking, the HTH is located from position 88 to 110 in *Apismar4.1* and from 90 to 112 in *Apismar4.3*.

The classification and phylogenetic tree showed the presence of a monophyletic clade related to *rosa* DD41D (43% ± 0.016 of similarity), designated *LTIR*. This monophyletic group, characterized by long sequences (> 2.3 kb) with a long TIR (> 460 bp), can be divided into two tribes based on the transposase similarities. The NLS motif is absent and in the catalytic domain the distance between the second and the third D is of 40 aa for *Apismar5.2* and 41 aa for *Apismar5.1*. Otherwise, HTH motif is only present in *LTIR* DD41D (Figure 2, Additional file 2). The phylogenetic tree also indicated that *rosa* DD41D and *LTIR* DD40-41D elements are closer to *maT* and *Tc1* than to *mariner* subfamilies (Figure 1). The comparison of the sequences surrounding the catalytic site is summarized in Figure 3. The flanking sequences of the second D is clearly distinct between the different groups (*rosa/ LTIR/maT vs* the *mariner* subfamilies).

MITEs occurrence: structure and evolution

MITEs are defined as short non-autonomous copies which are known to derive from autonomous ones. They do not encode functional transposase but can be *trans*-mobilized thank to the transposase of complete copies.

MITEs, detected in the present work, represent 43 copies *i.e.* 23.5% of all extracted sequences. Only the *Dnomar-like* tribe is free of MITEs (Table 2). For the others, there is a large-size polymorphism, and MITEs are clustered into eleven sublineages based on the breaking points of the main internal deletion and the TIR sequences. All of these sequences, except one (*MITE1.1 sub2*), can be related to a full-length copy (Figure 1, 4 and Additional file 3). Microhomologies have been found at the breaking points of the internal deletions for most of the MITEs. According to the nomenclature proposed by Negoua *et al.* [49], they are of the BPEE type for seven sublineages of MITE, and of the BPNN type for two other sublineages (Table 2). For the remaining (*MITE1.1*) no microhomology can be detected.

In the *irritans* clade, represented by the *Macrosiphinimar* tribes and *Batmar-like* elements, only *A. pisum* and *M. persicae* contain MITEs, with size varying between 908 and 1165 bp. The first tribe (*MITE1.1*) includes nine copies from the pea aphid clustered in two sublineages (*sub1* and *sub2*) which only share the first 12 nucleotides of the TIRs. An additional lineage (*MITE1.2*), closely related to *MITE1.1sub1,* is found in *M. persicae*. These two sublineages present similar TIRs and an average identity of 81.8%. However, they do not have similar

breaking points (Figure 4). These two types of MITEs are related to putative autonomous copies found in each species (*Apismar1.1* and *Mpmar1.1* respectively) showing 99% of identity.

A similar situation is observed for the *rosa* clade when *MITE4.1sub1* and *MITE4.2* are compared. The *MITE4.1* lineage, includes twelve copies with lengths from 349 to 548 bp, comprised two sublineages. Although clearly related, these sublineages seemed to result from independent internal deletions of the *Apismar4.1* complete element. The *D. noxia* genome contains two copies of a MITE of 578 bp (*MITE4.2*) which are also closely related to the autonomous element *Apismar4.1* (Figure 4).

For the *LTIR* DD41D tribe, *MITE5.1,* only found in *D. noxia,* comprises five copies (790-822 bp) with the same breakpoints, and are related to the autonomous element *Apismar5.1*. No *MITE5.1* was retrieved in the *A. pisum* genome. Furthermore, *MITE5.2* of *LTIR* DD40D tribe identified in the pea aphid is composed of seven short copies (411 and 441 bp). They are divided into two sublineages depending on the breakpoint positions, probably resulting from independent internal deletions (Figure 4).

Globally, these results show that (i) MITEs in aphid species are less frequent than in *Drosophila ananassae* (about 240 copies) [41] and in *Rhodnius prolixus* (about 400 copies) [20]; (ii) *irritans* clades do not generate MITEs smaller than 900bp, in contrast to *rosa* and *LTIR-like* elements clades; (iii) three MITE sublineages (*MITE2.2*, *MITE4.2* and *MITE5.1*) are closely related to autonomous copies found in other species; (iv) orphan MITE sublineages can be detected with no full-length partner (*MITE1.1 sub2*). In the later case, it cannot be excluded that active copies still exist in other populations or closely related species.

The distribution of MITEs and their relationship with full-length elements show that their phylogeny is inconsistent with that of the species. Several scenarios involving the existence of ancestral polymorphism, current population polymorphism (presence/absence of autonomous copies and/or MITEs), stochastic loss of autonomous copies and/or horizontal transfers can be proposed.

To infer the dynamics of MITEs identified in the aphid genomes, we generated consensus sequences for each sublineage in order to estimate their period of amplification from their percentage of divergence, as proposed by Le Rouzic *et al.* [50] and Wallau *et al.* [41]. Except for two sequences of the *MITE4.1 sub2* showing 69 and 72% of identity with the consensus of this lineage, all others exhibit a level of identity higher than 85% (Figure 5). While the transposition rate (*trans*-mobilization) of these copies is unknown, we observed that some of them are almost identical (97-99% of identity) suggesting that these copies are still *trans*-mobilizable or were recently inactivated. The remaining sequences (identity level from 85% to

95%) are less conserved and probably correspond to ancient *trans*-mobilization, and are no longer mobilizable.

Discussion

The three species of aphids, *A. pisum*, *D. noxia* and *M. persicae,* present different genome sizes (541Mb, 393Mb and 398Mb respectively), which correspond to different TE equipment [35,37], *i.e.* 38% and 11.5% for the first two species (no information being available for *M. persicae*), suggesting as previously proposed that the contribution of TEs to genome size variation is greater relative to other sources of variation [41,51,52].

This present work focused on a survey of *MLE*-related elements in aphid genomes. A total of 115, 45 and 23 sequences, extracted from *A. pisum*, *D. noxia* and *M. persicae*, respectively, are clustered into 22 lineages. The relative abundance of *MLE*-related elements in these three aphids' genomes is low compared to other insect genomes. For instance, *mariner* subfamilies are represented by 10836 copies in the 700Mb genome of the Hemiptera *Rhodnius prolixus* [20] and 642 copies in the 156Mb genome of the *Drosophila eugracilis* [41]. Otherwise, the *Tc1-mariner* superfamily is poorly represented in each aphid genome compared to other superfamilies of DNA transposons, such as *piggyBac* or *hAT* (personal data). This observation might be an illustration of the competition that may occur between superfamilies as described by Abrusán and Krambeck [53]. However, today without a complete and detailed overview of TE equipment of these genomes, we do not have strong arguments to conclude that such a result is due to competition.

In the *mariner* family, only members of the *irritans* subfamily are identified in the aphid's genomes. They belong to the *Macrosiphinimar*, *Batmar-like* and *Dnomar-like* tribes, and are characterized by the DD34D catalytic site. Moreover, only three lineages might still be active (*Apismar1.1*, *Mpmar1.1* and *Apismar2.1*). No sequence related to other *mariner* subfamilies (*i.e. mauritiana, mellifera, cecropia, elegans*) is found in these genomes, although they have been identified *in vitro* in other species belonging to a closely related aphid species such as *Aphis glycines* [33] and seven tree aphids [34].

However, sequences belonging to the *rosa* family (initially closely related to the *mariner* family [9]) have been detected in *A. pisum* and *D. noxia*; and a novel clade (*LTIR-like*) has been identified. This clade is closely related to the *rosa* subfamily but is characterized by long TIRs (about 460bp *vs* 28-32bp). Moreover, conservation of some specific amino acid residues in their catalytic region, especially the final aspartic acid (D) rather than glutamic acid (E), and phylogenetic analysis revealed that *rosa* and *LTIR-like* elements are more closely related to *maT*

elements than to *Tc1* and *mariner* ones. Therefore, we suggest that *rosa* DD41D and *LTIR-like* elements constitute a large new family belonging to *Tc1/mariner*.

Distribution, diversity and phylogeny of these elements in the three aphids' genomes are probably the result of vertical transmissions associated to an ancestral polymorphism. In such a situation, closely related sequences derived from the same ancestral copy can be found in several species, while copies derived from different ancestral copies and found in the same genome, can be more distantly related (see for instance [54-56]). Host genomes are also able to repress TE activity [57,58], leading to their elimination by stochastic loss or vertical extinction. Therefore, the absence of members of the *rosa* family may be due to a stochastic loss during the evolutionary trajectory of *M. persicae*. A similar observation was illustrated in some *Drosophila* species for *mariner* subfamilies [41,59].

The high level of similarity between MITEs and autonomous partner indicates that short sequences are internally deleted elements, deriving from complete copies. Most of them exhibit direct repeat microhomologies exactly (BPE) or nearly (BPN) to the deletion breakpoints, suggesting that these internal deletions are probably due to abortive gap repair [49,60,61]. However, MITEs and related complete copies can be found in two different species, as described in the *R. prolixus* and *Drosophila* genus [20,41]. This is the case for *MITE2.2, MITE4.2 and MITE5.1*. To explain such observations, two scenarios can be proposed. First, the ancestral autonomous element at the origin of MITEs may have been lost after the MITE amplification, but was maintained in another species. Another hypothesis consists in the emergence of MITEs after internal deletion(s) of a complete copy, these MITEs being then mobilized by the transposase of another copy closely related to the first one.

Finally, horizontal transfer may also occur for all these sequences between distantly related species. For instance, the *mariner* autonomous transposon *Dnomar2.2* from *D. noxia* is closely related to the sequence of *Agrilus planipennis*. Despite a divergence time of about 361 Mya between these two species (http://www.timetree.org/index.php), the phylogenetic tree of these elements is inconsistent with that of the species. Moreover, HT could also explain the patchy distribution of MITE elements in aphids. However, in all these cases, the transfer mechanism(s) remain unknown and only propositions are suggested, like those proposed in Silva *et al.* [62] and Loreto *et al.* [63].

Conclusion

Our results represent the first *in silico* evidence of diversity and possible evolutionary scenarios of elements belonging to the three clades: *irritans*, *rosa* and a new one named *LTIR-like* elements in aphid genomes. This latter clade is characterized by long TIRs and subdivided into two distinct subgroups based on the catalytic domain signature DD40D or DD41D. Moreover,

based on protein and phylogenetic analyses, the *rosa* and *LTIR* transposons are related to *maT* DD37D elements, indicating a recent common ancestor. We also demonstrated the presence of several MITE lineages deriving from internal deletion of autonomous elements. Finally, this study proposes an update of the classification of the *Tc1/mariner* superfamily. Data analyses will offer a basis for future research aiming to understand the role of transposable elements during evolution and to develop biotechnological applications for the genetic control of aphid species.

Competing interests: The authors declare that they have no competing interests.

Authors' contributions: MB, MM and PC conceived and designed research. MB performed research. MB, JF, MMK and JDR contributed analysis tools. MB, MM and PC drafted the manuscript. AHV and JF helped draft the manuscript. All authors have read and approved the final manuscript.

Acknowledgements: This work was financially supported by the Tunisian Ministry of Higher Education and Scientific Research, the University of Tunis El Manar, the Centre National de la Recherche Scientifique and the Paris-Sud University.

References

1. Biémont C, Vieira C. Genetics: junk DNA as an evolutionary force. Nature. 2006;443(7111):521-24.

2. Feschotte C, Pritham EJ. DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet. 2007;41:331-68.

3. Chénais B, Caruso A, Hiard S, Casse N. The impact of transposable elements on eukaryotic genomes: from genome size increase to genetic adaptation to stressful environments. Gene. 2012;509(1):7-15.

4. Feschotte C, Zhang X, Wessler SR. Miniature inverted-repeat transposable elements (MITEs) and their relationship with established DNA transposons. In *Mobile DNA II*. Edited by Craig NL, Craigie R, Gellert M, Lambowitz AM. Washington DC.:ASM Press; 2002. p. 1147-1158.

5. Brillet B, Bigot Y, Augé-Gouillou C. Assembly of the *Tc1* and *mariner* transposition initiation complexes depends on the origins of their transposase DNA binding domains. Genetica. 2007;130(2):105-20.

6. Plasterk RH, Izsvák Z, Ivics Z. Resident aliens: the *Tc1/mariner* superfamily of transposable elements. Trends Genet. 1999;15(8):326-32.

7. Lohe AR, Hartl DL. Autoregulation of *mariner* transposase activity by overproduction and dominant-negative complementation. Mol Biol Evol. 1996;13(4):549-55.

8. Shao H, Tu Z. Expanding the diversity of the *IS630-Tc1-mariner* superfamily: discovery of a unique DD37E transposon and reclassification of the DD37D and DD39D transposons. Genetics. 2001;159(3):1103-15.

9. Gomulski LM, Torti C, Bonizzoni M, Moralli D, Raimondi E, Capy P et al. A new basal subfamily of *mariner* elements in *Ceratitis rosa* and other tephritid flies. J Mol Evol. 2001;53(6):597-06.

10. Claudianos C, Brownlie J, Russell R, Oakeshott J, Whyard S. *maT* a clade of transposons intermediate between *mariner* and *Tc1*. Mol Biol Evol. 2002;19(12):2101-09.

11. Medhora MM, MacPeek AH, Hartl DL. Excision of the *Drosophila* transposable element *mariner*: identification and characterization of the *Mos* factor. EMBO J. 1988;7(7):2185.

12. Robertson HM, MacLeod EG. Five major subfamilies of *mariner* transposable elements in insects, including the Mediterranean fruit fly, and related arthropods. Insect Mol Biol. 1993;2(3):125-39.

13. Auge-Gouillou C, Bigot Y, Pollet N, Hamelin MH, Meunier-Rotival M, Periquet G. Human and other mammalian genomes contain transposons of the mariner family. FEBS Lett. 1995;368(3):541-46.

14. Sinzelle L, Chesneau A, Bigot Y, Mazabraud A, Pollet N. The *mariner* transposons belonging to the *irritans* subfamily were maintained in chordate genomes by vertical transmission. J Mol Evol. 2006;62(1):53-65.

15. Laha T, Loukas A, Wattanasatitarpa S, Somprakhon J, Kewgrai N, Sithithaworn P et al. The *bandit*, a new DNA transposon from a hookworm-possible horizontal genetic transfer between host and parasite. PLoS Negl Trop Dis. 2007;1(1):e35.

16. Dupeyron M, Leclercq S, Cerveau N, Bouchon D, Gilbert C. Horizontal transfer of transposons between and within crustaceans and insects. Mobile DNA. 2014;5:4.

17. Wallau GL, Capy P, Loreto E, Le Rouzic A, Hua-Van A. VHICA, a new method to discriminate between vertical and horizontal transposon transfer: application to the *mariner* family within Drosophila. Mol Biol Evol. 2016;33(4):1094-09.

18. Robertson HM, Soto-Adames FN, Walden KK, Avancini RM, Lampe D. The *mariner* transposons of animals: horizontally jumping genes. In: Kado CI, editor. Horizontal gene transfer. Academic Press; San Diego, CA: 2002. p. 173-185.

19. Rouault JD, Casse N, Chénais B, Hua-Van A, Filée J, Capy P. Automatic classification within families of transposable elements: application to the *mariner* Family. Gene. 2009;448(2):227-32.

20. Filée J, Rouault JD, Harry M, Hua-Van A. Mariner transposons are sailing in the genome of the blood-sucking bug Rhodnius prolixus. BMC genomics. 2015;16(1):1.

21. Bigot Y, Brillet B, Auge-Gouillou C. Conservation of palindromic and mirror motifs within inverted terminal repeats of *mariner*-like elements. J Mol Biol. 2005;351(1):108-16.

22. Kidwell MG, Lisch DR. Perspective: transposable elements, parasitic DNA, and genome evolution. Evolution. 2001;55(1):1-24.

23. Jacobson JW, Medhora MM, Hartl DL. Molecular structure of a somatically unstable transposable element in Drosophila. Proc Natl Acad Sci USA. 1986;83(22):8684-88.

24. Maruyama K, Hartl DL. Evidence for interspecific transfer of the transposable element *mariner* between *Drosophila* and *Zaprionus*. J Mol Biol. 1991;33(6):514-24.

25. Medhora M, Maruyama K, Hartl DL. Molecular and functional analysis of the *mariner* mutator element *mos1* in Drosophila. Genetics. 1991;128(2): 311-18.

26. Barry EG, Witherspoon DJ, Lampe DJ. A bacterial genetic screen identifies functional coding sequences of the insect *mariner* transposable element *Famar1* amplified from the genome of the earwig, *Forficula auricularia*. Genetics. 2004;166(2):823-33.

27. Muñoz-López M, Siddique A, Bischerour J, Lorite P, Chalmers R, Palomeque T. Transposition of Mboumar-9: identification of a new naturally active mariner-family transposon. J Mol Biol. 2008;382(3):567-72.

28. Lampe DJ, Akerley BJ, Rubin EJ, Mekalanos JJ, Robertson HM. Hyperactive transposase mutants of the *Himar1 mariner* transposon. Proc Natl Acad Sci USA. 1999;96(20):11428-33.

29. Rholl DA, Trunck LA, Schweizer HP. In vivo *Himar1* transposon mutagenesis of *Burkholderia pseudomallei*. Appl Environ Microb. 2008;74(24):7529-35.

30. O'brochta DA, Atkinson PW. Transposable elements and gene transformation in non-drosophilid insects. Insect Biochem Mol Biol. 1996;26(8):739-53.

31. Wang W, Swevers L, Iatrou K. *mariner* (*mos1*) transposase and genomic integration of foreign gene sequences in *Bombyx mori* cells. Insect Mol Biol. 2000;9(2):145-55.

32. Delaurière L, Chénais B, Hardivillier Y, Gauvry L, Casse N. *mariner* transposons as genetic tools in vertebrate cells. Genetica. 2009;137(1):9-17.

33. Mittapalli O, Rivera-Vega L, Bhandary B, Bautista MA, Mamidala P, Michel AP et al. Cloning and characterization of *mariner*-like elements in the soybean aphid, *Aphis glycines* Matsumura. Bull Entomol Res. 2011;101(6):697-704.

34. Kharrat I, Mezghani M, Casse N, Denis F, Caruso A, Makni H et al. Characterization of *mariner*-like transposons of the mauritiana Subfamily in seven tree aphid species. Genetica. 2015;143(1): 63-72.

35. Richards S, Gibbs RA, Gerardo NM, Moran N, Nakabachi A, Stern D et al. Genome sequence of the pea aphid *Acyrthosiphon pisum*. PLoS Biol. 2010;8(2): e1000313.

36. Jurka J. *mariner* families from *Acyrthosiphon pisum*. Repbase Reports. 2008;8(3):340.

37. Nicholson SJ, Nickerson ML, Dean M, Song Y, Hoyt PR, Rhee H et al. The genome of *Diuraphis noxia*, a global aphid pest of small grains. BMC genomics. 2015;16(1):1.

38. Kim H, Lee S, Jang Y. Macroevolutionary patterns in the Aphidini aphids (Hemiptera: Aphididae): diversification, host association, and biogeographic origins. Plos One. 2011;6(9):e24749.

39. Larsson A. AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics. 2014;30(22):3276-78.

40. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26(19):2460-61.

41. Wallau GL, Capy P, Loreto E, Hua-Van A. Genomic landscape and evolutionary dynamics of mariner transposable elements within the Drosophila genus. BMC genomics. 2014;15(1):1.

42. Bannai H, Tamada Y, Maruyama O, Nakai K, Miyano S. Extensive feature detection of N-terminal protein sorting signals. Bioinformatics. 2002;18(2), 298-05.

43. Gao Y, Mathee K, Narasimhan G, Wang X. Motif detection in protein sequences. String Processing and Information Retrieval Symposium, 1999 and International Workshop on Groupware. doi: 10.1109/SPIRE.1999.796579.

44. Narasimhan G, Bu C, Gao Y, Wang X, Xu N, Mathee K. Mining protein sequences for motifs. J Comput Biol. 2002;9(5):707-20.

45. Abascal F, Zardoya R, Posada D. ProtTest: selection of best-fit models of protein evolution. Bioinformatics. 2005;21(9):2104-05.

46. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol. 2013;30(12): 2725-29.

47. Robertson HM. The *mariner* transposable element is widespread in insects. Nature. 1993;362(6417):241-45.

48. Hartl DL, Lozovskaya ER, Nurminsky DI, Lohe AR. What restricts the activity of *mariner*-like transposable elements?. Trends Genet. 1997;13(5):197-01.

49. Negoua A, Rouault JD, Chakir M, Capy P. Internal deletions of transposable elements: the case of Lemi elements. Genetica. 2013;141(7-9):369-79.

50. Le Rouzic A, Boutin TS, Capy P. Long-term evolution of transposable elements. Proc Natl Acad Sci USA. 2007;104(49):19375-80.

51. Kidwell MG. Transposable elements and the evolution of genome size in eukaryotes. Genetica. 2002;115:49-63.

52. Vieira C, Nardon C, Arpin C, Lepetit D, Biémont C. Evolution of genome size in Drosophila. Is the invader's genome being invaded by transposable elements?. Mol Biol Evol. 2002;19(7):1154-61.

53. Abrusán, G, Krambeck HJ. Competition may determine the diversity of transposable elements. Theor Popul Biol. 2006;70(3):364-75.

54. Capy P, Anxolabéhère D, Langin T. The strange phylogenies of transposable elements: are horizontal transfers the only explanation?. Trends Genet. 1994;10(1):7-12.

55. Green CL, Frommer M. The genome of the Queensland fruit fly *Bactrocera tryoni* contains multiple representatives of the *mariner* family of transposable elements. Insect Mol Biol. 2001;10(4):371-86.

56. Prasad MD, Nurminsky DL, Nagaraju J. Characterization and molecular phylogenetic analysis of *mariner* elements from wild and domesticated species of silkmoths. Mol Phylogenet Evol. 2002;25(1):210-17.

57. Rozhkov NV, Aravin AA, Zelentsova ES, Schostak NG, Sachidanandam R, McCombie WR et al. Small RNA-based silencing strategies for transposons in the process of invading Drosophila species. RNA. 2010;16(8):1634-45.

58. Tóth KF, Pezic D, Stuwe E, Webster A. The piRNA pathway guards the germline genome against transposable elements. In: Non-coding RNA and the Reproductive System. Springer Netherlands: 2016. p. 51-77.

59. Capy P, David JR, Hartl DL. Evolution of the transposable element mariner in the Drosophila melanogaster species group. Genetica. 1992;86:37-46.

60. Rubin E, Levy AA. Abortive gap repair: underlying mechanism for *Ds* element formation. Mol Cell Biol. 1997;17(11):6294-02.

61. Brunet F, Giraud T, Godin F, Capy P. Do deletions of Mos1-like elements occur randomly in the Drosophilidae family?. J Mol Evol. 2002;54(2): 227-34.

62. Silva JC, Loreto EL, Clark JB. Factors that affect the horizontal transfer of transposable elements. Curr Issues Mol Biol. 2004;6(1):57-71.

63. Loreto ELS, Carareto CMA, Capy P. Revisiting horizontal transfer of transposable elements in Drosophila. Heredity. 2008;100(6):545-54.

64. Nicholas KB, Nicholas HBJ, Deerfield DW. GeneDoc: Analysis and Visualization of Genetic Variation. EMBNEW.NEWS. 1997;4(1):14.

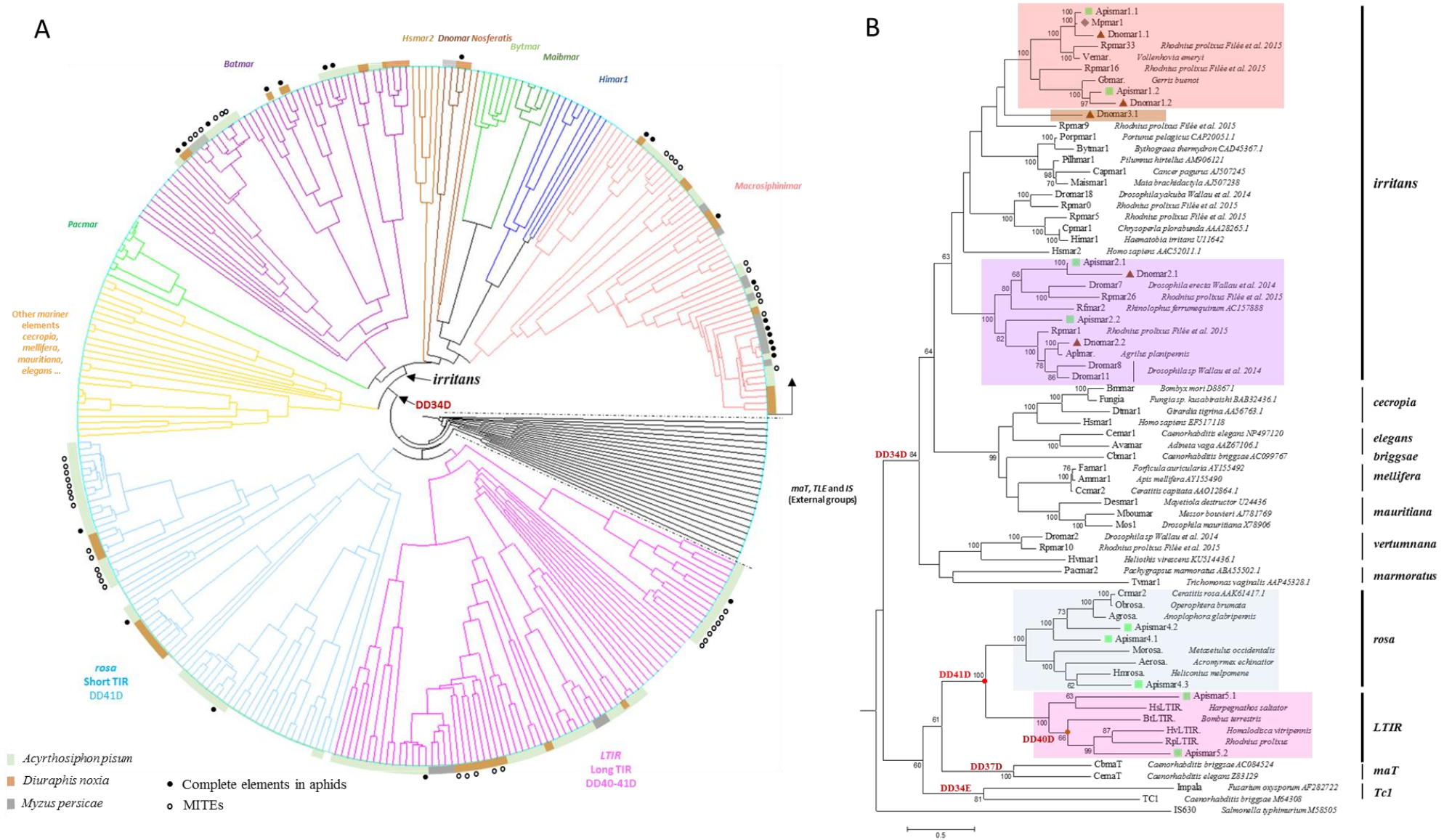


Figure 1. Classification and phylogenetic tree of sequences identified in *Macrosiphini* tribe.

A. Classification of the 115 sequences obtained from *Acyrthosiphon pisum*, 45 from *Diuraphis noxia* and 23 from *Myzus persicae*. These 183 sequences, along with 146 elements belonging to the *Tc1/mariner* superfamily were classified using the UPGM-VM method [19]. References and positions of all these sequences are given in Additional file 3 according to the reading sense indicated by the arrow in the circular tree. Sequences found in *A. pisum*, *D. noxia* and *M. persicae* are given in colour, in green, brown and grey, respectively. Complete sequences are marked by a full black circle and MITEs by an empty circle. **B.** The phylogeny based on amino-acid sequences of the 15 lineages. After a search of the best evolutionary scenario (ProTest 2.4), this tree was generated in MEGA6 with the Maximum likelihood (ML) method, using the WAG+F+I+G model. Only bootstrapping values (1000 replications) higher than 60% are written on the branch. Families and subfamilies are indicated in the right-hand part of the tree. The colored rectangles correspond to the different tribes as in A. Green squares, grey lozenges and brown triangles refer to the aphid species *A. pisum*, *D. noxia* and *M. persicae*, respectively. The designation of unpublished sequences extracted from other species than those of the three aphids includes a point (*i.e. Vemar.* from *Vollenhovia emeryi*). Sequences name: *Apismar*: elements from *Acyrthosiphon pisum*, *Dnomar:* elements from *Diuraphis noxia* and *Mpmar*: elements from *Myzus persicae.*

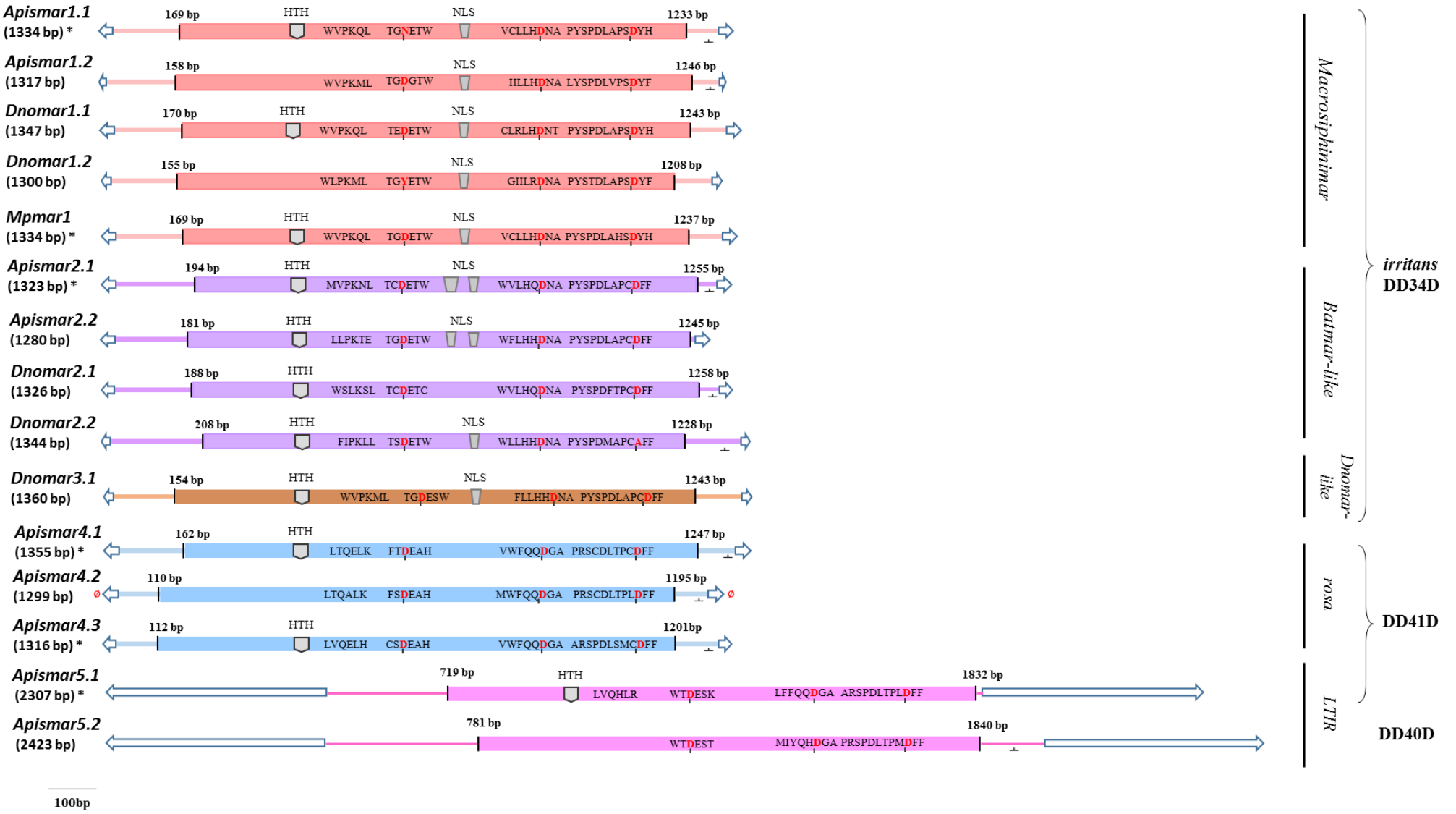


Figure 2. Schematic representation of the 15 lineages corresponding to complete sequences found in aphid's genomes.
The elements are arranged and colored (as in Figure 1) according to the clades they belong to. Potentially active copies are marked with asterisks. The lack of TA (TSD) is marked by a slashed zero in red. Blue arrows indicate TIR, while bold lines represent UTRs. A turned T shows the presence of polyAdenylation site "AATAAA". In transposase gene, the three catalytic residues containing aspartic amino acids marked in red are indicated. The helix turn helix (HTH) region, the nuclear localization signal (NLS), and motifs related to WVPHEL are also mentioned.
Sequences name: *Apismar*: elements from *Acyrthosiphon pisum*, *Dnomar:* elements from *Diuraphis noxia* and *Mpmar*: elements from *Myzus persicae.*

			D	D	D/E
rosa **DD41D**	Crmar2	:	FSDEA	QQDGA	SCDLTPLDF
	Apismar4.1	:	FTDEA	QQDGA	SCDLTPCDF
	Apismar4.2	:	FSDEA	QQDGA	SCDLTPLDF
	Aerosa.	:	FSDEA	QQDGA	SPDLTPCDF
	Apismar4.3	:	CSDEA	QQDGA	SPDLSMCDF
LTIR **DD40-41D**	BtLTIR.	:	WTDER	QQDGA	SPDLTPLDF
	HasLTIR.	:	FSDEA	QQDGA	SPDLTSLDF
	Apismar5.1	:	WTDES	QQDGA	SPDITPLDF
	Apismar5.2	:	WTDES	QHDGA	SPDLTPMDF
maT **DD37D**	CemaT	:	FTDEK	QQDGA	SPDLNPMDY
	CbmaT	:	WTDEK	QQDWA	SPDLNPMDF
Tc1 **DD34E**	Impala	:	WSDEC	MHDNA	SPDLNPIE-
	CbTc1	:	FSDES	QQDND	SPDLNPIE-
mariner **subfamilies DD34D**	Mos1	:	TGDEK	LHDNA	SPDLAPSDY
	Hsmar1	:	TCDEK	LHDNA	SPDLSPTDY
	Cemar1	:	TGDEK	LHDNA	SPDLAPTDY
	Famar1	:	TGDEK	HHDNA	SPDLAPSDY
	Cpmar1	:	TMDET	HQDNA	SPDLAPSDF
	Dnomar3.1	:	TGDES	HHDNA	SPDLAPCDF
	Rfmar2	:	TGDET	HHDNA	SPDLAPNDF
	Apismar2.1	:	TCDET	HQDNA	SPDLAPCDF
	Dnomar2.1	:	TCDET	HQDNA	SPDFTPCDF
	Dnomar2.2	:	TSDET	HHDNA	SPDMAPCAF
	Apismar2.2	:	TGDET	HHDNA	SPDLAPCDF
	Apismar1.1	:	TGNET	LHDNA	SPDLAPSDY
	Dnomar1.1	:	TEDET	LHDNT	SPDLAPSDY
	Mpmar1	:	TGDET	LHDNA	SPDLAHSDY
	Apismar1.2	:	TGDGT	LHDNA	SPDLVPSDY
	Dnomar1.2	:	TGYET	LRDNA	STDLAPSDY

Figure 3. Multiple alignments of catalytic motifs of *Tc1*, *mariner*, *maT* families with the 15 lineages identified in aphids.

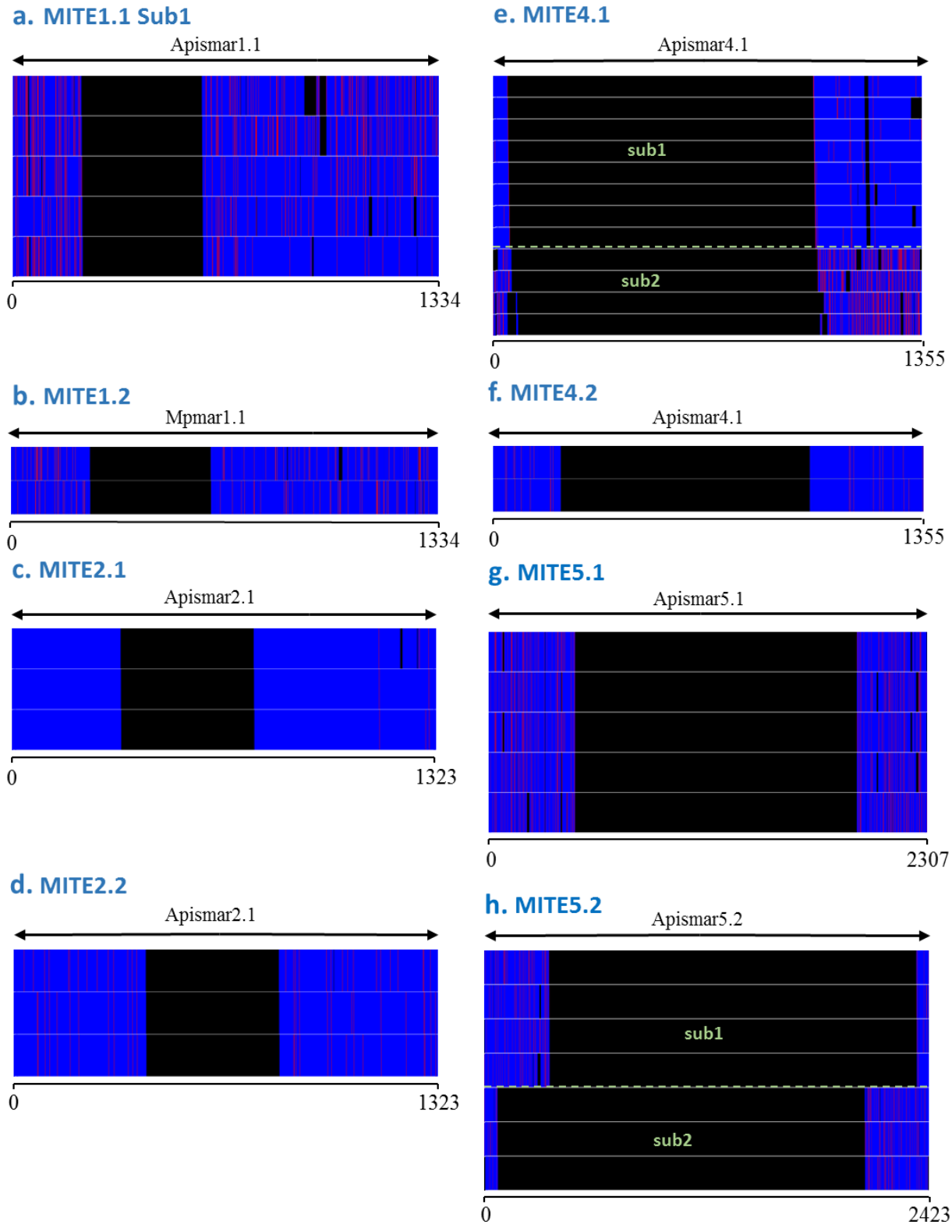


Figure 4. Sequence alignments of MITE lineages with a longer autonomous partner.
For each alignment (a-h), sequences are in blue, showing substitutions in red and gaps in black. The autonomous copies related to MITE and the global structure of the copies are shown on top, with arrowheads corresponding to TIR. Similar copies in length and sequence-defined sublineages (numbered in green). Given the lack of homology with the full potential element, *MITE1.1 sub2* is not represented. a, c, e and h are found in *A. pisum*, b and d in *M. persicae*, f and g in *D. noxia*.

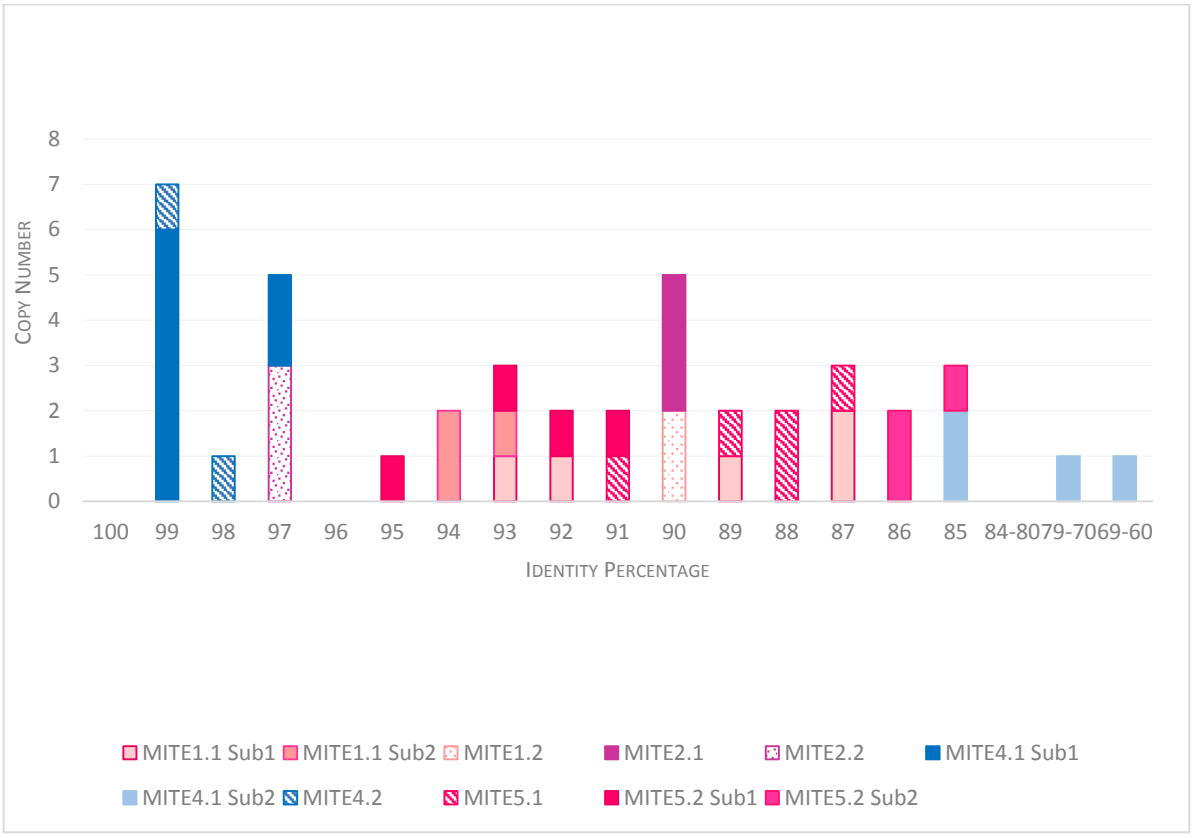


Figure 5. Evolution analysis of different MITEs sublineages.
Based on the comparison of consensus with copies, the similarity rates are identified. While copy sublineages with a high level of similarity present recent invasion, the decrease of this percentage refers to an ancient element. Filled, hatched and dotted patterns correspond to *A. pisum*, *D. noxia* and *M. persicae,* respectively. Colors match to the different tribes as in Figure 1.

Table 1. Characteristics of 15 lineages corresponding to complete elements.

The copy number, clade, length of the element, TIR and ORF, as well as the presence of potentially active copies, are specifically indicated for each complete sequence. The number of copies not truncated by "N" is mentioned in the fifth column. Potentially active copy = existence of at least copy with a complete ORF, with no frameshift or codon stop. In TIR sequences, the mirror sites are mentioned in bold.

Clade	Tribe	Species	Lineage name	Complete copy number	Length (bp)	TIR		ORF Length (aa)	Potentially active copy
						Length (bp)	Sequences		
irritans DD34D	Macrosiphinimar	*A. pisum*	Apismar1.1	1	1334	28	CGAGGCGTGTCCAGAAAGTAAGTGTACT	354	Yes
			Apismar1.2	1	1317	15	TTCGAAAAGTAAGGG	355	No
		D. noxia	Dnomar1.1	1	1347	28	CGAGGCGTGTCCAGAAAGTAAGTGTACT	354	No
			Dnomar1.2	1	1300	20	TWCGAAAAKTAAGGGCCGTT	347	No
		M. persicae	Mpmar1.1	6	1334	28	CGAGGCGTGTCCAGAAAGTAAGTGTACT	355	Yes
	Batmar-like	*A. pisum*	Apismar2.1	1	1323	30	CGAGGTATGA**CAATAAAATAAY**GAGACTTT	354	Yes
			Apismar2.2	2	1280	22	AAYACCCAGACAAMAWKTATTA	354	No
		D. noxia	Dnomar2.1	2	1326	27	YGAKGTGWS**AMATAAAATAAA**CGAGAC	357	No
			Dnomar2.2	2	1344	24	CSWGGTGTGTTCAAAAAGWACYCG	339	No
	Dnomar-like	*D. noxia*	Dnomar3.1	1	1360	26	CGAGGGCGGGCTGATAAGTAATGCCT	362	No
rosa DD41D	Crmar2-like	*A. pisum*	Apismar4.1	1	1355	32	AAGGGTGTCTCAAAAAGAACGCCGGATTTRAA	361	Yes
			Apismar4.2	1	1299	32	GGGTTTTTCAATARRAGCGCTCGAWSTTTSAT	361	No
			Apismar4.3	1	1316	27	GGTGCGGCAGAGCCRACTGACGAGTTT	362	Yes
LTIR	DD41D	*A. pisum*	Apismar5.1	1	2307	466	TCACCAATTTAGGGAACACTGAATTTCTCGGCT…	370	Yes
	DD40D	*A. pisum*	Apismar5.2	1	2423	460	AATGTGTCAAACTTCTAGAGGTGTTTCTACACC…	351	No

Table 2. List of MITEs detected in the aphid's genomes.

NR = no related autonomous copy identified. Presence of short direct repeat (microhomologies) in the region of deletion breakpoints are indicated: BPEE for Breaking Point Exact Exact and BPNN for Breaking Point Near Near (according to the nomenclature proposed by Negoua *et al.* [49]).

Clade	Tribe	Species	ID MITE	Length (bp)	Sublineage	Copy number	TIR sequences	Autonomous element related to MITE		Breakpoints
								sequences	Average identity (%)	
irritans DD34D	Macrosiphinimar	*A. pisum*	MITE1.1	923-1165	sub1	5	CGAGGCRTGTCCAGAAAGTAAGTGTACT	Apismar1.1	90.8	-
					sub2	4	CGAGGCGTGTCCCAAAARTAAGGTCTCCAT	NR	-	
		M. persicae	MITE1.2	959, 1007	sub1	2	CGAGGCGTGTCCWGAAAGWAAGTGTACT	Mpmar1.1	92	BPEE
	Batmar-like	*A. pisum*	MITE2.1	908-931	sub1	3	CGAGGTRTGACAATAAAATAACGAGACTTT	Apismar2.1	98.6	BPEE
		M. persicae	MITE2.2	908-912	sub1	3			97	BPNN
rosa DD41D	Crmar2-like	*A. pisum*	MITE4.1	349-548	sub1	8	AAGGGTGTCTCAAAAAGAACGCCGGATTTRAA	Apismar4.1	94.5	BPEE
					sub2	4	RGGRTRYCWCAAAAARAAGSGYGGATTTKRAA		74.6	
		D. noxia	MITE4.2	578	sub1	2	WAGGGTGTCTCAAAAAGAACGCCGGATTTRAA		97	
LTIR	DD41D	*D. noxia*	MITE5.1	790-822	sub1	5	TCACCAATTTAGGGATCACTGAATTTCTCGGC…	Apismar5.1	86.2	BPEE
	DD40D	*A. pisum*	MITE5.2	411-441	sub1	4	AATGTGTCAAACTTCTAGAGGTGTTTCTACAC…	Apismar5.2	90.25	BPEE
					sub2	3			90	BPNN

Additional file 1. *mariner and rosa* transposases sequences used as queries in the tBLASTN search (Species, Clades, Accession number).

Transposase	Species	Clades		Accession number
		Catalytic motif	ID name	
Cbmar1	*Caenorhabditis briggsae*	DD34D	*briggsae*	AC099767
Avamar	*Adineta vaga*		*elegans*	AAZ67106.1
Cemar1	*C. elegans*			NP497120.1
Gtmar1	*Girardia tigrina*		*cecropia*	CAA56763.1
Hsmar1	*Homo sapiens*			U52077
Bytmar	*Bythograea thermydron*		*irritans*	CAD45367.1
Cpmar1	*Chrysoperla plorabunda*			AAA28265.1
Fungia	*Fungia sp. Kusabiraishi*			BAB32436.1
Himar1	*Haematobia irritans*			U11642
Hsmar2	*H. sapiens*			AAC52011.1
Rfmar	*Rhinolophus ferrumequinum*			AC157888
Ammar	*Apis mellifera*		*mellifera*	AY155490
Ccmar2	*Ceratitis capitata*			AAO12864
Famar1	*Forficula auricularia*			AY155492
Desmar1	*Mayetiola destructor*		*mauritiana*	U24436
Mboumar	*Messor bouvieri*			AJ781769
Mos1	*Drosophila mauritiana*			X78906
Crmar2	*Ceratitis rosa*	DD41D	*rosa*	AAK61417.1

Additional file 2. Amino acid sequences of the transposase of the 15 complete elements. The three aspartic residues of the catalytic domain are marked in red, the sequences related to the WVPHEL-specific motif of the DNA binding domain are indicated in black bold as well as the helix turn helix (HTH) region (underlined) and the NLS (in blue). Stop codons are represented by asterisks (*).

Apismar1.1

MSIESAAKCEIRAVIRYLVAKEKSPHEIFNEVRTVYGEGHMNRTSVYKWCREFKNDRTNVHDDLRSGRPSILTDD
IVKKSRMRFVDRRLTLDELSAMFPQLSRSLLHETSITETSLGFHKLCAR**WVPKQL**TEQHMLNRVQASREFLERYE
LDGDNFLKSIVTG**N**ETSWVAHYTPETSKRQSEQWRHTTS**PSTKKFN**TTISAKKIMASVFWDHKGIILIEYLPQGE
TSINAARYCETSLKKLRRAIQNKRRGLLTSGVCLLH**D**NARPHTANVTKQLLDSFGWDVLNHTPYSPDLAPS**D**YHL
FTSLKKHMGGKKFSADEEVKGAVDKWTKEMAAEFYEAGIKKLICRLTTCIERNGDYVEK

Apismar1.2

MSAIVAAPASCEVRTVIRFLCAKRSSAAEIHGELCLAYGLTVMSEGKIRWCRDFKNGRTNVHDEERSGRPSMQTD
EIVSLVDQKLRFDCRLTISALADEFPNLARTTVYTIITEKLGYHKLYAR**WVPKML**TDQHKEQRIMSSGREFLNRY
RQDGDNLFSHIVTG**D**GTWISYINPETSKQQSMQWRHSTS**PKQKKFK**QTPYTSRKMMATVFWDEKSVLLVDFMERG
TTITAQVYCETSLNKLRCAIQNRRRGKLSSSIILLH**D**NARPHTAAKTQEKIDFR*ELFNHPLYSPDLVPS**D**YFLF
FHFKKWLVGQRFENDKELNAVENWFNSQAANFYADGLRKLVKRYEKCFEINGNYVEK

Dnomar1.1

MSIESATMCEIRAVIRYLDAKEKSP*GIFNEVRTTYIEGNNINRTSVYKWCREFKNNRTNVHDDLRSGRPSVLTD
DIVKKVENAVCDDRRLTLDELSAIFPQLPRSLIHETSITETSLGFHKLCAR**WVPKQL**TDQHMLNRVQASREFLER
YELDGDNFLKSIITE**D**ETSWVAHYTLETSKRQSEQWHHTTS**PSTKKFK**TIISAKKIMALVFWDHKGIIFIKYLPQ
GETSINAARYFETSLKKLRRAIQNKKRGLLTSGDCLRLH**D**NTKPHMANVSKQLLDSFGWDVLNHPPYSPDLAPS**D**
YHFFTSLKKLVVGKKFSTDEEVKGAVDKWTKEVFYGAGIKKFICCLITCI*RDGDYAEK

Dnomar1.2

MSTIIPAPTSCEIRAVIRFLCAKRSSAAEIHQELFLVYGPNVMSEGKIRQWCRDLKNGRTNVHSMQTDEIVSLMD
QKPRTNRRLTFSTLADEFPNL*RTIVYTIVTEKLGYHKLCAR**WLPKML**TDQHQE*RITSG*EFLIHYRQDGDNLF
SHIATG**Y**ETsWISYINPETsKQQSMQWCRSTSSK**PKKFK**PTPCTSRKMIATVFWDEKGVLLVDFMERGTTITAEV
YCETsLNKLRRAIQN*RSGK*SGIILR**D**NARPHTAAKTQEKIQDFRWKLLNHPPYSTDLAPS**D**YFLFLHFKKWLG
EQRFENDKELKNAVENWFKSQNEFYTDELRKLVKWYEKCLEVNGDYVEK

Mpmar1.1

MSIESAAKCEIRAVIRYLVAKEKSPHEIFNEVRTVYGEGHMNRTSVYKWCREFKIGCTNVHDDLRSGRPSILTDD
IVKIVENAVRDDRRLTLDELSAMFPQLSRSLLHETSITETSLGFHKLCAR**WVPKQL**TEQHMLNRVQASREFLERY
ELDGDNFLKSIVTG**D**ETSWVAHYTPETSKKQSEQWRHTTS**PSTKKFK**TTISAKKIMASVFWDLKGIILIEYLPQG
ETSINAARYCETSLKKLRRAIQNKRRGLLTSGVCLLH**D**NARPHTANVTKQLLDSFGWDVLNHPPYSPDLAHS**D**YH
LFTSLKKHMGGKKFSADEEVKGAVDKWTKEMAAEFYEAGIKKLICRLTTCIERNGDYVEK

Apismar2.1

MLDIKIEQRVNIKFLVKLKKTAAESFRMLCEVYGEECLSRARVFEWHKRFCSGREDVEDDDRSERPTTSSTNENV
EKIDKIIRQDRRLSVRAVAEMVNIDRESVRKILVENLNMNKVCAK**MVPKNL**TIDQKFNRKEICSDTLKIIKDDPS
FINNIITC**D**ETSWIFTYDPETS**KRQSMHWKTPTSPRMKK**ARMSKSKFKAMLIVFFDIKGIIFVEWVPSGQTVNQY
YYKEVLIKLRERV**RKKRP**DLWKNGWVLHQ**D**NAPAHSAFSIQRYLTEKKISVLQHPPYSPDLAPC**D**FFLFPKIKSL
LKGTHFQTVDDVKMKTAELLKGLNESDWQHCFQEWQRRMQQCIDAEGRYFEGDNH

Apismar2.2

MDNITEQRACFKFCSKIGKNATETSLELIKLAFGDVSLSRCVTFDWFRRSKEGRISIEDDYRPGRPSSSKTNDTI
DLVRNKIRNYRRLTVREVANEVGISIGTCHSILSDELSMKRVSAK**LLPKTE**EQMEHRIEVCLELKNRVSNDPNFI
KSIITG**D**ETSWVYGYDPKTKVQSSQWKTANSSR**PKKCRQI**RSNIKAMLIVFFDFFGLVHYEFVPTGQTINQVFYK
*VVLKRLREKVC**RKRP**EVWKSKSWFLHH**D**NAPAHSALSIREFLASKNIPVVPHPPYSPDLAPC**D**FFLFPRLKSTL
KGHRFVDVNETSIHNATQELKAITMKEIQRCFKKWQDRWDHCIEAKGHYFEGDPFK

Dnomar2.1

MLGIKFEQR*NIKFFCEIKKKIAVESFHILCEFYGEERLSRVCVFELHKERMSKMMIVLDVLRPTTPSTNENVEK
IDTIIRK*RRLSVRAVTEMVNIDRESVRKILVENLNMKKMCAK**WSLKSL**TVDQKFNRKEICSDTLKIIKDDPSFI
NNIITC**D**ETSCIFTYDTETSKRQSMHWKTVRHQVSQE*RKHE*TSQNSKQFSLFSLTLRE*LFLEWVPSGQTVYQ
YYCKEILIKLKEHIRKK*LNLWKNGWVLHQ**D**NAPAHSAFSIQRFLTEKNIFILQHPPYSPDFTPC**D**FFLFSKIKS
LLKNGTNNFQTVDDVKM*TAELLKGLTESDWQYCFQEWQRRMQQCVDDE*RYFEGDNH

Dnomar2.2

MEELKSQRIFIKFCVKNEIKCSKVCELLQKAYGESAMKKTIVNEWYKRYQDGRKDVEDDKRSGRPSTSIIDANVK
KVEKMVVNDRRITIIREVADEVGISIVSCQNIFSNVLGLKRVAAK**FIPKLL**NFDQKNNRMNVAQELLNDVNVDPS
LLERVITS**D**ETSWVYGYEVETSKAQSSTWKHSTS**PRAKKARQV**RSNVKVLLTVFFDFNGIVHQEFLPQGRTVNKE
YYLEVMRRLREAIRKKRPDIWKNNSWLLHH**D**NAPAHSSLLVHNFLAKNNTAVMPQPPYSPDMAPCAFFLFPMLKR
HMKGQRFSSIEEIKAESLRVLKDMPKSEYQECFEDWKKSLA

Dnomar3.1

MESIITAPSDVRFKQRAVIEFLVAENVKPVDIHRRLLAVYGNQTLDVSSVRRWALRVKGSEVGKAIITDQDRSGR
PVTVTDEGLVTRPIKQKVDDLVKGNRRIKQSEIAIALGISKERVQHILCELEYRKICTR**WVPKML**TEEMKQNRVE
ICRQLLLRLNVRENFLNIMVTGDESWVHHYGPENKRQSMEFRHKTSPA**PKKFK**VQASAGKVMLTVFWDSKGVIHT
EYLEKGSTINSIRYIEALKKLKKRIKRVRPNLTQFLLHHDNARPHCSRATMTAIESLGFQVIPHPPYSPDLAPCD
FFLFPKLKEHLKGTKFNSDEKVKAEVKRWFNAQPEEFYLNGISKLVNRWQKCIALEGSYVEK

Apismar4.1

MERYSKEQRVLIVKTHYQNGEHYAVTVRKLRTILGHHNAPNESTVRRLIKKFEESGSTQDKKISGRHRSGRSEAN
VTVVHDSVTVSPRKSCRRRAQEMHMSPATMQRILTKDLHLHAYKVQ**LTQELK**PADHEKRRQFVEWILTRDRESEG
FAKRIIFTDEAHFHLNGFVNKQNCRIWGSENPRTIQEKEMHPERVTVWCGIWSGGLIGPYFFEDEEGNAVTVNGV
RYRAMLNHFLWPRLDQMNIENVWFQQDGATCHTSRETSIALLREKFPDTLISLRGDQSYPPRSCDLTPCDFFLWG
YTKSRVYQNKVRNVLELKQEIRCVLNELDGAMCDRVMVNFMERIIAYRASRGGHMPDVVFHC

Apismar4.2

MNGYSVEQRVRIIKFYYQNQCSVRETSFRAFTDFYPRHNRPAESTIRRLVAKF*STGSINYQPTPIRQRNARSIE
NIAAVRDSVRENPRQ*IPRRSQELGLSVTSTWRILRRDLGLHPYKIQ**LTQALK**VNDHTQRRVFADWVLGQLAVDP
NFAKKIIFSDEAHFWMNGYVNKQNCRIWDDTNPHKTHQNKMHPEEVTVWCGFWSGGIIGPYFFQNETSGIAITVN
GERYRSMINNFFWPKLDDMDTEDMWFQQDGATCHTARATMDILRERFEGMVISRNGDINWPPRSCDLTPLDFFLW
GYLKSQVYANKPQTIDALKVNIINTIKKIQPDVCNKVIENWTTRIRATKQSRGGHLNDVIFHK

Apismar4.3

MVWTGVHRSFVVRAYYENNHSLIATQRAFRIHFGIPRNESIPSANTIKFWIRQLEETSGSTLSELGHGAPRTVRT
PENVQLVRESIEQSPTRSARKHAVALGISVRSLRRILHEYLSFHPYKLM**LVQELH**ATDYDNRKNLCQQILLRIPP
TSTFFCSDEAHFHLSGTVNKQNFRYWAANNPQQLHERPLHSPKVTVWCGVSQFGVIGPYFFEDENRTVIVTPGRY
VVMLETSYLQQRLEEMAEYHNLENVWFQQDGATAHTAQISLGVLQQMFPGRLVSLRGDIGWPARSPDLSMCDFFL
WGYLKDKVFRHRPHTIEDLKQKITEEIEAIPVETSCRKSYESFRDRLQQCIDADGRHHRDIISKQ

Apismar5.1

MLVVILVRLVGLNVKMNNQEKVQMLLIYGKCDRNSRQSARMYAEQYPGRYHPPHTFFIKIEQLLINHGAFSVKVV
RNQQIRENNINEDVELQVLAYIRLNPRSSVRHVGREVGISFGLVHKILKKHKMHPYKPD**LVQHLR**PADPERRLNF
IAWLLVQIDTKPLFLNQILWTDESKFTNNGVINKQNNRMWSDVNPHWAVDNRYQTVWGTNVWCGLIGGKLLGPYF
YEENLTARRYLAFLTNVLPLMLENLPLATRQTLYFQQDGAPAHNAHIVRDYLNRVYEGKWLGTYGPIEWPARSPD
ITPLDFFLWGHLKTVVYADPPVNLADLKNKILVACNNLTESQIMSATNRGCLQRFQLCVDNHGANFEQFI

Apismar5.2

MPSYSNTELPDMHFIYGLCNGNTRASQREYENRFPHRRVPAPAMFSRIHQALRQRGNFRRSLRESVQNVDLEREI
LDEVNRDPETSSTRTLAHQFGVHHSTVWRTINREGLHPYHF*EFMA*RTQTINNVYSSVDGYFIMKLRIVVFSKV
LWTDESTFTREGVFNIHNSHHYAQENPRLVRQQRFQRRFSINVWMGIIGGVLIGPFLGLPRTVGGNSYLNFLQNE
LPGLLEDLPLEVRRRMIYQHDGAPPHFSRAVRQHLDETSFTCWIGRGGTIPWPPRSPDLTPMDFFVWGYLKERVY
HQEVDSEAELRQRILQAAIEIRRVVTAGVTGRHVRERARACLRQNGGHIEQLL

Additional file 3. Sequences classified by UPGM-VM method according to the reading sense indicated by the arrow in the circular tree. Deleted or truncated sequences are indicated by an asterisk (*).

***Macrosiphinimar* DD34D**			
1	*Diuraphis noxia*	gi984744883:140935-140721 strand-	*
2		gi984744980:13238-14654 strand+	*
3		gi984745316:3270-2865 strand-	*
4		gi984745505:879619-879165 strand-	*
5	*Acyrthosiphon pisum*	gi320446981 NW003383590.1 Scaffold101:202225-201649 strand -	*
6		gi320388812 NW003403215.1 Scaffold19726:243-904 strand -	*
7		gi320446984 NW003383587.1 Scaffold 98:39716-40668 strand +	MITE1.1 sub1
8	*Myzus persicae*	Scaffold419:92618-93582 strand-	*
9	*Acyrthosiphon pisum*	gi320446985 NW003383586.1 Scaffold97:172262-173595 strand +	Apismar1.1
10	*Myzus persicae*	Scaffold938 :36419-37755 strand-	Mpmar1.1
11		Scaffold42 :437239-438574 strand-	
12		Scaffold26 :74179-75510 strand+	
13		Scaffold353:292415-290211 strand -	
14		Scaffold167:418817-420155 strand+	
15		Scaffold425:148657-149615 strand +	MITE1.3
16	*Diuraphis noxia*	gi984745178:34523-35477 strand+	*
17	*Acyrthosiphon pisum*	gi320447035 NW003383536.1 Scaffold47:642566-643516 strand +	MITE1.1 sub1
18	*Myzus persicae*	Scaffold6:1582921-1583927 strand+	MITE1.3
19	*Acyrthosiphon pisum*	gi320446957 NW003383614.1 Scaffold125:833628-834589 strand -	MITE1.1 sub1
20	*Myzus persicae*	Scaffold344:286764-288093 strand+	Mpmar1.1
21		Scaffold50:754375-755343 strand+	*
22	*Acyrthosiphon pisum*	gi320446934 NW003383637.1 Scaffold 148:254344-255290 strand +	MITE1.1 sub1
23		gi320446899 NW003383672.1 Scaffold 183:345754-346918 strand +	
24	*Trachymyrmex cornetzi*	gi1006854094 LKEY01038001.1 Contig 38001:392-1605	
25	*Rhodnius prolixus*	Rpmar33 [20]	
26	*Vollenhovia emeryi*	gi763991541 BBUO01005411.1 Contig06950:10342-11676	
27	*Oncopeltus fasciatus*	gi641099433 JHQO01163363.1 ContigNC163363:6062-7386	
28	*Homalodisca vitripennis*	gi642862357 JJNS01041406.1 ContigNC41406:3310-4634	
29	*Myzus persicae*	Scaffold1103:23507-23906 strand+	*
30	*Diuraphis noxia*	gi984745202:272244-271837 strand-	*
31		gi984745488:452325-450979 strand-	Dnomar1.1
32		gi984745255:324638-325555 strand+	*
33	*Myzus persicae*	Scaffold131:48139-48677 strand-	*
34	*Acyrthosiphon pisum*	gi320447046 NW003383525.1 Scaffold36 :908811-909602 strand -	*
35		gi320447034 NW003383537.1 Scaffold48:638002-638846 strand +	*
36		gi320446875 NW003383696.1 Scaffold207:537540-538487 strand -	*
37	*Diuraphis noxia*	gi984745157:267924-268872 strand+	*
38		gi984744390:37892-38851 strand+	*
39	*Acyrthosiphon pisum*	gi320447089 NW003383510.1 Scaffold21:987841-988816 strand +	*
40		gi320446282 NW003384289.1 Scaffold800:146539-147106 strand -	MITE1.1 sub2
41		gi320446572 NW003383999.1 Scaffold510:194743-195684 strand +	
42		gi320446653 NW003383918.1 Scaffold429:453472-454426 strand +	
43		gi320446326 NW003384245.1 Scaffold756:48163-49727 strand +	
44		gi320447005 NW003383566.1 Scaffold77:451686-452577 strand +	*
45		gi320442452 NW003388040.1 Scaffold4551:1640-2495 strand +	*
46		gi320446993 NW003383578.1 Scaffold89:934560-935876 strand +	Apismar1.2
47	*Diuraphis noxia*	gi984745245:252450-251151 strand-	Dnomar1.2
48	*Dendroctonus ponderosae*	gi459605371 APGL01021548.1 Seq01021608:13322-14422	
49	*Gerris buenoi*	gi822390376 JHBY01085489.1 Contig85496: 109-1314	
50	*Homalodisca vitripennis*	gi642470880 JJNS01248885.1 ContigNC248885:1757-2877	
51	*Mesobuthus martensii*	gi553813549 AYEL01086727.1 Contig347321:2028-3230	
52	*Anoplophora glabripennis*	gi496870061 AQHT01063015.1 Contig63076:4332-5534	
***Himar-like elements* DD34D**			
53	*Aphis glycines*	GQ231493	
54	*Drosophila yakuba*	Dromar18 [41]	

55	*Rhodnius prolixus*	Rpmar0 [20]	
56	*Bactrocera dorsalis*	AF346541	
57	*Diachasmimorpha longicaudata*	AY601748	
58	*Chrysoperla plorabunda*	L06041	
59	*Haematobia irritans*	U11642	
60	*Mantispa pulchella*	U11649	
61	*Bactrocera dorsalis*	AY601743	
		***Maibmar-like elements* DD34D**	
62	*Cancer pagurus*	AJ507245	
63	*Eriphia verrucosa*	AM906106	
64	*Thalamita possoinii*	AM906155	
65	*Pilumnus hirtellus*	AM906121	
66	*Xantho hydrophilus*	AM906156	
67	*Maia brachidactyla*	AJ507238	
		***Bytmar-like element* DD34D**	
68	*Alvinella caudata*	AJ496120	
69	*Ventiella sulfuris*	AJ507232	
70	*Perisesarma bidens*	AM906146	
71	*Bythograea thermydron*	AJ507219	
72	*Portunus pelagicus*	AM906137	
73	*Alvinella pompejana*	AJ496135	
		***Nosferatis* DD34D**	
74	*Rhodnius prolixus*	Rpmar13 [20]	
75		Rpmar9 [20]	
		***Dnomar-like element* DD34D**	
76	*Diuraphis noxia*	gi984735891:2467-2015 strand-	*
77		gi984745098:54676-56035 strand+	Dnomar3.1
78	*Myzus persicae*	Scaffold280:15925-14979 strand -	*
79		Scaffold220: 401477-402088 strand +	*
		***Hsmar2-like element* DD34D**	
80	*Portunus pelagicus*	AM906137	
81	*Alvinella pompejana*	AJ496135	
82	*Homo sapiens*	U49974	
83	*Lemur catta*	AC133072	
84	*Gorilla gorilla*	AC145402	
		***Batmar-like element* DD34D**	
85	*Diuraphis noxia*	gi984745342:336477-336902 strand+	*
86		gi984744928:23881-24312 strand+	*
87		gi984745420:365693-366623 strand+	*
88		gi984745311:440129-439207 strand-	*
89	*Acyrthosiphon pisum*	gi320447037 NW003383534.1 Scaffold45:158750-158202 strand-	*
90		gi320446899 NW003383672.1 Scaffold183:261886-261320 strand-	*
91	*Diuraphis noxia*	gi984744320:45892-45397 strand-	*
92		gi984744320:47738-47243 strand-	*
93	*Acyrthosiphon pisum*	gi320446184 NW003384387.1 Scaffold898 :36060-36989 strand -	*
94		gi320446473 NW003384098.1 Scaffold609:132371-133021 strand+	*
95		gi320438813 NW003391455.1 Scaffold7966:4139-4600 strand+	*
96		gi320446392 NW003384179.1 Scaffold690:96269-97548 strand -	Apismar2.2
97		gi320447011 NW003383560.1 Scaffold71:493658-494941 strand +	
98		gi320447000 NW003383571.1 Scaffold82:64692-64037 strand -	*
99	*Heliconius melpomene*	gi378865014 CAEZ01008735.1 Contig7180001235928:24625-25792	
100	*Drosophila eugracilis*	gi449842783 AFPQ02005657.1 Contig5655:975265-976459	
101	*Neodiprion lecontei*	gi914279877 LGIB01001307.1 Scaffold1307:12116-13443	
102	*Lasius niger*	gi861599989 LBMM01019009.1:53-1374	
103	*Diuraphis noxia*	gi984745480:182128-183428 strand+	*
104		gi984745116:79279-80622 strand+	Dnomar2.2
105	*Agrilus planipennis*	gi648140536 JENH01008607.1 Contig8616:12243-13577	

106	*Diuraphis noxia*	gi984745529:295648-297002 strand+	Dnomar2.2
107	*Rhodnius prolixus*	Rpmar1 [20]	
108	*Dinoponera quadriceps*	gi938133368 JPHR01007292.1 Scaffold1145:2209123415	
109	*Copidosoma floridanum*	gi619889135 JBOX01069944.1 Contig69949:1149212815	
110	*Ceratitis capitata*	gi488305875 NW004523814.1 Contig13099:32902-34140	
111	*Drosophila ficusphila*	Dromar8 [41]	
112	*Acyrthosiphon pisum*	gi320445628 NW003384943.1 Scaffold1454:8232- 8961 strand -	*
113		gi320445628 NW003384943.1 Scaffold1454:3454-4361 strand-	MITE2.1
114		gi320447015 NW003383556.1 Scaffold67:738385-739292 strand-	
115		gi320445203 NW003385294.1 Scaffold1805:8796-7866 strand-	
116		gi320446728 NW003383843.1 Scaffold354:173658-174753 strand -	*
117		gi320446952 NW003383619.1 Scaffold130:212179-213501 strand -	Apismar2.1
118	*Myzus persicae*	Scaffold544:202614-203522 strand+	MITE2.2
119		Scaffold166:292461-293374 strand+	
120		Scaffold6:1479795-1480706 strand+	
121	*Diuraphis noxia*	gi984744707:110647-111912 strand+	Dnomar2.1
122		gi984744972:124915-123590 strand-	
123	*Acyrthosiphon pisum*	gi320446088 NW003384483.1 Scaffold994:76058-76816 strand+	*
124	*Blattella germanica*	gi692674178 JPZV01249368.1 ContigNC249368:8211803	
125	*Trabutina mannipara*	gi1044319939 FKYK01006678.1 :1-1114	
126	*Camponotus floridanus*	gi304581076 AEAB01024585.1 Contig622:1601-2726	
127	*Trionymus perrisii*	gi1010807036 FIZV01000272.1 :504-1644	
128	*Trachymyrmex septentrionalis*	gi1006956206 LKEZ01022017.1 Contig22017:3294-4437	
129	*Homalodisca vitripennis*	gi642782498 JJNS01106073.1 ContigNC106073:3898-5039	
130	*Rhinolophus ferrumequinum*	AC157888	
131	*Wasmannia auropunctata*	gi780611046 XM011690486.1	
132	*Rhodnius prolixus*	Rpmar26 [20]	
133	*Carollia perspicillata*	AC148202	
	***Pacmar-like element* DD34D**		
134	*Pachygrapsus marmoratus*	AM231069	
135		AM231072	
136	*Portunus granulatus*	AM906131	
137	*Pachygrapsus marmoratus*	AM983536	
138	*Portunus granulatus*	AM906134	
139		AM906132	
140	*Thalamita possoinii*	AM906151	
141	*Paromola bathyalis*	AM906119	
142	*Perisesarma bidens*	AM906150	
143	*Atelecyclus undecimdentatus*	AM906092	
	***Other mariner elements* DD34D**		
144	*Papilio xuthus*	AB055185	
145	*Attacus atlas*	AB006464	
146	*Hyalophora cecropia*	M63844	
147	*Bombyx mori*	D88671	
148	*Antheraea yamamai*	AB247378	
149	*Antheraea mylitta*	AF126011	
150	*Homo sapiens*	EF517118	
151	*Apis mellifera*	AY155490	
152	*Forficula auricularia*	AY155492	
153	*Ceratitis capitata*	U76903	
154	*Caenorhabditis elegans*	U10438	
155	*Meloidogyne chiwoodi*	AJ437557	
156	*Caenorhabditis briggsae*	AC099767	
157	*Solenopsis invicta*	AF518170	
158	*Solenopsis saevissima*	AF518177	
159	*Myrmica ruginodis*	AY652423	
160	*Bombus terrestris*	AJ312712	
161	*Drosophila mauritiana*	M14653	

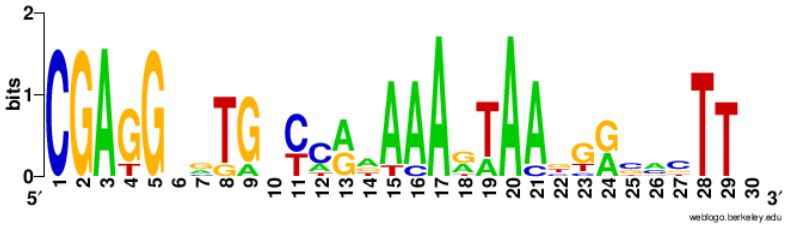
162		X78906	
163	*Drosophila simulans*	X89927	
164	*Mamestra brassicae*	AF465247	
165	*Messor bouvieri*	AJ781769	
166	*Mayetiola destructor*	gi30778846 AEGA01027875.1	
167		U24436	
168	*Rhynchosciara sp.*	GU442128	
	***Crmar2-like element* DD41D**		
169	*Acyrthosiphon pisum*	gi320446057 NW003384514.1 Scaffold1025:1012-1344 strand+	*
170		gi320446057 NW003384514.1 Scaffold1025:2902-3244 strand+	*
171		gi320401144 NW003399120.1 Scaffold15631:349-1 strand-	MITE4.1 sub1
172		gi320447076 NW003383518.1 Scaffold29:1329370-1329000 strand-	
173		gi320446843 NW003383728.1 Scaffold239:535418-535797 strand+	
174		gi320446434 NW003384137.1 Scaffold648:49392-49775 strand-	
175		gi320446106 NW003384465.1 Scaffold976:179217-179591 strand+	
176		gi320446648 NW003383923.1 Scaffold434:173770-174148 strand+	
177		gi320446083 NW003384488.1 Scaffold999:5967-6345 strand+	
178		gi320447199 NW003383491.1 Scaffold2:2053341-2053729 strand+	
179		gi320447197 NW003383492.1 Scaffold3:641055-641464 strand -	*
180		gi320446904 NW003383667.1 Scaffold178:3884-4294 strand -	*
181		gi320441128 NW003389140.1 Scaffold5651:1032-1552 strand -	*
182		gi320447041 NW003383530.1 Scaffold41:488837-490191 strand -	Apismar4.1
183	*Diuraphis noxia*	gi984744291:33825-34544 strand+	*
184		gi984744522:27562-28176 strand+	*
185		gi984745358:548730-549307 strand+	MITE4.2
186		gi984745382:435077-434500 strand-	
187	*Acyrthosiphon pisum*	gi320447027 NW003383544.1 Scaffold55:451821-452338 strand+	*
188		gi320446929 NW003383642.1 Scaffold153:555302-554950 strand-	MITE4.1 sub2
189		gi320446959 NW003383612.1 Scaffold123:513503-513873 strand+	
190		gi320446812 NW003383759.1 Scaffold270:349975-349604 strand-	
191		gi320445944 NW003384627.1 Scaffold1138:18647-18271 strand-	
192	*Locusta migratoria*	AVCP010119604.1	
193	*Mesobuthus martensii*	gi553824729 AYEL01075810.1 Contig333646:1103812150	
194	*Stegodyphus mimosarum*	gi602493518 AZAQ01086514.1 Contig86514:26900-28235	
195	*Acyrthosiphon pisum*	gi320446734 NW003383837.1 Scaffold348:294674-294127 strand-	*
196		gi320445699 NW003384872.1 Scaffold1383:15032-14708 strand-	*
197		gi320447104 NW003383505.1 Scaffold16 :729100-731500 strand+	Apismar4.3
198	*Diuraphis noxia*	gi984745081:151016-151527 strand+	*
199		gi984745384:382222-382755 strand+	*
200		gi984745429:267344-266809 strand-	*
201		gi984745389:349961-349421 strand-	*
202		gi984745380:533248-533974 strand+	*
203		gi984745519:793804-794330 strand+	*
204		gi984745503:728303-727753 strand-	*
205	*Nilaparvata lugens*	gi688034042 AOSB01116314.1 Scaffold930:1140012616	
206	*Heliconius melpomene*	KU514436.1 GI:974707777 32163-33486	
207	*Acromyrmex echinatior*	XM011050935.1	
208	*Acyrthosiphon pisum*	gi320446279 NW003384292.1 Scaffold803:93845-94145 strand+	*
209		gi320445889 NW003384682.1 Scaffold1193:81099-81559 strand -	*
210		gi320446954 NW003383617.1 Scaffold128:616437-616899 strand-	*
211		gi320447044 NW003383527.1 Scaffold38:870577-871352 strand-	*
212		gi320446184 NW003384387.1 Scaffold898:33860-34319 strand+	*
213		gi320447097 NW003383508.1 Scaffold19:246121-246683 strand+	*
214		gi320446974 NW003383597.1 Scaffold108:252404-252965 strand-	*
215		gi320447085 NW003383512.1 Scaffold23:1752408-1753706 strand-	Apismar4.2
216		gi320447044 NW003383527.1 Scaffold38:871366-871812 strand+	*
217	*Atta cephalotes*	gi295962919 ADTU01003898.1 Contig03898:64257703	
218	*Trachymyrmex cornetzi*	gi1006767341 LKEY01058653.1 Contig58653:2926-4236	
219	*Atta colombica*	gi1006787365 LKEW01025605.1 Contig25605:1490716198	

220	*Danaus plexippus*	gi357604284 AGBW01012371.1 :21719-22917	
221	*Dufourea novaeangliae*	gi919891211 LGHO01003380.1 Contig3380:23632-24843	
222	*Harpegnathos saltator*	gi304616639 AEAC01015863.1 Contig6619:7536976586	
223	*Anoplophora glabripennis*	gi496927062 AQHT01044136.1 Contig44173:546-1789	
224	*Anastrepha suspensa*	AY034629	
225		AY034630	
226	*Operophtera brumata*	gi914552887 JTDY01008752.1 2357-3647	
227	*Ceratitis rosa*	AY034623	
228	*Homalodisca vitripennis*	gi642494493 JJNS01240973.1 ContigNC240973:3989-5223	
229	*Anoplophora glabripennis*	gi496912294 AQHT01049420.1 Contig49462:67688047	
230	*Dendroctonus ponderosae*	gi459588353 APGL01038566.1	
231	*Acyrthosiphon pisum*	gi320446058 NW003384513.1 Scaffold1024:28892-29972 strand-	*
232		gi320446789 NW003383782.1 Scaffold293:337514-338606 strand +	*
233	*Metaseiulus occidentalis*	gi391326941 XM003737920.1	
		***LTIR-like element* DD40-41D**	
234	*Acyrthosiphon pisum*	gi320446742 NW003383829.1 Scaffold340:1803-2145 strand+	*
235		gi320446952 NW003383619.1 Scaffold130:17385-18186 strand+	*
236		gi320442315 NW003388177.1 Scaffold4688:7007-6080 strand-	*
237		gi320446920 NW003383651.1 Scaffold162:358972-359636 strand+	*
238		gi320447020 NW003383551.1 Scaffold62:504081-503547 strand-	*
239		gi320447044 NW003383527.1 Scaffold38:494414-493847 strand-	*
240		gi320447038 NW003383533.1 Scaffold44:245172-244580 strand-	*
241		gi320442393 NW003388099.1 Scaffold4610:542-1295 strand+	*
242		gi320433456 NW003396117.1 Scaffold12628:816-1 strand-	*
243		gi320446328 NW003384243.1 Scaffold754:22810-24151 strand+	*
244		gi320446852 NW003383719.1 Scaffold230:461388-461756 strand+	*
245		gi320447106 NW003383503.1 Scaffold14:910161-909675 strand-	*
246		gi320447031 NW003383540.1 Scaffold51:481283-479084 strand-	*
247		gi320446462 NW003384109.1 Scaffold620:170760-168538 strand-	*
248		gi320447015 NW003383556.1 Scaffold67:301784-299478 strand-	Apismar5.1
249	*Myzus persicae*	Scaffold228: 352709..353122 strand+	*
250		Scaffold16:1450072-1450945 strand+	*
251		Scaffold176:395279-395722 strand+	*
252		Scaffold10:1542469-1543323 strand-	*
253	*Diuraphis noxia*	gi984745065:239246-238443 strand-	MITE5.1
254		gi984745517:765990-765175 strand-	
255		gi984745390:259040-259829 strand+	
256		gi984745032:193491-194325 strand+	*
257		gi984745175:253583-254179 strand+	*
258		gi984745400:243420-244241 strand+	MITE5.1
259		gi984745050:86317-85518 strand-	
260		gi984745081:249492-250419 strand+	*
261	*Acyrthosiphon pisum*	gi320446972 NW003383599.1 Scaffold110:354919-355343 strand+	*
262		gi320447020 NW003383551.1 Scaffold62:503547-504119 strand+	*
263		gi320446800 NW003383771.1 Scaffold282:207301-208570 strand+	*
264		gi320446154 NW003384417.1 Scaffold928:21356-19996 strand-	*
265		gi320446987 NW003383584.1 Scaffold95:302694-301750 strand-	*
266		gi320447105 NW003383504.1 Scaffold15:1038009-1036576 strand-	*
267	*Diuraphis noxia*	gi984745508:982719-984070 strand+	*
268	*Acyrthosiphon pisum*	gi320447008 NW003383563.1 Scaffold74:94584-96017 strand+	*
269		gi320442315 NW003388177.1 Scaffold4688:6080-7007 strand+	*
270		gi320446319 NW003384252.1 Scaffold763:18048-18455 strand+	*
271		gi320447082 NW003383514.1 Scaffold25:979065-978002 strand-	*
272		gi320446163 NW003384408.1 Scaffold919:2730-3812 strand-	*
273		gi320447077 NW003383517.1 Scaffold28:658930-658046 strand-	*
274		gi320447191 NW003383495.1 Scaffold6:867979-866906 strand-	*
275	*Myzus persicae*	Scaffold1040:29624-30705 strand+	*
276		Scaffold1981:5634-7149 strand+	*
277	*Acyrthosiphon pisum*	gi320446879 NW003383692.1 Scaffold203:311633-311055 strand-	*

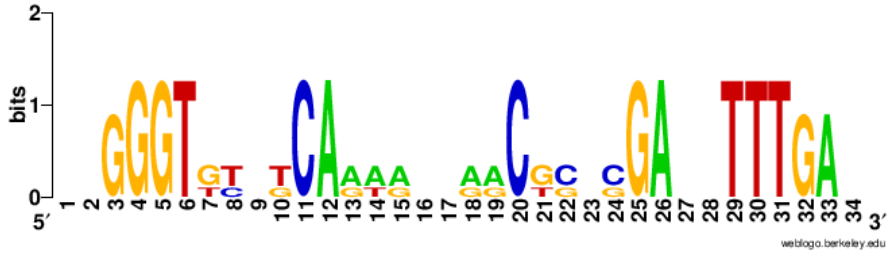
278		gi320447177 NW003383497.1 Scaffold8:947336-948127 strand+	*
279		gi320446579 NW003383992.1 Scaffold503:316494-315441 strand-	*
280	*Diuraphis noxia*	gi984745469:593111-591544 strand-	*
281	*Acyrthosiphon pisum*	gi320447082 NW003383514.1 Scaffold25:163596-162726 strand-	*
282		gi320434094 NW003395907.1 Scaffold12418:13-883 strand+	*
283		gi320392825 NW003401900.1 Scaffold18411:6357-5063 strand-	*
284	*Bombus terrestris*	gi339751187 AELG01000709: 6387911-6385650	
285	*Nasonia vitripennis*	gi154053393:26728-27797	
286	*Harpegnathos saltator*	NW011646526.1 Scaffold1231:24830-28107	
287	*Pogonomyrmex barbatus*	NW011933482.1 Scaffold7180000350099:45895-49565	
288	*Calycopis cecrops*	LUGF01035226.1	
289	*Cimex lectularius*	JRLE01000220.1 Scaffold310843	
290	*Rhodnius prolixus*	KQ034065 Scaffold9:2936024-2933697	
291	*Anoplophora glabripennis*	gi496979286 AQHT01025496.1 Contig25506:11686-12762	
292	*Homalodisca vitripennis*	KK962044.1 Scaffold551:1019188-1021387	
293	*Megachile rotundata*	gi383080544 Scaffold0130:560690-562464	
294	*Acyrthosiphon pisum*	gi320447045 NW003383526.1 Scaffold37:902476-902066 strand+	MITE5.2 sub1
295		gi320446511 NW003384060.1 Scaffold571:438535-438949 strand+	
296		gi320446471 NW003384100.1 Scaffold611:61351-61779 strand+	
297		gi320446667 NW003383904.1 Scaffold415:160458-160887 strand+	
298		gi320446129 NW003384442.1 Scaffold953:162133-162549 strand+	MITE5.2 sub2
299		gi320446884 NW003383687.1 Scaffold198:591126-591545 strand+	
300		gi320446815 NW003383756.1 Scaffold267:153990-154430 strand+	
301		gi320447083 NW003383513.1 Scaffold24:949317-949750 strand+	*
302		gi320447001 NW003383570.1 Scaffold81:754854-757276 strand+	Apismar5.2
303		gi320446306 NW003384265.1 Scaffold776:57772-58831 strand+	*
304		gi320446984 NW003383587.1 Scaffold98:28691-29172 strand+	*
305		gi320446449 NW003384122.1 Scaffold633:131583-130926 strand-	*
306		gi320446999 NW003383572.1 Scaffold83:210237-208812 strand-	*
307		gi320446763 NW003383808.1 Scaffold319:416807-419008 strand+	*
***maT* DD37D**			
308	*Caenorhabditis elegans*	AF038612	
309		Z83129	
310	*Caenorhabditis briggsae*	AC084524	
311	*Bombyx mori*	U43131	
312	*Anopheles gambiae*	AAAB01008975	
Outgroup TLE and IS630			
313	*Fusarium oxysporum*	AF282722	
314	*Caenorhabditis briggsae*	M64308	
315	*Drosophila virilis*	CH940657	
316	*Salmo salar*	AJ249090	
317	*Anopheles gambiae*	U89802	
318	*Aedes aegypti*	AF208675	
319	*Caenorhabditis elegans*	M77697	
320	*Aedes atropalpus*	AY038027	
321	*Fusarium oxysporum*	AF076632	
322	*Fusarium solani*	AF443562	
323	*Fusarium oxysporum*	AF076631	
324		AJ608703	
325	*Aspergillus niger*	U58946	
326	*Sinorhizobium meliloti*	AF143444	
327	*Pseudomonas sp.*	U15298.1	
328	*Catharanthus roseus*	DQ852611	
329	*Salmonella typhimurium*	M58505	

Additional file 4. TIR sequences for each clade. TIR consensus per clade was generated using the Web-Logo server (http://weblogo.berkeley.edu/logo.cgi). At each position the nucleotides are stacked one on top of another with the most frequent one on the top. It displays the frequency of bases at each position with height indicating the proportion of occurrence. The vertical scale is in bits with maximum of two bits possible at each position indicating that there can be possibility of four different bases at each position. For LTIR- like element, only the first 57 nucleotides are presented.

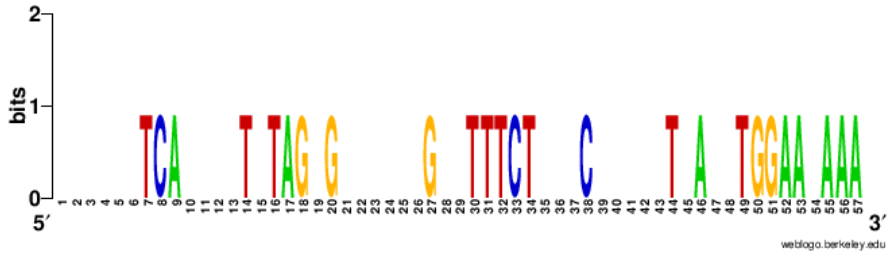
Consensus of TIR-*irritans*



Consensus of TIR-*rosa*



Consensus of TIR-*LTIR*



Additional file 5. Pairwise divergence matrix between amino acid lineages.
Fifteen complete sequences have been aligned using Aliview. The alignment was then transferred in GENEDOC software [64] to obtain the identity percentage.
Sequences name: *Apismar*: elements from *Acyrthosiphon pisum*, *Dnomar:* elements from *Diuraphis noxia* and *Mpmar*: elements from *Myzus persicae.*

	Macrosiphinimar					*Batmar-like element*				*Dnomar-like element*	*Rosa*			*LTIR-like element*	
	Apismar1.1	Apismar1.2	Dnomar1.1	Dnomar1.2	Mpmar1.1	Apismar2.1	Apismar2.2	Dnomar2.1	Dnomar2.2	Dnomar3.1	Apismar4.1	Apismar4.2	Apismar4.3	Apismar5.1	Apismar5.2
Apismar1.1		66%	88%	62%	95%	54%	49%	47%	52%	58%	32%	31%	30%	29%	33%
Apismar1.2			62%	79%	67%	50%	48%	43%	48%	54%	34%	32%	31%	32%	34%
Dnomar1.1				59%	89%	53%	48%	46%	52%	55%	32%	32%	29%	28%	33%
Dnomar1.2					63%	49%	45%	42%	43%	52%	32%	33%	30%	31%	34%
Mpmar1.1						55%	51%	47%	52%	59%	33%	33%	31%	30%	33%
Apismar2.1							66%	78%	64%	53%	36%	37%	36%	34%	34%
Apismar2.2								56%	66%	50%	35%	35%	37%	31%	32%
Dnomar2.1									53%	44%	32%	34%	33%	34%	31%
Dnomar2.2										52%	35%	34%	35%	32%	32%
Dnomar3.1											36%	32%	38%	35%	33%
Apismar4.1												68%	60%	44%	42%
Apismar4.2													56%	46%	44%
Apismar4.3														43%	41%
Apismar5.1															52%
Apismar5.2															

Analyse globale et Conclusions

Dans ce chapitre, nous nous sommes intéressés à identifier et à caractériser les éléments transposables *mariner* DD34D et *rosa* DD41D dans les génomes d'*Acyrthosiphon pisum*, *Diuraphis noxia* et *Myzus persicae* qui appartiennent à la tribu des Macrosiphini. Pour cela, 18 séquences de transposases représentant ces deux groupes ont été utilisées comme requêtes dans chacun de ces génomes (TBLASTN). Après filtration des séquences obtenues (élimination des séquences ayant une taille inférieure à 250 pb, ou avec un motif catalytique de type DDxE), les résultats ont permis d'extraire un total de 115 séquences, regroupées en 11 clusters (groupes de séquences présentant une homologie ≥ 75%) chez *A. pisum*, 45 séquences réparties en sept clusters chez *D. noxia* et 23 séquences distribuées en 4 clusters chez *M. persicae*.

Tous les éléments identifiés ont été classés selon la méthode UPGM-VM (Rouault *et al.* 2009) avec 146 autres séquences nucléotidiques représentant la diversité des séquences de la superfamille *Tc1-mariner-IS630*. Parmi les 183 séquences extraites, 118 présentent des délétions internes ou sont dépourvues de TIR et sont incluses dans cette classification qui ne tient pas compte de la taille de l'élément mais du pourcentage de similitude.

Au total, 11 clusters ont été classés au sein de la sous-famille *irritans* de la famille des *mariner* DD34D. Ils sont subdivisés en trois tribus : *Macrosiphinimar* (*irritans* spécifique chez ces pucerons), *Batmar-like* element, proches de l'élément de référence *Rfmar* découvert chez la chauve-souris *Rhinolophus ferrumequinum* et *Dnomar-like* element, comprenant une séquence complète de *D. noxia* avec des séquences délétées/tronquées chez *M. persicae* et *D. noxia*. Ces éléments présentent des motifs spécifiques des transposons *mariner* avec des TIR assez proches de 15 à 30 pb suggérant l'existence d'un ancêtre récent. Toutefois, il faut noter que les autres sous-familles de *mariner* (*i.e. mauritiana, mellifera, cecropia, elegans*…) ne sont pas présentes dans ces génomes, bien qu'elles aient été trouvées *in vitro* chez d'autres espèces appartenant à des tribus phylogénétiquement proches telles que le puceron du soja et les pucerons des arbres fruitiers (Mittapalli *et al.* 2011; Kharrat *et al.* 2015).

Les autres clusters (11) sont éloignés de la branche de la famille des *mariner*. Certains éléments, identifiés uniquement chez *A. pisum* et *D. noxia* se sont regroupées dans le clade *rosa* représenté par la séquence de référence *Crmar2* identifiée chez le diptère *Ceratitis rosa*. Ils sont caractérisés par la présence d'un motif DD41D au niveau de la transposase ainsi que de TIRs de tailles de 28 à 32 pb.

D'autres éléments trouvés dans les trois génomes, se distinguent par de long TIRs de tailles d'environ 460pb et sont divisés en deux groupes selon le motif de la triade catalytiques DD40D et DD41D. Ils forment ainsi un nouveau clade phylogénétiquement proche de *rosa* que nous avons désigné *LTIR-like* elements.

Parmi l'ensemble des 22 clusters définis, 15 d'entre eux présentent des copies complètes. Ils sont répartis dans différents clades monophylétiques à savoir : *irritans* (10), *rosa* (3) et *LTIR* (2). Parmi ces clusters, cinq chez *A. pisum* et un cluster chez *M. persicae* comprennent des séquences potentiellement actives (absence de codons stop, d'indels ou de déphasage du cadre de lecture).

Il faut noter que des éléments appartenant au moins à quatre clusters coexistent dans le même génome. Toutefois la grande diversité de ces éléments entre les espèces peut refléter leur histoire évolutive indépendante. En effet, la distribution et la phylogénie de ces ETs sont probablement le résultat de transmissions verticales associées à un polymorphisme ancestral. Ainsi des séquences étroitement apparentées dérivées de la même copie ancestrale peuvent être présents chez plusieurs espèces, tandis que des copies dérivées de différentes copies ancestrales et trouvées dans le même génome, peuvent être plus éloignées phylogénétiquement (Capy *et al.* 1992a; Green et Frommer 2001; Prasad *et al.* 2002). De plus, le génome de ces pucerons est capable de réprimer l'activité de transposition (Rozhkov *et al.* 2010; Tóth *et al.* 2016), ce qui conduit à l'élimination des ETs par la perte stochastique ou l'extinction verticale (Le Rouzic *et al.* 2007). Par conséquent, l'absence de membres de la famille *rosa* peut être due à une perte stochastique pendant la trajectoire évolutive de *M. persicae*.

Par ailleurs, pour déterminer l'existence éventuelle de transferts horizontaux, les séquences nucléotidiques des éléments complets identifiés précédemment chez ces pucerons ont été utilisées comme requêtes dans la base de données NCBI-nr (BLASTN) et WGS des Eucaryotes. Une seule séquence détectée chez le coléoptère *Agrilus planipennis* présente une homologie supérieure à 90% couvrant 90% de la séquence *Dnomar2.2* de *D. noxia*, appartenant à la sous-famille *irritans*.

En outre, l'analyse de l'arbre phylogénétique basé sur les séquences complètes protéiques et la comparaison des motifs de la triade catalytique ont révélé que les deux clades *rosa* et *LTIR* sont plus étroitement liés à la famille de *maT* (DD37D) que de celle de *Tc1* et de *mariner*. Ceci suggère que ces deux clades peuvent constituer une nouvelle famille DD40-41D appartenant à la superfamille des *Tc1-mariner-IS630*.

En plus des séquences complètes autonomes, des éléments courts ou miniatures (MITE = *Miniature Inverted Transposable Elements*) ont été décrits. Ces copies sont dépourvues de la totalité ou d'une partie de l'ORF mais restent *trans*-mobilisables par des séquences capables de

coder une transposase fonctionnelle. Un total de 43 séquences distribuées en 11 sous-groupes et appartenant à quatre tribus (*Macrosiphinimar*, *Batmar-like* element, *rosa* et *LTIR-like* elements) ont pu être détectées. Souvent, des copies autonomes proches (sur la base des TIRs) ont été identifiées indiquant que les MITEs proviendraient de délétions internes à partir des éléments complets. Plus précisément, sept sous-groupes (*MITE1.1 sub1, MITE1.2, MITE2.1, MITE4.1 sub1-2, MITE5.2 sub1-2*) présentent une forte homologie avec les éléments complets trouvés dans la même espèce. En revanche, trois sous-groupes à savoir *MITE2.2* identifiés chez *M. persicae,* et *MITE4.2, MITE5.1* chez *D. noxia* sont fortement reliés aux éléments autonomes d'une autre espèce à savoir *A. pisum.* Un seul sous-groupe *MITE1.1 sub2* reste sans partenaire potentiel.

Ces résultats suggèrent donc que :

- le nombre de MITE par clade et par espèce de pucerons reste faible par rapport à ce qui est connu, notamment chez les plantes (Feschotte *et al.* 2002).
- la plupart des MITEs présentent des microhomologies internes au niveau des points de rupture de la délétion (*Breaking Point Exact Exact* BPE, *Breaking Point Near Near* BPNN) suggérant que ces délétions internes sont probablement dues à une interruption de la réparation de l'excision de l'ETs (*abortive gap repair*).
- seuls les éléments de la sous-famille *irritans* semblent incapables de générer des MITEs de taille inférieure à 900pb.
- plusieurs scénarios peuvent être proposés pour expliquer la distribution, pas toujours congruentes, des MITEs et des séquences autonomes associées. Par exemple, l'existence d'un polymorphisme ancestral de séquences génératrices de MITEs et de pertes stochastiques de ses séquences pourrait expliquer la présence de MITEs dans une espèce et des séquences complètes associées dans une autre espèce.
- le transfert horizontal peut expliquer la distribution aléatoire des éléments MITEs chez ces pucerons.

En conclusion, cette partie a permis de décrire précisément la composition en *MLE* au sein de trois espèces de pucerons. L'abondance relative de ces éléments est faible par rapport à celle observée dans d'autres génomes d'insectes, par exemples, le génome de 700 Mb de l'hémiptère *Rhodnius prolixus* et celui de 156MB du diptère *Drosophila eugracilis* comprenant respectivement 10836 et 642 copies de *MLE* (Wallau *et al.* 2014; Filée *et al.* 2015).

La classification et l'analyse phylogénétique des 183 séquences ont suggéré que ces éléments répartis en 22 clusters sont divisés en trois groupes : la sous-famille *irritans* DD34D, *rosa* DD41D et un nouveau groupe DD40/41D proche de *rosa*, caractérisé par de longs TIRs.

Par ailleurs, les résultats montrent que des transferts verticaux avec des pertes stochastiques et/ou des événements de transferts horizontaux ont pu exister. Ces observations soulignent la grande diversité d'éléments et de structures que l'on peut observer au sein et entre génomes même phylogénétiquement proches, fournissant de nouvelles informations sur l'histoire évolutive de ces éléments transposables chez ces pucerons.

Chapitre II

Recherche et caractérisation des éléments *irritans* de la sous-famille *mariner* chez quatre espèces de pucerons des céréales. Et si ces transposons sont aussi présents chez les plantes hôtes ?

Characterization of *irritans* subfamily in four cereal aphid species. What if these transposons occur in their host plants?

Chapitre II

Recherche et caractérisation des éléments *irritans* de la sous-famille *mariner* chez quatre espèces de pucerons des céréales. Et si ces transposons sont aussi présents chez les plantes hôtes ?

Après avoir analysé *in silico* les génomes des trois espèces *A. pisum, M. persicae* et *D. noxia* appartenant à la tribu des Macrosiphini, trois clades d'éléments transposables (*irritans, rosa* et *LTIR-like* elements) ont été identifiés. Alors que la sous-famille *irritans*, subdivisée en trois tribus, est commune aux trois génomes, les clades *rosa* et *LTIR-like* elements semblent être plus répandus dans le génome d'*A. pisum*. Par ailleurs, nous avons remarqué, à partir des données de la littérature, que bien que la sous-famille *irritans* appartenant à la famille des transposons *mariner* DD34D soit largement distribuée au sein des espèces, la majorité de ces éléments n'est pas forcément fonctionnelle.

Dans ce chapitre, nous nous sommes intéressés à la recherche et la caractérisation *in vitro* des éléments de la sous-famille *irritans* DD34D chez les pucerons des céréales *Rhopalosiphum padi, R. maidis, Schizaphis graminum, Sitobion avenae* en utilisant des amorces déduites des résultats obtenus précédemment (chapitre I). La connaissance de la structure de ces éléments et de leur évolution pourrait être à la base du développement de nouvelles stratégies de contrôle génétique.

Par ailleurs, nous avons investigué *in silico* et *in vitro* chez les plantes-hôtes (céréales : orge, blé, avoine, égilope) la présence des éléments *irritans* identifiés chez les pucerons. Il est toujours important de développer ce type de recherche entre des espèces qui interagissent fortement. En effet, si des éléments proches sont retrouvés chez un puceron et sa plante-hôte, cela pourrait être le résultat d'un transfert horizontal (TH). La bonne connaissance de ce type de situation pourrait également permettre d'aborder les mécanismes à l'origine de ces transferts qui pour l'instant ne sont qu'au stade des hypothèses. Plusieurs exemples de TH de ETs ont été suggérés chez les Eucaryotes entre espèces appartenant à un même règne animal ou végétal (Robertson et Lampe 1995; Gilbert *et al.* 2010; Dupeyron *et al.* 2014; El Baidouri *et al.* 2014). Toutefois, aucune étude de TH entre animal et plante n'a été menée.

Il faut signaler que la recherche *in vitro* a nécessité la vérification préalable de l'absence de contamination au niveau des feuilles des céréales par du tissu d'origine animale en utilisant des amorces ciblant le gène de la chitine synthase1 spécifique des insectes et des champignons.

Les résultats de ce chapitre font l'objet d'un article en cours de rédaction en vue d'une soumission prochaine.

Characterization of *irritans* subfamily in four cereal aphid species. What if these transposons occur in their host plants?

Maryem Bouallègue[1, 2], Wafa Ben Lazhar-Ajroud[2], Maha Mezghani-Khemakhem[2], Hanem Makni[2, 3], Aurélie Hua-Van[1], Mohamed Makni[2] and Pierre Capy[1]

[1] Laboratoire Evolution, Génomes, Comportement, Ecologie CNRS, Univ. Paris-Sud, IRD, Université Paris-Saclay, 1 avenue de la Terrasse, 91198 Gif-sur-Yvette Cedex, France;

[2] Université de Tunis El Manar, Faculté des Sciences de Tunis, UR11ES10 Génomique des Insectes Ravageurs de Cultures, 1002, Tunis, Tunisie;

[3] Institut Supérieur de l'Animation pour la Jeunesse et la Culture de Bir-El-Bey, Université de Tunis, 2055, Tunis, Tunisie.

Abstract

The *mariner* family of transposable elements (TEs) is one of the most widespread in Metazoans. Like other TEs, their activity thanks to a self-promoting mobility and duplication lead to increase their copy number in genomes ensuring their own perpetuation. Moreover, horizontal transfers (HT) is also a frequent phenomenon encountered during TE evolution. All these events are source of structural and functional diversity. Previous *in silico* study, based on the analysis of three aphid genomes belonging to Macrosiphini tribe, showed that three monophyletic clades of *irritans* subfamily of element containing complete elements and MITEs can be detected. In the present work, *in vitro* identification of members of this subfamily, in four cereal aphid species whose genomes are still absent in databases, has been done. Firstly, two elements with internal deletions, corresponding to MITEs, have been isolated using their terminal inverted repeats (TIRs) as primers. All detected *irritans* sequences are clustered together suggesting a common ancestor. Secondly, the sequenced genomes of *Hordeum vulgare*, *Brachypodium distachyon*, *Triticum aestivum* and *Aegilops tauschii* were explored to detect whether sequences belonging to the *irritans* subfamily were present in the aphid's hosts. Interestingly, one contig from barley cultivar *barke*, includes a truncated element closely related to one complete sequence of aphids *Apismar1.1*. Two types of sequences specific to *Roho* cultivar were detected. The first one corresponds to the truncated element identified in *silico* from cultivar *barke*, the other one shows 98% of homology with the deleted *irritans* sequence from *Sitobion avenae*. Therefore, our results strongly suggest of the occurrence of *irritans* subfamily in cereal aphids and a possible HT between cereal aphids/*H. vulgare*. However, the pathway of this transfer remains to elucidate.

Key words: *irritans* subfamily; aphids; cereals; horizontal transfer (HT).

Introduction

Transposable elements (TEs) are ubiquitous and involved in genome architecture evolution (size, adaptability and structure) as well as gene regulation (Biémont and Vieira 2006; Feschotte and Pritham 2007; Bennetzen and Wang 2014).

Within *Class II* (DNA transposons), the *mariner* is probably one of the best known families. It has been isolated from a vast range of phylogenetically distant organisms, such as protozoa (Silva *et al.* 2005), arthropods (Robertson 1993; Yamada 2015) and mammals (Auge-Gouillou *et al.* 1995). According to the current classification (Wicker *et al.* 2007), *mariner*-like elements *(MLEs*) belong to the *Tc1-mariner* superfamily, using a "cut and paste" mechanism (Robertson 1993; Hartl *et al.* 1997). The full-length of an *MLE* is about 1.3 kb. It contains terminal inverted repeats (TIRs), untranslated terminal regions (UTR) and one intronless open reading frame (ORF) that encodes a transposase of about 340 amino acids with a DD34D catalytic motif (Robertson and Macleod 1993; Shao and Tu 2001). Initially, phylogenetic analyses led to define five *mariner* subfamilies *mauritiana, cecropia, elegans/briggsae, mellifera/capitata* and *irritans* (Robertson 1993; Robertson and MacLeod 1993). Other 16 minor subfamilies were then proposed by Rouault *et al.* (2009).

Due to independent accumulation of mutations, which increases divergence between copies, *MLE* sequences become nonfunctional, and compared to autonomous element, inactive copies are more frequent in genomes (Muñoz-López and García-Pérez 2010; Filée *et al.* 2015; Robillard *et al.* 2016). Especially, the Miniature Inverted repeat Transposable Elements (MITEs) emerge from complete elements, by internal deletions that are flanked by microhomologies boarding the breakpoint of deletions (Brunet *et al.* 2002; Negoua *et al.* 2013). MITEs are characterized by a short length, with TIRs and no coding potential. As non-autonomous elements, they can be *trans*-mobilized by autonomous copies (Feschotte and Pritham 2007). MITEs were first described from plant genomes (Bureau and Wessler 1992, 1994; Feschotte and Mouches 2000) but have also been identified in several animals (see references in Lu *et al.* 2012). Among taxa, the high sequence similarity of autonomous/non-autonomous *MLEs* and their non-uniform distribution, strongly suggested that horizontal transfers (HT) have occurred among species (Schaack *et al.* 2010; Dupeyron *et al.* 2014; Wallau *et al.* 2016). To date, HT events between either plant or animal species have been described (Robertson and Lampe 1995; Loreto *et al.* 2008; Gilbert *et al.* 2010; Dupeyron *et al.* 2014; El Baidouri *et al.* 2014). Generally, these transfers are considered as rare events, but it remains difficult to estimate their frequency since only methods based on homology allows us to detect the successful ones. However, as suggested by Loreto *et al.* (2008), they are

probably much more frequent than initially suspected even if the mechanisms of transfers remain unknown in eukaryotes.

Previously, comparative genomics of the three available aphids (pea aphid *Acyrthosiphon pisum*, peach aphid *Myzus persicae* and Russian wheat aphid *Diuraphis noxia*), which belong to Macrosiphini Tribe, reveal the prevalence of *irritans* subfamily. In this study, taking *A. pisum* as a control, we reported the *in vitro* identification and characterization of *irritans* copies in four most devastating cereal aphids belonging to two tribes: Aphidini (*Rhopalosiphum padi*, *R. maidis*, *Schizaphis graminum*) and Macrosiphini (*Sitobion avenae*) which have diverged approximately since 44.5 Mya (Vea and Grimaldi 2016).

Because these aphids ravaged cereal crops such as barley, wheat as well as oat and, can transmit several phytoviruses leading to the decrease of agricultural productions (Miller and Pike 2002), an *in silico* and *in vitro* exploration of the plant genomes was done to check whether TE of the *irritans* subfamily were present. All obtained sequences were then compared to evaluate their dynamics and evolutionary history as well as the putative existence of HT.

Material and methods

Sample collections and DNA isolation

During 2012 to 2014, four species belonging to the Aphididae family *i.e. Rhopalosiphum padi, R. maidis, Sitobion avenae* and *Schizaphis graminum* were collected on leaves of several host plants (barley, wheat and oat) at the CRRGC (Centre Régional des Recherches en Grandes Cultures) (36.44N, 9.13E) in Beja, in Tunisia. Species were determined on the basis of the identification keys (Blackman and Eastop 2000). Leaves of host plants as well as plant samples, including *Aegilops tauschii* seeds provided by CIMMYT (The International Maize and Wheat Improvement Center, Mexico) and leaves of several barley varieties (Manel, Martin, Bowman, Roho, HD29) and wheat (BLG, Salambo) sampled from CRRGC were also collected. The pea aphid *Acyrthosiphon pisum* was used as a positive control.

Each sample was kept at -80°C until DNA extraction. Genomic DNA was extracted using the Cetyl Trimethyl Ammonium Bromide (CTAB) method (Doyle and Doyle 1987) or using an extraction kit following the protocol provided by the manufacturer (*i.e.* Wizard® Genomic DNA Extraction Kit, Promega® or NucleoSpin® Tissue DNA Kit, Macherey-Nagel).

Polymerase chain reaction (PCR/ inversePCR), cloning and sequencing

To amplify *irritans* subfamily, six TIR degenerated primers including the TA dinucleotide target site of duplication (**Table 1**) were designated from the alignment of TIRs identified in the previous

genomic comparison (data unpublished). PCR reactions were performed using 50 to 100 ng of genomic DNA in a 25µl reaction mixture comprising 0.1 U of GoTaq polymerase (Promega), 1X PCR buffer, 2 mM MgCl2, 0.1 mM of each primer, and 0.2 mM dNTPs, with the following protocol: an initial denaturing step at 94 °C for 5 min, 35 cycles [95 °C for 60 s; Th (temperature of hybridization) for 60 s; 72 °C for 90 s] and ending with a final extension at 72 °C for 7 min.

In order to determine the true flanked regions of specific copies, inverse PCR (iPCR) was done, using the same amplification steps (Ochman *et al.* 1988) with an initial genomic DNA (500 ng) digested with 10 U of restriction endonuclease SmaI or HaeII (Promega®) at 37°C for 5 h. Restriction fragments were then circularized in 150 µl of ligation buffer with 100U of T4 DNA ligase (Promega®) at 37°C for 5 h. After inactivation of the endonucleases at 65 °C for 10 min, 10 µL of the ligated gDNA template was used for iPCR using specific primers *InvF1*/*InvR1* (**Table 1**).

To confirm non-contamination of the plant DNA with insect one, the amplification of chitin synthase1 gene was performed using primers designated from the alignment of the conserved catalytic domain of this gene found in arthropod species (**Table 1**).

The PCR/iPCR products were purified using Wizard® SV Gel and PCR Clean up System kit, Promega® and cloned into the pGEM T Easy Vector (Promega®) according to the manufacturer's protocol. *E. coli DH5a* cells (New England Biolabs) were then transformed. Positive clones were screened during a subsequent PCR using T7 and SP6 primers. Plasmids were isolated and purified using a Wizard Plus Minipreps DNA Purification system (Promega), and inserts were sequenced on both strands.

Data mining from genomes of cereals

The complete nucleotide sequences and MITEs, belonging to *irritans* subfamily, previously identified *in silico* (unpublished data) and *in vitro* were used as queries in BLASTn searches, with the default parameters, against the WGS of *Hordeum vulgare* and *Triticum aestivum* databases available on the NCBI (**Table 2**). The analysis was extended to two sequenced genomes of *Brachypodium Distachyon* (a small genome) and *Aegilops tauschii* (wheat ancestor). Copies located at the end of scaffolds and/or less than 300 bp were removed.

Sequence analyses and classification

The sequence analyses, including alignment were done with the Aliview 1.18 (Larsson 2014). The percentage of identity between sequences was determined using the GeneDoc software (Nicholas *et al.* 1997). The amino acid sequences were deduced by ExpasyTool (http://web.expasy.org/translate/) and then manually optimized. The nuclear localization sequence

(NLS) and the helix turn helix (HTH) domains were searched using PSORTII (Bannai *et al.* 2002) and GYM2.0 (Gao *et al.* 1999; Narasimhan *et al.* 2002), respectively.

In order to classify nucleotide sequences, copies of *irritans* subfamily were added to a set of already known sequences belonging to the *Tc1-mariner-IS630* superfamily. The classification is based on the Unweighted Pair Group Method with Variation of Metric UPGM-VM (Rouault *et al.* 2009) and allows gathering in the same group a complete sequence with the corresponding truncated and/or deleted sequences such as MITEs (including several lengths of TEs). In fact, this method is an ascending hierarchical classification analogous to UPGMA method, with twdo main differences: (i) there is no arithmetical mean, the nucleotide sequences are aligned by pairs, (ii) the metric varies with the ascending classification and gap is considered as a fifth nucleotide.

Results

Characterization of *irritans* subfamily in cereal aphids

Using six degenerated TIR primers (**Table 1**), *irritans* subfamily elements were searched in *Rhopalosiphum padi*, *R. maidis*, *Sitobion avenae* and *Schizaphis graminum*. The pea aphid *Acyrthosiphon pisum*, whose genome has revealed several copies of *irritans* TE, from an *in silico* investigation, was used as a positive control. Only PCR amplifications with TIR-G1, designated from the lineage of *Apismar1.1* (unpublished data) generates two types of products. The first one, of about 900 bp was obtained in all aphid samples, whereas the second one, of about 600 bp, was specific to *S. avenae* (**Figure 1A1**). All PCR products were cloned and five random clones per individual (for three individuals per species) and per species (given a total of 15 clones per PCR product per species) were sequenced.

The 75 sequences, ranging from 943 to 946 bp, were identified from the five aphids' species and using BLASTX, belonged to the *MLE* transposase family (pfam01359 in NCBI). While alignment between these sequences showed specific synapomorphies per species (*Apmar1*, *Rpmar1*, *Rmmar1*, *Samar1* and *Sgmar1*- more details in **Figure 2**), pairwise comparison revealed a high degree of identity upper to 98%, allowing to constitute a single cluster named *Aphidmar*. In addition, elements present 94.5% ± 1.2 of similarity with the sequence of *Apismar1.1*, extracted from *A. pisum* genome. The 5′- and 3′-UTRs are 143 and 79 bp long, respectively with a polyAdenylation signal conserved. The transposase gene contained internal deletions of 378 bp and point mutations all along its sequence. The binding domain (HTH, WVPKQL) was lost and the second canonical motif (VCLLH**D**NA) of the transposase was modified (**Figure 1B1**), with small deletions in *Rpmar1*, *Rmmar1* and *Samar1* as well as a substitution in *Apmar1* and *Sgmar1*.

The 15 remaining sequences from *S. avenae* constitute a single identical sequence of 653 bp (*Samar2*). Due to the significant deletion of 682 bp near the 3' end, leading to the absence of the catalytic triad core, the membership of this element was difficult to recognize by BLASTX. However, its nucleotide alignment with *Apismar1.1*, revealed 92% of identity. Contrary to *Aphidmar*, no NLS motif was detected for *Samar2* while HTH was found.

Thus, all the 90 sequences amplified present internal deletions. They are unable to transpose autonomously but can be *trans*-mobilizable by autonomous partner and can be considered as MITEs. Small direct repeat (SDR) microhomologies were manually searched at the breaking points (BP) of the internal deletions, as proposed by Negoua *et al.* (2013). SDR are detected in *Aphidmar* and *Samar2*, located near the BP (BPNN) and exactly in the breaking points (BPEE), respectively (**Figure 2)**, suggesting that these short copies were originated from an abortive gap repair after an excision event.

In silico investigation of *irritans* subfamily in host plant species

The Whole Genome Sequence (WGS) of four cereal species *Brachypodium distachyon, Aegilops taushii, Triticum aestivum* and *Hordeum vulgare*, available in the GenBank database were investigated to characterize potential copies *irritans* subfamily (**Table 2**). The ten complete elements belonging to *Macrosiphinimar*, *Batmar-like elements* and *Dnomar-like elements* identified from the three aphids genome namely *A. pisum*, *M. persicae*, *D. noxia* (unpublished data) as well as MITEs *Aphidmar* and *Samar2* (**Figure 2)** were used as queries.

While the majority of contigs/scaffolds exhibit truncated copies of elements belonging to *irritans* subfamily at their extremities, only one contig (#289869) from the genome of *H. vulgare* cultivar *barke* shows 76% and 79% of identity with *Aphidmar* and *Apismar1.1* respectively, over 38% and 27% (i.e. more than 300 bp) of their length. As expected, similar results were obtained with *Mpmar1.1* and *Dnomar1.1*, *i.e.* copies showing about 92% of similarity with *Apismar1.1*. Therefore, only *Apismar1.1*will be used in the following analyses.

Interestingly, this contig of 1471 bp, includes truncated *irritans* element (320 bp), flanked by genomic DNA (gDNA) from aphids (*i.e A.pisum, M. persicae, D. noxia*), the whole surrounded by gDNA from *H. vulgare* host cultivar *barke* (**Figure 3**). Nevertheless, analysis of these flanking regions showed that these sequences are not characterized as EST. This structural organization could be explained by molecular recombination during evolution, followed by the insertion of an *irritans* copy.

This truncated element (named *Hvumar1*) corresponds to the 3'end of an *irritans* copy and comprises only the third aspartic residue of the catalytic domain of the transposase that was slightly

modified on PYSP-LAPT**D**YH from the conserved one as well as the same 3'TIR. In addition, it displayed several mutations and *indels.*

When *Samar2* is used as query, a single fragment shorter than 230 bp can be detected. Otherwise, in the other cultivars and plant species explored during this work, whatever the queries used, no sequence similar to those of the *irritans* subfamily were found.

***In vitro* verification of *irritans* subfamily in plants**

The absence of insect DNA in gDNA extracted from cereal plants (barley, *Triticum aestivum*, oat and aegilops) was confirmed by a PCR targeting an arthropod-specific gene encoding *chitin synthase1*. The amplification was carried out using a pair of degenerated primers, designated from a conserved region of 410 bp corresponding to the catalytic site of *chitin synthase 1* (**Table 1**). The four cereal aphids were considered as positive controls of the amplification. A single product of about 410 bp is detected in the genomes of the four aphid species (only *R. padi* in shown in **Figure 1A2**) and no similar amplification is observed in cereals. Sequencing of the 410 bp fragment, clearly confirmed its similarity with the *chitin synthase 1* gene. The average between the PCR products *chitin synthase 1* gene is higher than to 92% (**Supplemental data 1**).

In order to identify the gDNA flanked regions surrounding the 320 bp of *Hvumar1* identified in cultivar *barke*, PCR/iPCR amplifications were performed with specific primers defined in the contig 289869 (**Figure 3B** and **Table 1**). For iPCR, *Hae II* and *Sma I* were used, since there is no internal cutting-site within the truncated element, to digest the gDNA of several cultivars of barley (Roho, Rihane, HD29, Martin and Bowman). After cloning and sequencing of PCR and iPCR products, the results showed either retrotransposons or uncharacterized sequences.

In parallel, to identify members of the *irritans* subfamily in plant species, PCR amplifications with degenerated TIRs and *Tria3D* primers designated from the same contig were done (**Figure 3B** and **Table 1**). Several products were obtained (*e.g.* Barley cultivar *Roho* **Figure 1A2**), cloned and sequenced. BLASTn analysis revealed only two deleted sequences, obtained with TIR-G1 and Triad3D primers in *H. vulgare* cultivar *Roho* and corresponding to *mariner* family. These sequences of 267 bp and 646 bp were named *Hvumar1* and *Hvumar2*, respectively.

To classify *Hvumar1* and *Hvumar2* with MITEs previously identified in aphids (*Aphidmar, Samar2*) and other *irritans* elements from several aphids species (unpublished data), known nucleotide sequences of *Tc1-mariner-IS630* superfamily were extracted from Genbank and compared through a UPGM-VM method of classification allowing a comparison between all these sequences whatever their length.

Results revealed that all identified elements in this study were clustered in a monophyletic group of *irritans* subfamily namely *macrosiphinimar* (**Figure 4** and **Supplemental data 2**). Specifically, *Hvumar1* (*H. vulgare*) determined both *in silico* and *in vitro* as well as *Aphidmar*, were closely related to *MITE1.1sub1* (fragment of about 950 bp, found in aphids from an *in silico* investigation, which share the same putative breakpoint with these latter elements) and *Apismar1.1* identified in *A. pisum* genome, whereas, *Samar2* (*S. avenae*) and *Hvumar2* (*H. vulgare*) were more connected to *Dnomar1.1* found in *D. noxia*. Interestingly, *Hvumar2* differs from *Samar2* by a longer internal deletion sequence of 7 bp near the 3' breaking point (**Figure 1B2** and **Figure 2**). Compared to the sequence detected in the contig289869 of the cultivar *barke* of *H. vulgare*, *Hvumar1* of the cultivar *Roho*, has four substitutions and smaller size (267 bp) due to the position of the primers (**Figure 1B2** and **Figure 2**). The similarity of these sequences with that of *Apismar1.1* (from *A. pisum*) suggest a possible horizontal transfer between aphids and barley.

Discussion

Previously, three groups of *irritans* subfamily, including ten lineages of full-length elements and five sublineages of MITEs, were detected in the genome of three aphid species belonging to Macrosiphini tribe *i.e Acyrthosiphon pisum*, *Myzus persicae* and *Diuraphis noxia* (unpublished data). This *in silico* analysis allowed to designate degenerated TIR primers, in order to identify *irritans* subfamily elements in four species of cereal aphids namely *Rhopalosiphum padi*, *R. maidis*, *Sitobion avenae* and *Schizaphis graminum*, whose genomes are still unsequenced.

PCR amplification results reveal two inactive types of elements (*Aphidmar* and *Samar2*) obtained by *TIR-G1* primers, which are determined from potentially active elements belonging to *macrosiphinimar* clade. In fact, internal sequences were highly similar to those of *Apismar1.1*, *Dnomar1.1* as well as *Mpmar1* (unpublished data), indicating that internal deletions from an autonomous partner appear to be a major process in the generation of short elements.

The comparison of *Aphidmar* sequences has shown high degree of identity (up to 98%) among cereal aphids and *A. pisum irritans* copies. This suggests (i) an early invasion of complete copies into the host genome, followed by stochastic loss in the common ancestor of aphid, then a vertical transmission, at around the same time, during the process of speciation, or (ii) a horizontal transfer of deleted copies, either they are transmitted at the same time as autonomous copies or through endogenous active elements, found in donor and receptor species (Brunet *et al.* 1999; Le Rouzic *et al.* 2007), or (iii) there are hotspots of deletions within ETs, thus, deletions observed in cereal aphid species are not homologous but rather correspond to convergences. Noteworthy, several

inactive *mauritiana* elements have been detected in seven tree aphid species, exhibiting the same high rate of identity (Kharrat *et al.* 2015).

Moreover, due to the absence of amplification and/or detection of active copy in these species, *Aphidmar* and *Samar2* are considered as "orphan", suggesting that their ancestral autonomous elements may have been lost after the MITEs amplification but it cannot be excluded that they still exist and maintain in other populations.

By comparative genomics of several species belonging to *Drosophila* genus, Macrosiphini tribe and *Rhodnius prolixus*, the majority of MITEs belonging to *irritans* subfamily, *lato sensu mariner* family, have a size about 950 bp, whereas few of them present less than 900 bp (Wallau *et al.* 2014; Filée *et al.* 2015 and personal unpublished data). If shorter MITEs are obtained (like *Samar2*), they are obviously in very low copy number or in few species, so not quite appropriate to amplification. These observations suggest that *mariner* elements are prone to internal deletion but the ability to transpose is likely highly constrained, by a minimum size about 900 bp and/ or by unable structure (large deletion on the 3' side), preventing the amplification (Brunet *et al.* 1996; Lohe and Hartl 1996). In addition, the comparison between *in silico* and *in vitro* results from *A. pisum*, used as reference, reveals a decrease in the diversity of the *irritans* elements. This can be explained by the exploitation, during these works, of different populations, which are probably polymorphic in terms of presence/absence of elements. The same case was also described by Barrón *et al.* (2014) who deduced that prevalence, diversity, distribution and localization of elements within *Drosophila melanogaster* can greatly vary between populations.

To test the possibility that plant host can include *irritans* elements similar to those identified in aphids (MITEs and or complete elements), a first screening had concerned *in silico* research, by a similarity approach, in several species of cereal plants. Only one contig found in the genome of the cultivar *barke* of *H. vulgare* contains a truncated element stuck on either side by aphid gDNA followed by gDNA of the host plant. Several recombination steps could be at the origin of this specific DNA organization after an earlier invasion of complete element into the aphid genome, followed by stochastic loss and mutations.

A second screening used TIR primers and other ones, designated in accordance with the truncated element obtained, in order to check its presence in the various cereal. Two types of sequences were characterized in barley cultivar *Roho*. The same truncated sequence was found and interestingly, a deleted element closely near to *Samar2* (lacking 7 bp) was identified. Their presence in barley, a species phylogenetically distant from brachypodium, oat and the common ancestor of aegilops and wheat, might be an initial evidence of an ancient HT between cereal aphids and barley.

A last screening aimed to find flanking regions as the specific organization identified *in silico*, nevertheless it did not reveal any shared copies. This absence of amplification could be due to a lack of conservation in the regions flanking the truncated element between cultivars, since a different cultivar has been used.. In all cases, it would be necessary to target the flanking regions of the two element types obtained *in vitro*, to support that HT has occurred.

Sequence similarity, tree incongruence, or patchy distributions are the main evidences used for detecting HT, but what are the mechanisms of the HT and the direction of exchange between plants and aphids? These questions remain opened. However, it must be stressed that host-parasite relationships facilitate such transfers (Gilbert *et al.* 2010), and probably general analysis of the holobionts of the species will be useful to answer such questions.

The cereal aphid species, here investigated, can infest only plant family belonging to cereals, they are oligophagous species. As long as these species feed on the same plants, several vectors of horizontal transfer can be considered like viruses, bacteria, or parasites. Indeed, viruses are able to inject DNA/RNA into host cells (Jehle *et al.* 1995; Routh *et al.* 2012) and specifically two TEs inserted in the genome of an iridovirus infecting dipteran were identified (Piégu *et al.* 2013). Furthermore, in crustaceans (terrestrial isopods), endosymbiotic *Wolbachia* strains are known to be frequently horizontally transferred (Cordaux *et al.* 2001). In addition, several eukaryotic genes were integrated in *Wolbachia* genomes (Duplouy *et al.* 2013), suggesting that this bacteria is an ideal vector/shuttle of HT.

To conclude, the present study provides the first *in vitro* characterization of *irritans* subfamily in several cereal aphids and interesting clues of the existence of a HT of a transposable element between barley and aphids. These results could be relevant for biotechnological applications to genetically control some aphid species and to test potential vectors of horizontal transfer, in future studies.

Authors' contributions: MB, MM and PC conceived and designed research. MB performed research. WBLA contributed in *in vitro* research. MB, HM collected samples. MMK, HM, AHV helped in the analysis of data. MB, MM and PC drafted the manuscript.

Acknowledgements: This work was financially supported by the Tunisian Ministry of Higher Education and Scientific Research, the University of Tunis El Manar, the Centre National de la Recherche Scientifique and the Paris-Sud University. We thank Yosra Habachi for assistance to collect several plant cultivars and the INRAT (Institut National de Recherche Agronomique de Tunis) for the supply of *Acyrthosiphon pisum* samples.

References

Auge-Gouillou, C., Bigot, Y., Pollet, N., Hamelin, M. H., Meunier-Rotival, M., & Periquet, G. (1995). Human and other mammalian genomes contain transposons of the mariner family. FEBS letters, 368(3), 541-546.

Bannai, H., Tamada, Y., Maruyama, O., Nakai, K., & Miyano, S. (2002). Extensive feature detection of N-terminal protein sorting signals. Bioinformatics, 18(2), 298-305.

Barrón, M. G., Fiston-Lavier, A. S., Petrov, D. A., & González, J. (2014). Population genomics of transposable elements in Drosophila. Annual review of genetics, 48, 561-581.

Bennetzen, J. L., & Wang, H. (2014). The contributions of transposable elements to the structure, function, and evolution of plant genomes. Annual review of plant biology, 65, 505-530.

Biémont, C., & Vieira, C. (2006). Genetics: junk DNA as an evolutionary force. Nature, 443(7111), 521-524.

Blackman, R. L., & Eastop, V. F. (2000). Aphids on the world's crops: an identification and information guide. 2nd edn. New York: Willey Ltd., Chichester.

Brillet B., Bigot Y., Augé Gouillou C. (2007). Assembly of the Tc1 and mariner transposition initiation complexes depends on the origins of their transposase DNA binding domains. Genetica, 130(2), 105-120.

Brunet, F., Giraud, T., Godin, F., & Capy, P. (2002). Do deletions of Mos1-like elements occur randomly in the Drosophilidae family?. Journal of molecular evolution, 54(2), 227-234.

Brunet, F., Godin, F., Bazin, C., & Capy, P. (1999). Phylogenetic analysis of Mos1-like transposable elements in the Drosophilidae. Journal of molecular evolution, 49(6), 760-768.

Brunet, F., Godin, F., Bazin, C., David, J. R., & Capy, P. (1996). The mariner transposable element in natural populations of Drosophila teissieri. Journal of molecular evolution, 42(6), 669-675.

Bureau, T. E., & Wessler, S. R. (1992). Tourist: a large family of small inverted repeat elements frequently associated with maize genes. The Plant Cell, 4(10), 1283-1294.

Bureau, T. E., & Wessler, S. R. (1994). Stowaway: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. The Plant Cell, 6(6), 907-916.

Cordaux, R., Michel-Salzat, A., & Bouchon, D. (2001). Wolbachia infection in crustaceans: novel hosts and potential routes for horizontal transmission. Journal of Evolutionary Biology, 14(2), 237-243.

Doyle, J. J. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochemical bulletin, 19, 11-15.

Dupeyron, M., Leclercq, S., Cerveau, N., Bouchon, D., & Gilbert, C. (2014). Horizontal transfer of transposons between and within crustaceans and insects. Mobile DNA, 5(1), 1.

Duplouy, A., Iturbe-Ormaetxe, I., Beatson, S. A., Szubert, J. M., Brownlie, J. C., McMeniman, C. J., ... & Woolfit, M. (2013). Draft genome sequence of the male-killing Wolbachia strain w Bol1 reveals recent horizontal gene transfers from diverse sources. BMC genomics, 14(1), 1.

El Baidouri, M., Carpentier, M. C., Cooke, R., Gao, D., Lasserre, E., Llauro, C. *et al.* (2014). Widespread and frequent horizontal transfers of transposable elements in plants. Genome research, 24(5), 831-838.

Feschotte, C., & Mouches, C. (2000). Evidence that a family of miniature inverted-repeat transposable elements (MITEs) from the Arabidopsis thaliana genome has arisen from a pogo-like DNA transposon. Molecular Biology and Evolution, 17(5), 730-737.

Feschotte, C., & Pritham, E. J. (2007). DNA transposons and the evolution of eukaryotic genomes. Annual review of genetics, 41, 331.

Filée, J., Rouault, J. D., Harry, M., & Hua-Van, A. (2015). Mariner transposons are sailing in the genome of the blood-sucking bug Rhodnius prolixus. BMC genomics, 16(1), 1.

Gao Y, Mathee K, Narasimhan G, Wang X. Motif detection in protein sequences. String Processing and Information Retrieval Symposium, 1999 and International Workshop on Groupware. doi: 10.1109/SPIRE.1999.796579.

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(7293), 1347-1350.

Hartl, D. L., Lozovskaya, E. R., Nurminsky, D. I., & Lohe, A. R. (1997). What restricts the activity of mariner-like transposable elements?. *Trends in genetics*, *13*(5), 197-201.

Jehle, J. A., Fritsch, E., Nickel, A., Huber, J., & Backhaus, H. (1995). TCl4. 7: a novel lepidopteran transposon found in Cydia pomonella granulosis virus. Virology, 207(2), 369-379.

Kharrat, I., Mezghani, M., Casse, N., Denis, F., Caruso, A., Makni, H *et al.* (2015). Characterization of mariner-like transposons of the mauritiana Subfamily in seven tree aphid species. Genetica, 143(1), 63-72.

Larsson, A. (2014). AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics, 30(22), 3276-3278.

Le Rouzic, A., Boutin, T. S., & Capy, P. (2007). Long-term evolution of transposable elements. Proceedings of the National Academy of Sciences U S A, 104(49), 19375-19380.

Lohe, A. R., & Hartl, D. L. (1996). Germline transformation of Drosophila virilis with the transposable element mariner. Genetics, 143(1), 365-374.

Lohe, A. R., Moriyama, E. N., Lidholm, D. A., & Hartl, D. L. (1995). Horizontal transmission, vertical inactivation, and stochastic loss of mariner-like transposable elements. Molecular biology and evolution, 12(1), 62-72.

Loreto, E. L. S., Carareto, C. M. A., & Capy, P. (2008). Revisiting horizontal transfer of transposable elements in Drosophila. *Heredity*, *100*(6), 545-554.

Lu, C., Chen, J., Zhang, Y., Hu, Q., Su, W., & Kuang, H. (2012). Miniature inverted–repeat transposable elements (MITEs) have been accumulated through amplification bursts and play important roles in gene expression and species diversity in Oryza sativa. Molecular biology and evolution, 29(3), 1005-1017.

Miller, R. H., & Pike, K. S. (2002). Insects in wheat-based systems. In: Curtis BC, Rajaram S, Go´mez Macpherson H (eds) Bread wheat: improvement and production, plant production and protection series no. 30, FAO, Rome, pp 367–393.

Muñoz-López, M., & García-Pérez, J. L. (2010). DNA transposons: nature and applications in genomics. Current genomics, 11(2), 115-128.

Narasimhan, G., Bu, C., Gao, Y., Wang, X., Xu, N., & Mathee, K. (2002). Mining protein sequences for motifs. Journal of Computational Biology, 9(5), 707-720.

Negoua, A., Rouault, J. D., Chakir, M., & Capy, P. (2013). Internal deletions of transposable elements: the case of Lemi elements. Genetica, 141(7-9), 369-379.

Nicholas, K. B., Nicholas, H. B. J., & Deerfield, D. W. (1997). GeneDoc: analysis and visualization of genetic variation. Embnew. news, 4(1).

Ochman, H., Gerber, A. S., & Hartl, D. L. (1988). Genetic applications of an inverse polymerase chain reaction. Genetics, 120(3), 621-623.

Piégu, B., Guizard, S., Spears, T., Cruaud, C., Couloux, A., Bideshi, D. K *et al.* (2013). Complete genome sequence of invertebrate iridescent virus 22 isolated from a blackfly larva. Journal of General Virology, 94(9), 2112-2116.

Robertson, H. M. (1993). The mariner transposable element is widespread in insects. Nature, 362(6417), 241-245.

Robertson, H. M., & Lampe, D. J. (1995). Recent horizontal transfer of a mariner transposable element among and between Diptera and Neuroptera. Molecular Biology and Evolution, 12(5), 850-862.

Robertson, H. M., & MacLeod, E. G. (1993). Five major subfamilies of mariner transposable elements in insects, including the Mediterranean fruit fly, and related arthropods. Insect molecular biology, 2(3), 125-139.

Robillard, E., Le Rouzic, A., Zhang, Z., Capy, P., & Hua-Van, A. (2016). Experimental evolution reveals hyperparasitic interactions among transposable elements. *Proceedings of the National Academy of Sciences, 113*(51), 14763-14768.

Rouault, J. D., Casse, N., Chénais, B., Hua-Van, A., Filée, J., & Capy, P. (2009). Automatic classification within families of transposable elements: application to the mariner Family. Gene, 448(2), 227-232.

Routh, A., Domitrovic, T., & Johnson, J. E. (2012). Host RNAs, including transposons, are encapsidated by a eukaryotic single-stranded RNA virus. Proceedings of the National Academy of Sciences U S A, 109(6), 1907-1912.

Schaack, S., Gilbert, C., & Feschotte, C. (2010). Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. Trends in ecology & evolution, 25(9), 537-546.

Shao, H., & Tu, Z. (2001). Expanding the diversity of the IS630-Tc1-mariner superfamily: discovery of a unique DD37E transposon and reclassification of the DD37D and DD39D transposons. Genetics, 159(3), 1103-1115.

Silva, J. C., Bastida, F., Bidwell, S. L., Johnson, P. J., & Carlton, J. M. (2005). A potentially functional mariner transposable element in the protist Trichomonas vaginalis. Molecular biology and evolution, 22(1), 126-134.

Vea, I. M., & Grimaldi, D. A. (2016). Putting scales into evolutionary time: the divergence of major scale insect lineages (Hemiptera) predates the radiation of modern angiosperm hosts. Scientific reports, 6.

Wallau, G. L., Capy, P., Loreto, E., & Hua-Van, A. (2014). Genomic landscape and evolutionary dynamics of mariner transposable elements within the Drosophila genus. BMC genomics, 15(1), 1.

Wallau, G. L., Capy, P., Loreto, E., Le Rouzic, A., & Hua-Van, A. (2016). VHICA, a new method to discriminate between vertical and horizontal transposon transfer: application to the mariner family within Drosophila. Molecular biology and evolution, 33(4), 1094-1109.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B *et al.* (2007). A unified classification system for eukaryotic transposable elements. Nature Reviews Genetics, 8(12), 973-982.

Yamada, K., Yamada, A., Kawanishi, Y., Deep Gurung, R., Tokuda, G., & Maekawa, H. (2015). Widespread distribution and evolutionary patterns of mariner-like elements among various spiders and insects. Journal of Insect Biotechnology and Sericology, 84(2), 2029-2041.

Legends of figures, tables and supplemental data

Figure 1. PCR amplification and schematic representation of *in vitro* aphid and barley *Roho* cultivar elements.

A1. Amplification products obtained by PCR using degenerated TIR primers of *irritans subfamily* designated from the lineage of *Apismar1.1* (unpublished data). PCR amplifications were performed using genomic DNA from four cereal aphids with *A. pisum* as positive control. W = molecular weight marker 100 bp (Invitrogen). PCR products of about 1000 bp are present in the five species and a faint one about 600 bp is only detected in *S. avenae*.

A2. PCR Amplification with *chitin synthase 1 gene* primers: (1) T+ corresponds to DNA of *Rhopalosiphum padi* as positive control, (2) using genomic DNA from cereal plants. Amplification products obtained from genomic DNA of *Roho* cultivar by PCR/iPCR (3-9) using primers designated after *in silico* investigation in *Hordeum vulgare* genome (as mentioned in **Table 1** and **Figure 2**). W = molecular weight marker 100 bp (Invitrogen**)**.

B1. Schematic representation of the *Aphidmar* elements including *Apmar1*, *Sgmar1, Samar1*, *Rpmar1*, *Rmmar1* and *Samar2* compared to the full-length copy *Apismar1.1* used here a reference. Asterisk refers to a potentially active element. Blue arrows indicate TIRs, while bold lines represent UTRs. Internal deletions are mentioned and correspond to thin lines. The DNA binding domain including the helix turn helix (HTH) region, the related canonical WVPKQL motif, the three catalytic core and residues (in red), as well as the nuclear localization signal (NLS) are indicated. The turned T represents the polyAdenylation site AATAAA. Triangles refer to deletion.

B2. Schematic representation of *Hvumar1* and *Hvumar2* obtained by *TIR-G1*/ *Tria3D* primers compared to *Apismar1.1* as well as aphid elements detected from the *in vitro* investigation.

Figure 2. Alignment of nucleotide sequences.

This aligment contains the *Aphidmar* copies in brown (*Apmar1, Rpmar1, Rmmar1, Sgmar1* and *Samar1*), *Samar2*, *Hvumar* (*Hvumar1, Hvumar2, and Hvumar1 in silico*) and *Apismar1.1* extracted from the genome of *A. pisum.* Synapomorphies per species, internal deletions and Short Direct Repeats (SDR) microhomologies observed in the region of the breaking points (BP) of deletions were highlighted. Denomination of the SDR position in the region of BP, is that of Negoua *et al*. (2013). BPNN for breaking points near-near is observed in *Aphidmar* represented by blue boxes, whereas BPEE for breaking points exact-exact is located in *Samar2* and *Hvumar2* mentioned in yellow color. The *5'-TIR* and *3'-TIR* are in gray boxes. The dinucleotide TA flanking these elements corresponds to the target site of duplication (TSD).

Figure 3. Schematic representation of *in silico* plant element.

A. Structural organization of the contig 289869 extracted from *Hordeum vulgare* cultivar barke genome. It includes the truncated *irritans* element (320 bp) corresponding to the 3' end with the third catalytic motif and TIR. This element is compared to the full length *Apismar1.1* extracted *in silico* and *Aphidmar* identified *in vitro*. Blue arrows indicate TIRs, while bold lines represent UTRs. Internal deletions correspond to thin lines. The green and brown lines correspond to genomic DNA of the barley and genomic DNA of the *A. pisum*, respectively. The DNA binding domain including the helix turn helix (HTH) region, the related canonical WVPKQL motif, the three catalytic core and residues (in red), and nuclear localization signal (NLS) are indicated. The turned T represents the polyAdenylation site AATAAA.

B. Complete sequence of the contig 289869 extracted from *H. vulgare* cultivar barke genome. This sequence is 1471 bp long. The 3'TIR is in gray box. The third aspartic residue is mentioned in red. Arrows or lines represent different primers.

Figure 4. Classification of sequences identified in aphids and barley.

Classification of six sequences obtained *in vitro* from aphids including *Aphidmar* and *Samar2*, with three sequences identified in barley (*Hvumar1-2*), along with 184 elements belonging to the *Tc1-mariner-IS630* superfamily is based on the UPGM-VM method (Rouault *et al.* 2009). References and positions of all these sequences are given in **Supplemental data 3** according to the reading sense indicated by the arrow in the circular tree. Complete sequences and MITEs are marked by a full black circle and an empty one, respectively.

Table 1. Primers used for PCR and iPCR with hybridization temperature and products expected sizes. Asterisk corresponds to the expected size of PCR products with *Tria3D* and *TIR-G1* primers.

Table 2. Characteristics of the plant genome explored in this work.

Supplemental data 1. Chitin synthase I gene sequences from the cereal aphids

Supplemental data 2. Sequences (n=193) classified by UPGM-VM method according to the reading sense indicated by the arrow in the circular tree. Deleted or truncated sequences are indicated by an asterisk (*).

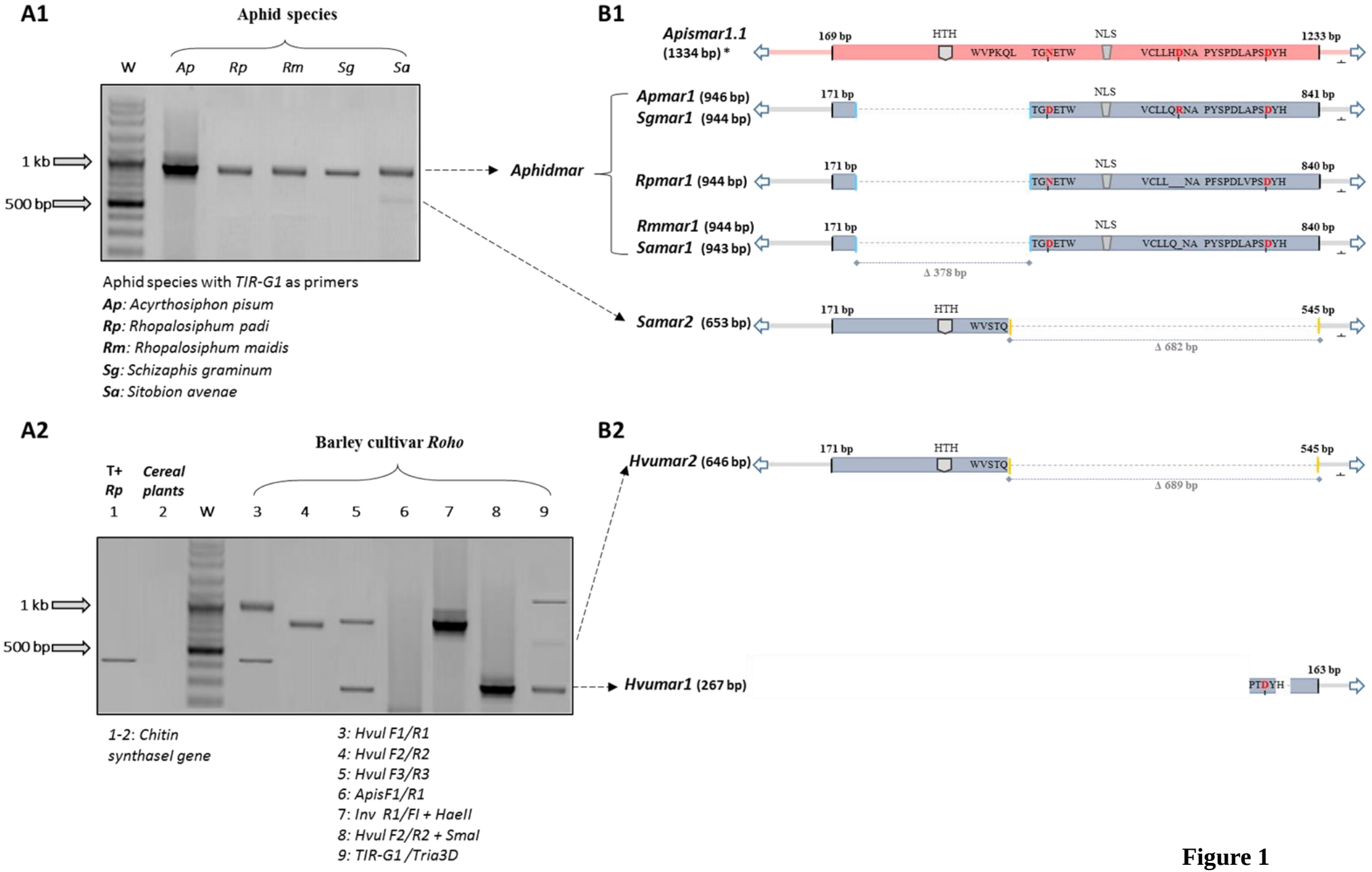

A1
Aphid species
W
Ap
Rp
Rm
Sg
Sa
1 kb
500 bp
Aphidmar
Aphid species with TIR-G1 as primers
Ap: Acyrthosiphon pisum
Rp: Rhopalosiphum padi
Rm: Rhopalosiphum maidis
Sg: Schizaphis graminum
Sa: Sitobion avenae
B1
Apismar1.1
(1334 bp) *
169 bp
HTH
NLS
1233 bp
WVPKQL
TGNETW
VCLLHDNA PYSPDLAPSDYH
Apmar1 (946 bp)
Sgmar1 (944 bp)
171 bp
841 bp
TGDETW
VCLLQRNA PYSPDLAPSDYH
Rpmar1 (944 bp)
840 bp
VCLL___NA PFSPDLVPSDYH
Rmmar1 (944 bp)
Samar1 (943 bp)
VCLLQ_NA PYSPDLAPSDYH
Δ 378 bp
Samar2 (653 bp)
545 bp
WVSTQ
Δ 682 bp
A2
Barley cultivar Roho
T+ Rp
Cereal plants
1
2
W
3
4
5
6
7
8
9
1 kb
500 bp
1-2: Chitin synthasel gene
3: Hvul F1/R1
4: Hvul F2/R2
5: Hvul F3/R3
6: ApisF1/R1
7: Inv R1/FI + HaeII
8: Hvul F2/R2 + SmaI
9: TIR-G1 /Tria3D
B2
Hvumar2 (646 bp)
Δ 689 bp
Hvumar1 (267 bp)
163 bp
PTDYH

Figure 1

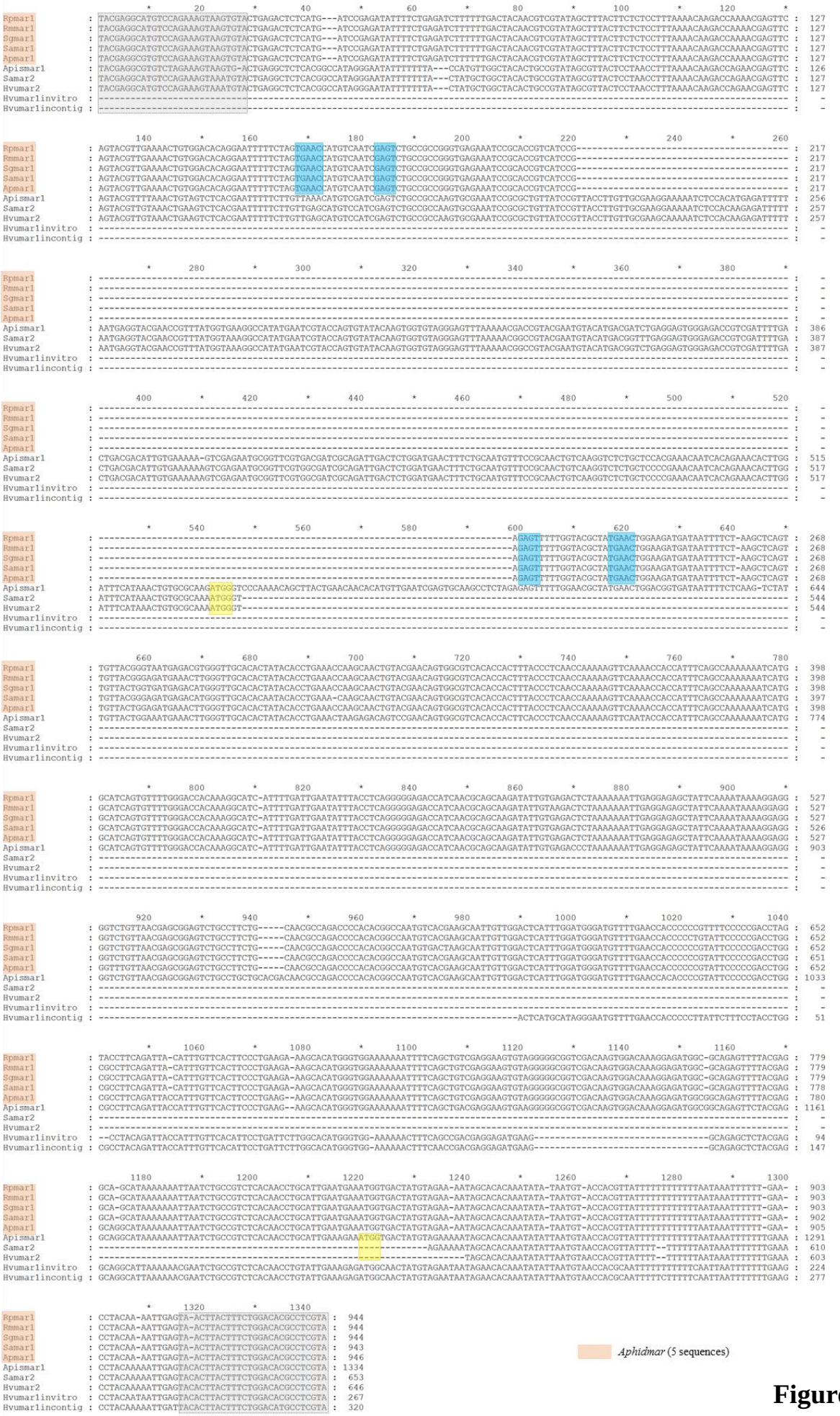
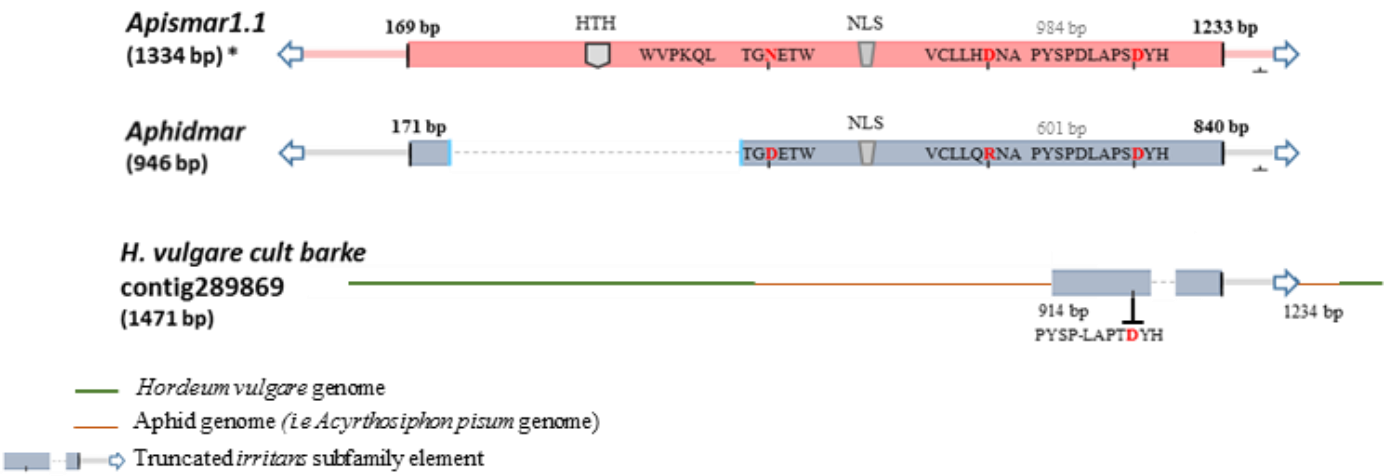


Figure 2

A



B

>CAJV010199892.1| *Hordeum vulgare subsp. vulgare* cultivar Barke WGS, contig289869 (strand-), **1471 bp**

ATATATACTCGTTTGGTGCTGATTGGTTGAAATTCCTTGGTATGAATCATCGCATCGTTGCTTAAATTTTGTAACTTTCTAAGTGCGTTTTCCCTCGTCATACTTTTGAAATTTAATTGTATA

AAATCAGTTTTTATGGTTGCCTTTTCCCCGCCATATGTACGCCACTCACTTGTAAAATTAGATTATTTATGGTCTTGATAAAGACCGCTTTCTCAGTATTTGTTCCGTAACACAGGAATTTAA

Hvul-F3

TATACTAAGAAGGTAATAGGTAATACCAAGAATACCAGAAACCATTCTTAAAGTTGTACAGCTTTTTCTGCTACAAAATATAATGACAGACACAAATATTTGTACACACATAATAAATTTTAA

Hvul-F2

AATTAATATTATATTCATTATAATTCATGAGAATCGAAAATCATTGTATAACACAATATTGAACGTTATTATTACTTAATTGGCTAGCTAGTAGATCATGACGTGTTCACTATAGCAATATTG

Hvul-F1

TTTAAAACTGAAATTAACTACAATATTTTTTTATAATAAATATAGTTTTTAACTATAACATACAATTTTAATTTTTTTATAGATCAATGCTAGATATGATTAAGAATATTTTAAAAGCAAAAA

TATTCATGGTATGGATGATATTCTTCACATTAATATATTGGATTTACTCATTTTATTCTTTGGCTGCATTTGAAGGGCTTCTTAGTAAACCACGTTCACTCTTCTTTGACCACGGTAAGTGTT

Apis-F1

AAAATACAATTCAATTTATATTAGATAAGTATACATATAATATGTGTGACACATTAATACGAAATTAAAGAAAATTACAGTTTATCTCTTTTATTATTTGACAGTATTAGTTTCATACTGATA

AATTAAGTATATTATTTAAATTCATGACATCAATATTGAAAATTCAAAAAGGTACTCATGCATAGGGAATGTTTTGAACCACCCCCTTATTCTTTCCTACCTGGCGCCTACAGATTACCATTT

Tria3D

GTTCACATTCCTGATTCTTGGCACATGGGTGGAAAAAACTTTCAACCGACGAGGAGATGAAGGCAGAGCTCTACGAGGCAGGCATTAAAAAACGAATCTGCCGTCTCACAACCTGTATTGAAA

*Inv*R1

GAGATGGCAACTATGTAGAATAATAGAACACAAATATATTAATGTAACCACGCAATTTTTCTTTTTCAATTAATTTTTTTGAAGCCTACAAAAATTGATTACACTTACTTTCTGGACATGCCT

*Inv*F1

TIR-G1

CGTATGTTTTAAATAATTAACAGAAAATTATAAATATTTTCGTTCACTTTAAATTCTAACTTAAATAACTAAACAGTCTGCACTTATTATCACATATTATTATGAATTAATTTAAAATGTTGA

Apis-R1

Hvul-R3

AACATTAAAATTACTATATATTTTTTTTTGACGCTTTTTAAACTTACCTATAAGTAGAGACCAGAGTTATATTCATTACAGAAATGCTAATTATGATGCAAATATGCATGATCAAAAG

Hvul-R1

Hvul-R2

Figure 3

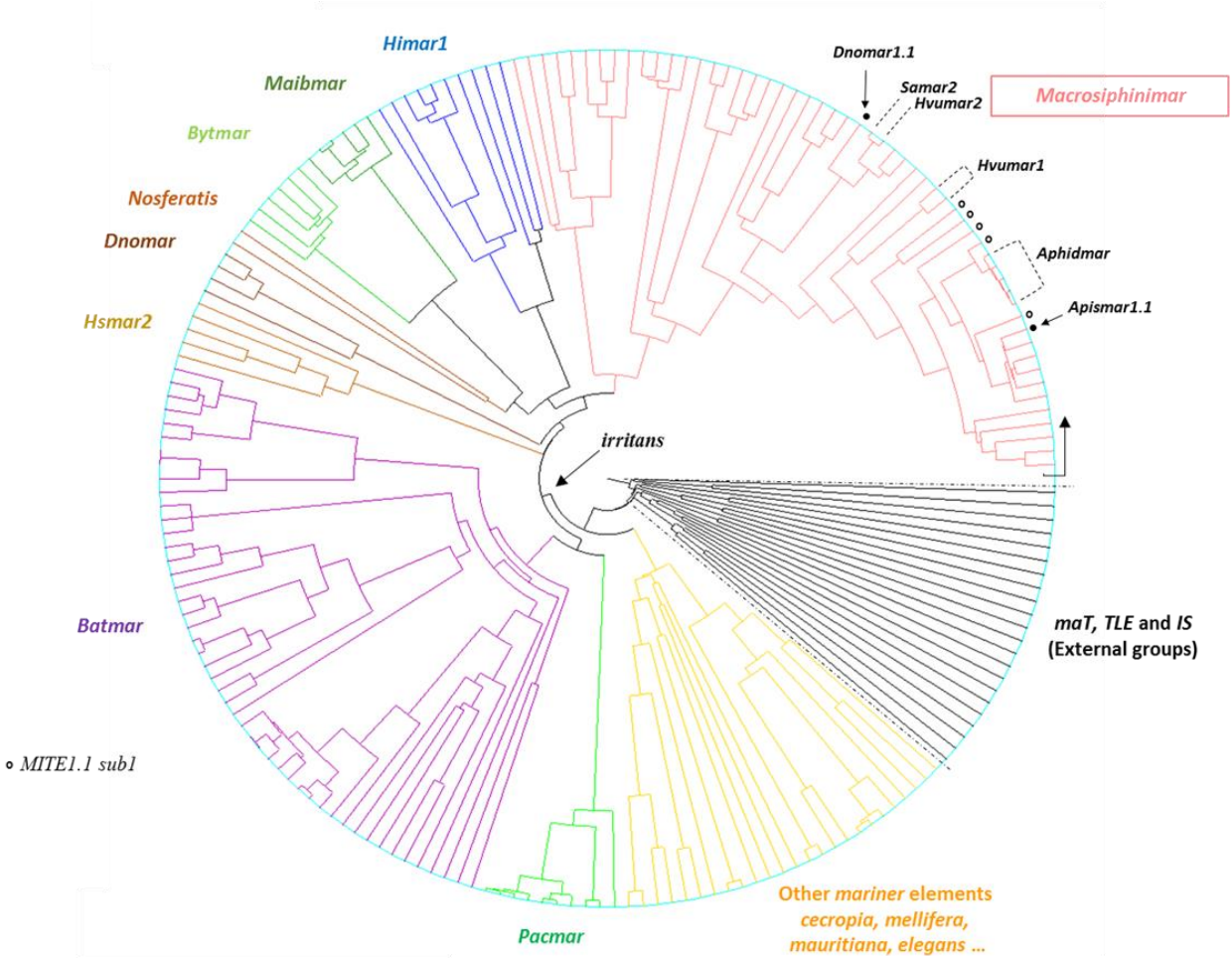

Himar1
Maibmar
Bytmar
Nosferatis
Dnomar
Hsmar2
Batmar
MITE1.1 sub1
Pacmar
Dnomar1.1
Samar2
Hvumar2
Macrosiphinimar
Hvumar1
Aphidmar
Apismar1.1
irritans
maT, TLE and IS
(External groups)
Other mariner elements
cecropia, mellifera,
mauritiana, elegans ...

Figure 4

Table 1. Primers used for PCR and iPCR with hybridization temperature and products expected sizes.
Asterisk corresponds to the expected size of PCR products with Tria3D and *TIR-G1* primers.

Applications	Primer names	Nucleotide sequences	T°	Expected size (bp)
irritans subfamily	*TIR-G1*	TACGARGCDTGYCYARAAART	60	920<size<1350
	TIR-G2	TATTCRAAAMKTAAGGRYYGT	54	~1300
	TIR-G3	TAATWAATTTTTCAWAATTWGG	58	920<size<1130
	TIR-G4	TACGAKGTRTGACAATAAAAT	54	~900, ~1320
	TIR-G5	TAAAYACCCAGACAAMAWRTAT	54	~1300
	TIR-G6	TACGAGRGCGGGCTGATAAG	62	~1360
	Tria3D	CCTACAGATTACCATTTGTTC	60	267*
iPCR	*InvR1*	GTGCCAAGAATCAGGAATG	52	>700
	InvF1	CCTGTATTGAAAGAGATGGC		
Chitin synthaseI gene	*F1*	CAGGTYATGTACATGTAYTAC	49	410
	R1	TGATCRTACTGCACGTAGTG		
Genomic DNA of *Hordeum vulgare* (Barke variety)	*Hvul F1*	GCTAGCTAGTAGATCATGAC	52	960
	Hvul R1	CTCTGGTCTCTACTTATAGG		
	Hvul F2	GCTACAAAATATAATGACAGAC	50	1130
	Hvul R2	GCATCATAATTAGCATTTCTG		
	Hvul F3	GACAGACACAAATATTTGTACAC	60	1140
	Hvul R3	CTTTTGATCATGCATATTTGCATC		
Genomic DNA of *Acyrthosiphon pisum*	*Apis F1*	GGTATGGATGATATTCTTCAC	56	700
	Apis R1	GTGATAATAAGTGCAGACTG		

R= A/G, **Y**= T/C, **D**= A/T/G, **W**= A/T, **M**= A/C, **K**= T/G

Table 2. Characteristics of the plant genomes explored in this work.

Plants	Cultivar	WGS accession	Project	Level and number	Size (Gb)
Brachypodium distachyon	Bd21	PRJNA32607	ADDN02	Scaffolds = 28	0.27
Aegilops tauschii	AL8/78	PRJNA182898	AOCO01	Scaffolds = 429,891	3.13
Triticum aestivum	Chinese Spring	PRJEB11773	FAOM01	Scaffolds = 735,943	13
Hordeum vulgare	Barke	PRJEB84	CAJV01	Contigs = 2,742,077	2.06
	Morex	PRJEB86	CAJW01	Scaffolds = 2,670,738	1.86
	Bowman	PRJEB88	CAJX01	Scaffolds = 2,077,901	1.78

Supplemental data 1. Chitin synthase I gene sequences from the cereal aphids

>Sequence of chitin synthase1 gene from *R. padi*

CAGGTCATGTACATGTACTACTTACTTGGTCATCGATTAATGGAACTACCGATTTCCGTTGAACGGAAAG
AAGTCATTGCTGAAAATACGTTTTTATTGACGCTTGATGGTGATATTGATTTCCAGCCACATGCGGTCAG
GCTTTTGATAGATTTGATGAAAAAAAATAAAAATTTGGGAGCCGCTTGTGGTAGAATCCATCCAATTGGA
GGAGGTCCATTGGCGTGGTATCAAGTTTTTGAATACGCCATTGGTCATTGGCTCCAAAAAGCTACTGAAC
ACATGATTGGTTGCGTTCTTTGTAGTCCTGGATGTTTCTCACTGTTCAGAGGTAAAGCTCTTATGGACGA
TAACGTGATGAAAAGATATACCACGCTGCCGGTTGAAGCCTTACATTATGTTCAATACGA

> Sequence of chitin synthase1 gene from *R. maidis*

CAGGTCATGTATATGTACTACTTACTTGGTCATCGATTAATGGAACTACCAATTTCCGTTGAACGTAAAG
AAGTCATTGCTGAGAATACGTTTTTGTTGACGCTTGATGGTGATATTGATTTCCAKCCACATGCGGTCAG
GCTTATGATAGATTTAATGAAAAAGAATAAAAATTTGGGAGCCGCTTGTGGTAGAATCCATCCAATTGGA
GGAGGTCCATTAGCGTGGTATCAAGTTTTTGAATATGCCATTGGTCATTGGCTCCAAAAAGCTACTGAAC
ACATGATCGGTTGCGTTCTTTGTAGTCCTGGATGTTTCTCGCTATTCAGAGGTAAAGCTCTTATGGACGA
TAACGTTATGAAAAGATATACCACGCTGCCGGTTGAAGCCTTACATTATGTTCAATACGA

>Sequence of chitin synthase1 gene from *S. graminum*

CAGGTCATGTACATGTACTACTTACTTGGTCATCGATTAATGGAACTACCGATTTCCGTTGAACGGAAAG
AAGTCATTGCTGAAAATACGTTTTTATTGACGCTTGATGGTGATATTGATTTCCAGCCACATGCGGTCAG
GCTTTTGATAGATTTGATGAAAACAAATAAAAATTTGGGAGCCGCTTGTGGTAGAATCCATCCAATTGGA
GGAGGTCCATTGGCGTGGTATCAAGTTTTTGAATACGCCATTGGTCATTGGCTCCAAAAAGCTACTGAAC
ACATGATTGGTTGCGTTCTTTGTAGTCCTGGATGTTTCTCACTGTTCAGAGGTAAAGCTCTTATGGACGA
TAACGTGATGACAAGATATACCACGCTGCCGGTTGAAGCCTTACATTATGTTCAATACGA

>Sequence of chitin synthase1 gene from *S. avenae*

CAGGTCATGTACATGTACTACTTACTTGGTCATCGATTAATGGAACTACCGATTTCCGTTGAACGGAAAG
AAGTCATTGCTGAAAATACGTTTTTATTGACGCTTGATGGTGATATTGATTTCCAGCCACATGCGGTCAG
GCTTTTGATAGATTTGATGAAAAAAAATAAAAATTTGGGAGCCGCTTGTGGTAGAATCCATCCAATTGGA
GGAGGTCCATTGGCGTGGTATCAAGTTTTTGAATACGCCATTGGTCATTGGCTCCAAAAAGCTACTGAAC
ACATGATTGGTTGCGTTCTTTGTAGTCCTGGATGTTTCTCACTGTTCAGAGGTAAAGCTCTTATGGACGA
TAACGTGATGAAAAGATATACCACGCTGCCGGTTGAAGCCTTACATTATGTTCAATACGA

Supplemental data 2. Sequences (n=193) classified by UPGM-VM method according to the reading sense indicated by the arrow in the circular tree. Deleted or truncated sequences are indicated by an asterisk (*).

		***Macrosiphinimar* DD34D**	
1	*Diuraphis noxia*	gi984744883:140935-140721 strand-	*
2		gi984744980:13238-14654 strand+	*
3	*Myzus persicae*	Scaffold6:1582921..1583927 strand+	MITE1.2
4		Scaffold425:148657-149615 strand +	
5	*Diuraphis noxia*	gi984745178:34523-35477 strand+	*
6		gi984745316:3270-2865 strand-	*
7		gi984745505:879619-879165 strand-	*
8	*Acyrthosiphon pisum*	gi320446981 NW003383590.1 Scaffold101:202225-201649 strand -	*
9		gi320388812 NW003403215.1 Scaffold19726:243-904 strand -	*
10	*Myzus persicae*	Scaffold938 :36419-37755 strand-	Mpmar1.1
11	*Acyrthosiphon pisum*	gi320446985 NW003383586.1 Scaffold97:172262-173595 strand +	Apismar1.1
12		gi320446984 NW003383587.1 Scaffold 98:39716-40668 strand +	MITE1.1 sub1
13	*Sitobion avenae*	*In vitro*	*Samar1*
14	*Rhopalosiphum maidis*	*In vitro*	*Rmmar1*
15	*Schizaphis graminum*	*In vitro*	*Sgmar1*
16	*Rhopalosiphum padi*	*In vitro*	*Rpmar1*
17	*Acyrthosiphon pisum*	*In vitro*	*Apmar1*
18		gi320447035 NW003383536.1 Scaffold47:642566-643516 strand +	MITE1.1 sub1
19		gi320446957 NW003383614.1 Scaffold125:833628-834589 strand -	
20		gi320446934 NW003383637.1 Scaffold 148:254344-255290 strand +	
21		gi320446899 NW003383672.1 Scaffold 183:345754-346918 strand +	
22	*Hordeum vulgare cultivar barke*	Contig 289869 strand-	*Hvumar1*
23	*H. vulgare cultivar Roho*	*In vitro*	
24	*Myzus persicae*	Scaffold50:754375-755343 strand+	*
25	*Diuraphis noxia*	gi984745255:324638-325555 strand+	*
26	*Myzus persicae*	Scaffold1103:23507-23906 strand+	*
27	*Diuraphis noxia*	gi984745202:272244-271837 strand-	*
28	*Hordeum vulgare cultivar Roho*	*In vitro*	*Hvumar2*
29	*Sitobion avenae*	*In vitro*	*Samar2*
30	*Diuraphis noxia*	gi984745488:452325-450979 strand-	Dnomar1.1
31	*Trachymyrmex cornetzi*	gi1006854094 LKEY01038001.1 Contig 38001:392-1605	
32	*Rhodnius prolixus*	Rpmar33 (Filée et *al.* 2015)	
33	*Vollenhovia emeryi*	gi763991541 BBUO01005411.1 Contig06950:10342-11676	
34	*Oncopeltus fasciatus*	gi641099433 JHQO01163363.1 ContigNC163363:6062-7386	
35	*Homalodisca vitripennis*	gi642862357 JJNS01041406.1 ContigNC41406:3310-4634	
36	*Myzus persicae*	Scaffold131:48139-48677 strand-	*
37	*Acyrthosiphon pisum*	gi320447046 NW003383525.1 Scaffold36 :908811-909602 strand -	*
38		gi320447034 NW003383537.1 Scaffold48:638002-638846 strand +	*
39		gi320446875 NW003383696.1 Scaffold207:537540-538487 strand -	*
40	*Diuraphis noxia*	gi984745157:267924-268872 strand+	*
41		gi984744390:37892-38851 strand+	*
42	*Acyrthosiphon pisum*	gi320447089 NW003383510.1 Scaffold21:987841-988816 strand +	*
43		gi320446282 NW003384289.1 Scaffold800:146539-147106 strand -	MITE1.1 sub2
44		gi320446572 NW003383999.1 Scaffold510:194743-195684 strand +	
45		gi320446653 NW003383918.1 Scaffold429:453472-454426 strand +	
46		gi320446326 NW003384245.1 Scaffold756:48163-49727 strand +	
47		gi320447005 NW003383566.1 Scaffold77:451686-452577 strand +	*
48		gi320442452 NW003388040.1 Scaffold4551:1640-2495 strand +	*
49		gi320446993 NW003383578.1 Scaffold89:934560-935876 strand +	Apismar1.2
50	*Diuraphis noxia*	gi984745245:252450-251151 strand-	Dnomar1.2
51	*Dendroctonus ponderosae*	gi459605371 APGL01021548.1 Seq01021608:13322-14422	
52	*Gerris buenoi*	gi822390376 JHBY01085489.1 Contig85496: 109-1314	
53	*Homalodisca vitripennis*	gi642470880 JJNS01248885.1 ContigNC248885:1757-2877	

54	*Mesobuthus martensii*	gi553813549 AYEL01086727.1 Contig347321:2028-3230	
55	*Anoplophora glabripennis*	gi496870061 AQHT01063015.1 Contig63076:4332-5534	
		***Himar-like elements* DD34D**	
56	*Aphis glycines*	GQ231493	
57	*Drosophila yakuba*	Dromar18 (Wallau et *al.* 2014)	
58	*Rhodnius prolixus*	Rpmar0 (Filée et *al.* 2015)	
59	*Bactrocera dorsalis*	AF346541	
60	*Diachasmimorpha longicaudata*	AY601748	
61	*Chrysoperla plorabunda*	L06041	
62	*Haematobia irritans*	U11642	
63	*Mantispa pulchella*	U11649	
64	*Bactrocera dorsalis*	AY601743	
		***Maibmar-like elements* DD34D**	
65	*Cancer pagurus*	AJ507245	
66	*Eriphia verrucosa*	AM906106	
67	*Thalamita possoinii*	AM906155	
68	*Pilumnus hirtellus*	AM906121	
69	*Xantho hydrophilus*	AM906156	
70	*Maia brachidactyla*	AJ507238	
		***Bytmar-like element* DD34D**	
71	*Alvinella caudata*	AJ496120	
72	*Ventiella sulfuris*	AJ507232	
73	*Perisesarma bidens*	AM906146	
74	*Bythograea thermydron*	AJ507219	
75	*Portunus pelagicus*	AM906137	
76	*Alvinella pompejana*	AJ496135	
		***Nosferatis* DD34D**	
77	*Rhodnius prolixus*	Rpmar13 (Filée et *al.* 2015)	
78		Rpmar9 (Filée et *al.* 2015)	
		***Dnomar-like element* DD34D**	
79	*Diuraphis noxia*	gi984735891:2467-2015 strand-	*
80		gi984745098:54676-56035 strand+	Dnomar3.1
81	*Myzus persicae*	Scaffold280:15925-14979 strand -	*
82		Scaffold220: 401477-402088 strand +	*
		***Hsmar2-like element* DD34D**	
83	*Portunus pelagicus*	AM906137	
84	*Alvinella pompejana*	AJ496135	
85	*Homo sapiens*	U49974	
86	*Lemur catta*	AC133072	
87	*Gorilla gorilla*	AC145402	
		***Batmar-like element* DD34D**	
88	*Diuraphis noxia*	gi984745342:336477-336902 strand+	*
89		gi984744928:23881-24312 strand+	*
90		gi984745420:365693-366623 strand+	*
91		gi984745311:440129-439207 strand-	*
92	*Acyrthosiphon pisum*	gi320447037 NW003383534.1 Scaffold45:158750-158202 strand-	*
93		gi320446899 NW003383672.1 Scaffold183:261886-261320 strand-	*
94	*Diuraphis noxia*	gi984744320:45892-45397 strand-	*
95		gi984744320:47738-47243 strand-	*
96	*Acyrthosiphon pisum*	gi320446184 NW003384387.1 Scaffold898 :36060-36989 strand -	*
97		gi320446473 NW003384098.1 Scaffold609:132371-133021 strand+	*
98		gi320438813 NW003391455.1 Scaffold7966:4139-4600 strand+	*
99		gi320446392 NW003384179.1 Scaffold690:96269-97548 strand -	Apismar2.2
100		gi320447011 NW003383560.1 Scaffold71:493658-494941 strand +	
101		gi320447000 NW003383571.1 Scaffold82:64692-64037 strand -	*
102	*Heliconius melpomene*	gi378865014 CAEZ01008735.1 Contig7180001235928:24625-25792	
103	*Drosophila eugracilis*	gi449842783 AFPQ02005657.1 Contig5655:975265-976459	
104	*Neodiprion lecontei*	gi914279877 LGIB01001307.1 Scaffold1307:12116-13443	
105	*Lasius niger*	gi861599989 LBMM01019009.1:53-1374	

106	*Diuraphis noxia*	gi984745480:182128-183428 strand+	*
107		gi984745116:79279-80622 strand+	Dnomar2.2
108	*Agrilus planipennis*	gi648140536 JENH01008607.1 Contig8616:12243-13577	
109	*Diuraphis noxia*	gi984745529:295648-297002 strand+	Dnomar2.2
110	*Rhodnius prolixus*	Rpmar1 (Filée et *al.* 2015)	
111	*Dinoponera quadriceps*	gi938133368 JPHR01007292.1 Scaffold1145:2209123415	
112	*Copidosoma floridanum*	gi619889135 JBOX01069944.1 Contig69949:1149212815	
113	*Ceratitis capitata*	gi488305875 NW004523814.1 Contig13099:32902-34140	
114	*Drosophila ficusphila*	Dromar8 (Wallau et *al.* 2014)	
115	*Acyrthosiphon pisum*	gi320445628 NW003384943.1 Scaffold1454:8232- 8961 strand -	*
116		gi320445628 NW003384943.1 Scaffold1454:3454-4361 strand-	MITE2.1
117		gi320447015 NW003383556.1 Scaffold67:738385-739292 strand-	
118		gi320445203 NW003385294.1 Scaffold1805:8796-7866 strand-	
119		gi320446728 NW003383843.1 Scaffold354:173658-174753 strand -	*
120		gi320446952 NW003383619.1 Scaffold130:212179-213501 strand -	Apismar2.1
121	*Myzus persicae*	Scaffold544:202614-203522 strand+	MITE2.2
122		Scaffold166:292461-293374 strand+	
123		Scaffold6:1479795-1480706 strand+	
124	*Diuraphis noxia*	gi984744707:110647-111912 strand+	Dnomar2.1
125		gi984744972:124915-123590 strand-	
126	*Acyrthosiphon pisum*	gi320446088 NW003384483.1 Scaffold994:76058-76816 strand+	*
127	*Blattella germanica*	gi692674178 JPZV01249368.1 ContigNC249368:8211803	
128	*Trabutina mannipara*	gi1044319939 FKYK01006678.1 :1-1114	
129	*Camponotus floridanus*	gi304581076 AEAB01024585.1 Contig622:1601-2726	
130	*Trionymus perrisii*	gi1010807036 FIZV01000272.1 :504-1644	
131	*Trachymyrmex septentrionalis*	gi1006956206 LKEZ01022017.1 Contig22017:3294-4437	
132	*Homalodisca vitripennis*	gi642782498 JJNS01106073.1 ContigNC106073:3898-5039	
133	*Rhinolophus ferrumequinum*	AC157888	
134	*Wasmannia auropunctata*	gi780611046 XM011690486.1	
135	*Rhodnius prolixus*	Rpmar26 (Filée et *al.* 2015)	
136	*Carollia perspicillata*	AC148202	
***Pacmar-like element* DD34D**			
137	*Pachygrapsus marmoratus*	AM231069	
138		AM231072	
139	*Portunus granulatus*	AM906131	
140	*Pachygrapsus marmoratus*	AM983536	
141	*Portunus granulatus*	AM906134	
142		AM906132	
143	*Thalamita possoinii*	AM906151	
144	*Paromola bathyalis*	AM906119	
145	*Perisesarma bidens*	AM906150	
146	*Atelecyclus undecimdentatus*	AM906092	
***Other mariner elements* DD34D**			
147	*Papilio xuthus*	AB055185	
148	*Attacus atlas*	AB006464	
149	*Hyalophora cecropia*	M63844	
150	*Bombyx mori*	D88671	
151	*Antheraea yamamai*	AB247378	
152	*Antheraea mylitta*	AF126011	
153	*Homo sapiens*	EF517118	
154	*Apis mellifera*	AY155490	
155	*Forficula auricularia*	AY155492	
156	*Ceratitis capitata*	U76903	
157	*Caenorhabditis elegans*	U10438	
158	*Meloidogyne chiwoodi*	AJ437557	
159	*Caenorhabditis briggsae*	AC099767	
160	*Solenopsis invicta*	AF518170	
161	*Solenopsis saevissima*	AF518177	
162	*Myrmica ruginodis*	AY652423	

163	*Bombus terrestris*	AJ312712	
164	*Drosophila mauritiana*	M14653	
165		X78906	
166	*Drosophila simulans*	X89927	
167	*Mamestra brassicae*	AF465247	
168	*Messor bouvieri*	AJ781769	
169	*Mayetiola destructor*	gi30778846 AEGA01027875.1	
170		U24436	
171	*Rhynchosciara sp.*	GU442128	
		***maT* DD37D**	
172	*Caenorhabditis elegans*	AF038612	
173		Z83129	
174	*Caenorhabditis briggsae*	AC084524	
175	*Bombyx mori*	U43131	
176	*Anopheles gambiae*	AAAB01008975	
		Outgroup TLE and IS630	
177	*Fusarium oxysporum*	AF282722	
178	*Caenorhabditis briggsae*	M64308	
179	*Drosophila virilis*	CH940657	
180	*Salmo salar*	AJ249090	
181	*Anopheles gambiae*	U89802	
182	*Aedes aegypti*	AF208675	
183	*Caenorhabditis elegans*	M77697	
184	*Aedes atropalpus*	AY038027	
185	*Fusarium oxysporum*	AF076632	
186	*Fusarium solani*	AF443562	
187	*Fusarium oxysporum*	AF076631	
188		AJ608703	
189	*Aspergillus niger*	U58946	
190	*Sinorhizobium meliloti*	AF143444	
191	*Pseudomonas sp.*	U15298.1	
192	*Catharanthus roseus*	DQ852611	
193	*Salmonella typhimurium*	M58505	

Analyse globale et Conclusions

Dans ce chapitre, nous nous sommes intéressés à la caractérisation des éléments de la sous-famille *irritans* aussi bien chez les pucerons des céréales *Rhopalosiphum padi, R. maidis, Sitobion avenae, Schizaphis graminum*, dont le génome n'a pas encore été séquencé, que chez les plantes-hôtes (céréales : orge, blé, avoine, égilope).

Après alignement des séquences *irritans* identifiées *in silico* chez trois espèces de pucerons qui sont proches phylogénétiquement des pucerons des céréales (chapitre I), des séquences consensus de TIR ont été définies. L'amplification par PCR a révélé deux types de copies, d'une part des communes entre ces espèces appelées *Aphidmar* et d'autre part, la séquence nommée *Samar2* spécifique à *S. avenae*. Ces deux types de copies sont de ~944 et 653 pb respectivement. La comparaison des séquences *Aphidmar* a montré un pourcentage d'identité supérieur à 98%. Un tel niveau de similitude a déjà été mentionné entre plusieurs copies délétées de la sous-famille *mauritiana* identifiées chez sept espèces de pucerons des arbres fruitiers (Kharrat *et al.* 2015). Ceci suggère soit une invasion ancienne de copies complètes dans le génome de l'hôte, suivie d'acquisition de mutations chez l'ancêtre commun des pucerons, soit de transfert de copies délétées par des éléments actifs endogènes (Brunet *et al.* 1999; Le Rouzic *et al.* 2007). Ces éléments *irritans* sont non fonctionnels et comprennent des délétions internes qui ont induit à la perte soit du domaine de liaison à l'ADN pour *Aphidmar* soit du domaine catalytique pour *Samar2.* En raison de ces délétions internes, ils correspondraient donc à des MITEs. Leur alignement avec des séquences complètes trouvées précédemment montrent qu'ils sont très proches des éléments *Apismar1.1*, *Dnomar1.1* et *Mpmar1.1* et appartiennent ainsi à la tribu des *macrosiphinimar*.
En raison de l'absence de copies actives chez ces espèces, les deux types d'éléments obtenus sont considérés comme "orphelins". Ceci suggère que leurs partenaires autonomes ancestraux auraient été perdus après l'amplification des MITEs, mais il ne peut être exclu qu'ils existent et se maintiennent dans d'autres populations dans la mesure où peu de populations naturelles ont été analysées.
Par ailleurs, à partir de la comparaison des génomes de plusieurs espèces tels que des *Drosophila* (Wallau *et al.* 2014), la punaise *Rhodnius prolixus* (Filée *et al.* 2015) et les pucerons appartenant à la tribu des *Macrosiphini* (données non publiées), la majorité des MITEs de la sous-famille *irritans* voire même de la famille *mariner* présentent une taille d'environ de 950pb. Peu d'entre eux ont une taille inférieure à 900 pb. Néanmoins, si des MITEs très courts comme *Samar2* sont détectés *in silico* ou *in vitro*, ils sont représentés par un très faible nombre de copies et sont restreints à quelques espèces. Par conséquent, la capacité des éléments *mariner* à transposer serait

fortement contrainte par leur taille et/ou par leur structure (Brunet *et al.* 1996; Lohe et Hartl, 1996), et des copies courtes seraient vraisemblablement inaptes à s'amplifier au sein d'un génome.
Il faut aussi noter que les résultats des recherches *in vitro* et *in silico* de la sous-famille *irritans* chez *Acyrthosiphon pisum*, espèce utilisée comme témoin positif au cours de ce travail, ont montré une grande variabilité. Ceci pourrait être expliqué par l'utilisation de deux populations différentes, au cours de ces recherches, qui sont probablement polymorphes. Le même cas décrit par Barrón *et al.* (2014), pour des ETs de drosophile, a amené ces auteurs à suggérer que la prévalence, la diversité, la distribution et la localisation des éléments au sein des populations appartenant à *Drosophila melanogaster* peuvent varier considérablement d'une population à l'autre.

Afin de vérifier l'existence des éléments *irritans* identifiés chez les pucerons au sein des céréales, nous nous sommes d'abord proposés de balayer les WGS de plusieurs espèces de céréales (*Hordeum vulgare*, *Triticum aestivum*, *Brachypodium Distachyon* et *Aegilops tauschii*). L'approche porte sur la similitude avec les séquences complètes ou MITEs de la sous-famille *irritans* trouvées chez les pucerons. Les copies de tailles inférieures à 300 pb et/ou localisés enfin de contigs ont été éliminés. Seul un contig trouvé dans le génome du cultivar *barke* de l'orge contient un élément tronqué de 320 pb entouré de part et d'autres par de l'ADN génomique de pucerons, le tout flanqué par de l'ADNg de la plante hôte. Cette organisation particulière pourrait être expliquée par des recombinaisons effectuées après une invasion ancienne d'un élément complet dans le génome de puceron, suivie par sa perte due à des mutations. Nous avons ensuite procédé à la vérification *in vitro* dans les feuilles de plusieurs variétés d'orge (*Roho, Martin, Bowman, HD29, Rihane*), de blé (*BLG, Salambo*) et celles de l'avoine. Il est à noter que les extraits de l'ADNg de plantes sont dépourvus de contamination par de l'ADN d'origine animale après amplification du gène de la *chitine synthase1*.

La recherche *in vitro* des copies *irritans* a porté sur l'utilisation des mêmes amorces TIRs et d'une amorce interne définie en fonction de l'élément *in silico* tronqué. Deux types d'éléments ont été caractérisés chez l'orge dans le cultivar *Roho*. Le même élément tronqué que celui décrit précédemment, a été trouvé (*Hvumar1*) ainsi qu'un autre élément délété, noté *Hvumar2*, de 647 pb a été obtenu présentant 99% d'identité avec *Samar2*. Leur présence chez l'orge, espèce phylogénétiquement distante de brachypodium, de l'avoine et de l'ancêtre commun des égilopes et du blé, pourrait être une première preuve d'un éventuel transfert horizontal ancien entre les pucerons des céréales et l'orge.

Enfin la dernière investigation a porté sur l'identification des régions flanquantes de l'élément tronqué. Toutefois cette étape n'a pas abouti (amplification de région inconnue ou de

rétrotransposons). L'absence d'amplification peut être expliquée par la non-conservation de la région flanquante entre les différentes variétés d'orge, étant donné que nous n'avons pas utilisé *in vitro* le même cultivar *barke* exploité *in silico*. Il serait nécessaire de cibler les régions flanquantes (ADNg de plantes) des deux types d'éléments trouvés, par la définition d'autres amorces pour *Hvumar2* et par leur recherche dans le cultivar *barke* pour l'élément tronqué, afin de renforcer l'existence d'un transfert horizontal.

Jusqu'à ce jour, les acteurs et mécanismes d'un tel transfert ne sont toujours pas connus chez les eucaryotes. Toutefois, des interactions hôtes-parasites durables, telles celles existant entre plantes et pucerons, sont des terrains favorables aux transferts horizontaux et à leur étude (Gilbert *et al.* 2010). Dans ce contexte, les interactions entre pucerons et céréales constituent un bon modèle pour ce type d'étude et plusieurs vecteurs de transferts peuvent être considérés. Ainsi, les virus sont capables d'injecter de l'ADN/ARN dans les cellules de l'hôte (Jehle *et al.* 1995; Routh *et al.* 2012) et on sait que les génomes de plusieurs virus (*i.e.* iridovirus) spécifiques des diptères peuvent contenir des éléments transposables (Piégu *et al.* 2013). Par ailleurs, les bactéries telle que l'endosymbiote *Wolbachia* sont également des vecteurs potentiels, sachant que ces bactéries sont fréquemment sujettes à des transferts horizontaux, qu'elles ciblent les lignées germinales et qu'elles peuvent contenir plusieurs gènes d'eucaryotes (Duplouy *et al.* 2013).

Chapitre III

Evolution moléculaire de la super-famille des *piggyBac* : du parasitisme à la domestication

Molecular evolution of *piggyBac* superfamily: from selfishness to domestication

Chapitre III

Evolution moléculaire de la super-famille des *piggyBac* : du parasitisme à la domestication.

L'élément transposable *piggyBac* a été initialement découvert chez le lépidoptère *Trichoplusia ni* (Fraser *et al.* 1983). Il est fonctionnel chez plusieurs espèces d'insectes ainsi que d'autres organismes y compris les levures, les protozoaires, les vertébrés et les plantes (Yusa 2015). D'autres éléments ont été identifiés chez différentes espèces de métazoaires, essentiellement chez les arthropodes (voir références dans Yusa 2015). A quelques exceptions près, la plupart des éléments *piggyBac* caractérisés jusqu'ici ne sont pas fonctionnels. Par conséquent, l'identification de nouveaux éléments actifs, ainsi qu'une vision générale de la distribution et de la structure de cet élément sera utile pour mieux comprendre son évolution et éventuellement contribuer au développement de vecteurs de transfert de gènes.

Notre objectif initial était de rechercher *in vitro* les éléments *piggyBac* chez les pucerons des céréales, *Rhopalosiphum padi*, *R. maidis* et *Schizaphis graminum* de la tribu des Aphidini, ainsi que *Sitobion avenae* de la tribu des Macrosiphini. Pour cela, nous avons utilisé les amorces élaborées par Luo *et al.* (2014) qui ont permis de mettre en évidence des séquences *piggyBac* complètes, potentiellement actives, chez le puceron du cotonnier *Aphis gossypii* de la tribu des Aphidini. Cependant et malgré les mêmes protocoles expérimentaux, nous avons obtenu, après amplification et visualisation sur le gel d'agarose, un smear indiquant la présence de plusieurs copies avec des tailles variant de 800 à 3000 pb. Nous avons alors décidé d'explorer le génome séquencé du puceron du petit pois *Acyrthosiphon pisum*, en utilisant comme requêtes les neuf séquences nucléotidiques trouvées chez *A. gossypii*. Au total, nous avons pu détecter 231 copies de *piggyBac* chez ce puceron (données personnelles).

La caractérisation du polymorphisme et la distribution de ces éléments sont richement illustrés dans la littérature. Toutefois, cette super-famille n'a pas fait l'objet d'une analyse approfondie de l'ensemble de ces éléments et de leur évolution. Deux principaux types de séquences peuvent être distingués :

- Les éléments *piggyBac-like* elements (PBLE) correspondant à des éléments complets comportant les deux TIR, les UTR en 5' et 3' ainsi que le gène codant la transposase avec deux domaines fonctionnels : (i) un domaine catalytique caractérisé par la présence de trois résidus aspartates (D) qui assurent les réactions catalytiques de coupure de l'ADN et la ligation, et (ii) un domaine riche en cystéines dans la région C-terminale qui forme des doigts de zinc assurant la

fixation à l'ADN. Ces éléments sont potentiellement actifs et susceptibles d'être utilisés en mutagenèse.

- Les éléments *piggyBac-derived* genes/proteins (PGBD) qui sont des éléments en copie unique dérivant du gène codant pour la transposase. Ils sont conservés au cours de l'évolution et correspondent à des éléments domestiqués.

Par ailleurs, face à l'explosion des génomes complets séquencés, le répertoire des éléments annotés *piggyBac* a été fortement enrichi dans les banques de données (NCBI, Repbase). Ceci nous a amené à rechercher et à caractériser cette superfamille incluant les PBLE et les PGBD, dans un large spectre d'organismes.

Pour cela, 107 séquences protéiques des PBLE et 10 séquences des PGBD, incluant cinq séquences de référence identifiées initialement chez l'Homme *Homo sapiens* (PGBD1à 5), deux séquences orthologues à PGBD5 identifiées chez l'agnate *Petromyzon marinus* (Pma) et chez le céphalocordé *Branchiostoma floridae* (Bfl), l'élément KOBUTA de *Xenopus sp* et deux séquences *piggyMac*, à savoir PGM de *Paramecium tetraurelia* et TPB2 de *Tetrahymena thermophila*, ont été utilisées comme requête pour rechercher des séquences homologues dans la base de données NCBI-nr et dans les génomes disponibles (WGS) par TBLASTN.

Les séquences détectées ont conduit à :

- l'étude de l'organisation, de la structure et de l'évolution des éléments PBLE
- la caractérisation des éléments PGBD
- l'analyse de la distribution, de l'évolution et des relations phylogénétiques entre ces deux groupes.

Les résultats de ce chapitre ont fait l'objet d'un article publié online dans la revue Genome Biology and Evolution (disponible dans la rubrique Advance access.

GBE

Molecular Evolution of *piggyBac* Superfamily: From Selfishness to Domestication

Maryem Bouallègue[1,2], Jacques-Deric Rouault[1], Aurélie Hua-Van[1], Mohamed Makni[2], and Pierre Capy[1,*]

[1]Laboratoire Evolution, Génomes, Comportement, Ecologie CNRS, Univ. Paris-Sud, IRD, Université Paris-Saclay, Gif-sur-Yvette, France

[2]Université de Tunis El Manar, Faculté des Sciences de Tunis, UR11ES10 Génomique des Insectes Ravageurs de Cultures, Tunis, Tunisie

*Corresponding author: E-mail: pierre.capy@egce.cnrs-gif.fr.

Accepted: December 13, 2016

Abstract

The *piggyBac* transposable element was originally isolated from the cabbage looper moth, *Trichoplusia ni*, in the 1980s. Despite its early discovery and specificity compared to the other Class II elements, the diversity and evolution of this superfamily have been only partially analyzed. Two main types of elements can be distinguished: the *piggyBac-like elements* (PBLE) with terminal inverted repeats, untranslated region, and an open reading frame encoding a transposase, and the *piggyBac-derived* sequences (PGBD), containing a sequence derived from a *piggyBac* transposase, and which correspond to domesticated elements. To define the distribution, their structural diversity and phylogenetic relationships, analyses were conducted using known PBLE and PGBD sequences to scan databases. From this data mining, numerous new sequences were characterized (50 for PBLE and 396 for PGBD). Structural analyses suggest that four groups of PBLE can be defined according to the presence/absence of sub-terminal repeats. The transposase is characterized by highly variable catalytic domain and C-terminal region. There is no relationship between the structural groups and the phylogeny of these PBLE elements. The PGBD are clearly structured into nine main groups. A new group of domesticated elements is suspected in *Neopterygii* and the remaining eight previously described elements have been investigated in more detail. In all cases, these sequences are no longer transposable elements, the catalytic domain of the ancestral transposase is not always conserved, but they are under strong purifying selection. The phylogeny of both PBLE and PGBD suggests multiple and independent domestication events of PGBD from different PBLE ancestors.

Key words: transposable element, piggyBac, molecular evolution, domestication.

Introduction

Transposable elements (TEs) are mobile and repetitive genetic elements, abundant in all eukaryotic genomes investigated so far. Two classes of elements are distinguished according to their respective transposition mechanisms (Wicker et al. 2007). With few exceptions (like *SINEs*, MITES, LARD elements), TEs encode their own transposition machinery. They use a reverse transcriptase and integrase, leading to a "copy and paste" mechanism (retrotransposons or *Class I*) or a transposase in a "cut and paste" mechanism (DNA transposons or *Class II*). Excluding dead copies (*i.e.* with no coding capacity and not mobilizable), both classes exist as autonomous active copies, which encode all the factors required for their mobility and as nonautonomous but *trans*-mobilizable copies depending on the transposition machinery of their autonomous relatives (Feschotte et al. 2002).

While TEs are generally considered as selfish sequences or genomic parasites, they are also important evolutionary factors in both structural and functional dynamics of the genomes. Indeed, one of the most direct contributions of TEs to host genome evolution is their potential role in the emergence of new genes and functions through an exaptation process also known as a "molecular domestication" where the use of TE sequences for a new function is usually associated to the loss of their mobility capacities (Britten 1996; Kidwell and Lisch 1997; Miller et al. 1999; Kidwell 2002; Volff 2006; Sinzelle et al. 2009; Joly-Lopez et al. 2016). In these cases, while "classical" TEs are present in multi-copies and inserted at different positions in the genome, within and between species, domesticated elements are generally found as single orthologous copies in different species (if the domestication is old enough). In addition, a low ratio ($Ka/Ks < 1$) of nonsynonymous (*Ka*) to synonymous (*Ks*) nucleotide

substitution rates suggests that the sequence is under strong purifying selection if the new function is already acquired, while a high ratio ($Ka/Ks > 1$) indicates that the sequences are under positive selection, that is, when the adaptive peak is not yet reached (Hurst 2002).

This process can be illustrated by *RAG1* in V(D)J recombination (Kapitonov and Jurka 2005; Hencken et al. 2012) and the *CENPB* centromere protein (Casola et al. 2008), which derive from the *Transib* and the *pogo* transposases, respectively. More recently, in the primate lineage, a *mariner-like* transposase was fused to an SET histone methyltransferase domain by *de novo* exonization. The fusion protein retains the ancestral DNA binding activity of the transposase and acts as a transcriptional regulator of dispersed *mariner-like* repeat elements (Cordaux et al. 2006; Liu et al. 2007).

The *piggyBac* element (formerly IFP2) is a typical Class II element, originally isolated from a mutant baculovirus in a cell culture of the Lepidopteran cabbage looper moth *Trichoplusia ni* (Fraser et al. 1983; Cary et al. 1989). More precisely, this element jumped from *T. ni* to the baculovirus. This 2,475 bp autonomous mobile element inserts in TTAA target sites and is bounded by 13 bp terminal inverted repeats (TIRs) and 19 bp sub-terminal asymmetric inverted repeats (STIR) located 3 and 31 bp from the 5′ and 3′ TIRs, respectively (Cary et al. 1989; Fraser et al. 1996; Lobo et al. 1999). The single open reading frame is 1,782 bp long, coding for a protein of 594 amino acids with a molecular weight of 64 kDa. Interestingly, *piggyBac* is also functional in other organisms, including yeasts, protozoa, vertebrates and plants (Yusa 2015). Due to its high transposition activity in several species of insects, it has become one of the most widely used systems for the germline transformations, as well as a genetic tool for gene tagging or trapping, and "an insertional mutagen" (Yusa 2015).

Since its discovery in 1983, *piggyBac* has for a long time remained the only member of the currently known *piggyBac* superfamily. The taxonomic distribution, initially believed to be restricted to the insect orders (Coleoptera, Diptera, Hymenoptera, Lepidoptera, and Orthoptera; see Wang et al. 2008 and references therein), has been significantly expanded to cover several eukaryotic groups. Indeed, a number of *piggyBac*-like elements (PBLE) from other species or from baculovirus have been identified (Penton et al. 2002; Arkhipova and Meselson 2005; Pritham et al. 2005; Wang et al. 2006; Xu et al. 2006; Hikosaka et al. 2007; Ray et al. 2008; Sun et al. 2008; Wang et al. 2008; Wu et al. 2008; Carpes et al. 2009; Wang et al. 2009; Daimon et al. 2010; Pagan et al. 2010; Luo et al. 2011; Wu et al. 2011; Luo et al. 2014; Wu and Wang 2014).

Most mammalian genomes also contain decayed *piggyBac* transposons. They ceased their activity due to several mutations or rearrangements (Pace and Feschotte, 2007; Pagan et al. 2010). The picture emerging from the initial analyses of the human, mouse, rat, and dog genomes shows that there is no evidence for *piggyBac* activity during the past 40 Ma (Lander et al. 2001; Gibbs et al. 2004; Lindblad-Toh et al. 2005; Pace and Feschotte 2007). However, *piggyBac* evolution can be different from one lineage to another. For instance, recent data suggest a continuous colonization of the vesper bat genomes. Several waves of amplification of *piggyBac* have succeeded over the past 40 Ma and the invasion seems to be ongoing, while a new member of PBLE, *piggyBat*, has been identified in the little brown bat *Myotis lucifugus*. This element is active in its native form and in transposition assays in bat and human cultured cells, as well as in *Saccharomyces cerevisiae* (Ray et al. 2008; Mitra et al. 2013).

Using computational analysis of genomic data, several species have also revealed a number of genes, derived from various TEs, including *piggyBac* transposases (Sarkar et al. 2003). Indeed, the human genome contains at least five domesticated *piggyBac*, designated from PGBD1 to PGBD5 (for *piggyBac-derived* genes). On the one hand, PGBD1 and PGBD2 were probably present in the common ancestor of mammals, while PGBD3 and PGBD4 are restricted to primates. On the other hand, PGBD5, the only sequences interrupted by multiple introns, are not only orthologous in mammals and fish (Sarkar et al. 2003) but also in lamprey and lancelet suggesting an ancient domestication event about 525 Ma before cephalocordates and vertebrates split from urochordates (Pavelitz et al. 2013).

Otherwise, domestication of *piggyBac* transposases can be observed in several evolutionary lineages. In the ciliates *Paramecium tetraurelia* (Baudry et al. 2009) and *Tetrahymena thermophila* (Cheng et al. 2010), the genome undergoes massive DNA amplification during macronucleus development, and extensive programmed genome rearrangement, including elimination of TEs and internal eliminated sequences (IES). These eliminations, essential to reconstruct functional genes, are due to domesticated *piggyBac* transposases, named *piggyMac* (*PGM*) in *P. tetraurelia* and *TPB2* in *T. thermophila* (Baudry et al. 2009; Cheng et al. 2010). In *Xenopus*, the *KOBUTA*-domesticated transposase has been conserved for 100 Ma and seems to be involved in DNA-binding or DNA-recombination activity. Moreover, it can inactivate the *Uribo* autonomous transposase through heterodimerization (Hikosaka et al. 2007).

The general features shared by the members of the *piggyBac* superfamily are the TTAA integration target sites and the precise excision of the element leading to the restoration of the pre-integration site. In addition, highly conserved blocks can be detected in the core region, including several aspartic acid (D) and glutamic acid (E) residues (Sarkar et al. 2003; Keith et al. 2008). Although this region does not readily show similarity to the widespread DDE catalytic domains of many Class II transposases and Class I integrases, a weak similarity to the *IS4* family protein was identified (Sarkar et al. 2003), leading to the prediction that D268 and D346 in the *T. ni* transposase might be the conserved aspartic acid of a

DDE/D catalytic domain. Mutational analyses of these positions, as well as another highly conserved D447, revealed that these residues are absolutely required for all steps of transposition. While not conserved in *piggyMac* and *TPB2*, a fourth aspartic acid D450 could also be involved for the excision of the element in cell cultures, while a glutamate substitution can be tolerated (Keith et al. 2008). Another peculiar feature of all *piggyBac* transposases is the conserved Cysteine residues, forming a putative zinc-binding homeodomain (PHD) finger in the C-terminal region (Sarkar et al. 2003; Keith et al. 2008).

The availability of numerous almost complete eukaryotic genome sequences has considerably enriched the repertoire of annotated *piggyBac* elements, providing an opportunity to better characterize the origin, distribution, diversity, structure and evolution of this superfamily as well as those of the *piggyBac-derived* genes. Until now, in many cases the evolutionary scenarios leading to the presumed domestications have not been fully reconstructed particularly because the ancestral copies were difficult to identify unambiguously.

The objective of this work is to identify and characterize *piggyBac*-related elements, including domesticated sequences, in a large spectrum of organisms. Structural and sequence comparisons suggest that PBLE (*bona fide* transposons) can be grouped in four different structures due to the presence or absence of subterminal repeats with highly divergent catalytic domains and C-terminal regions. Concerning domesticated *piggyBac*, we identified a new group of PGBD sequences, besides the eight groups already described. Evolutionary scenarios based on the structural features, phylogenetic relationships and fate of these elements are discussed.

Materials and Methods

Data Mining

One hundred and seventeen *piggyBac* transposases were extracted from databases (NCBI, Repbase, and genomes available). Among those, 107 sequences are related to PBLE including 28 sequences from literature and 79 consensus sequences from Repbase with a transposase longer than 300 aa. Sequences containing long truncations, insertions or deletions were not retained. The other ten sequences are related to *piggyBac-derived* elements (PGBD) including five sequences from *Homo sapiens*, two sequences orthologous to PGBD5 of Humans, namely, *Pma* from the agnathe *Petromyzon marinus* and *Bfl* from the cephalochordate *Branchiostoma floridae*, the *KOBUTA* element from *Xenopus sp* and two *piggyMac* sequences, namely *PGM* from *P. tetraurelia* and *TPB2* from *T. termophila* (fig. 1 and supplementary material S1, Supplementary Material online).

Each of these copies was used as query to look for homologous sequences in the NCBI nr database and genomes available by TBlastN. The new sequences identified were in turn used as query in order to identify more PBLE and PGBD. The stringency of the mining steps was between 0 and $1E^{-100}$ (TBlastN) and only transposases of at least 250 amino acids were retained. Sequences used as vector were removed and in case of isoform proteins or identical sequences only one sequence was selected.

Structural Analysis of Copies

The annotation of *piggyBac* sequences in databases can sometime be confusing. For instance, the term PGBD-like can be used indifferently for PBLE, PGBD or for degenerated copies (with internal deletion, truncation...). In the present work, PBLE will refer to sequences (active or not) with two TIRs, two untranslated regions (UTRs) and a transposase. PGBD will correspond to domesticated sequences (i.e., sequences in single copy in orthologous position in different species and a *Ka/Ks* indicating the existence of positive or purifying selection). Finally, the PGBD-like term will be restricted to the remaining sequence *i.e.* with one or no TIRs or UTRs and a transposase that can be partially deleted or truncated.

The BLAST searches were done using the transposase sequence as query. Therefore, to discriminate between PBLE sequences from PGBD-like elements, TBlastN was used to align each sequence on the host genome. Then, 5 kb of both flanking regions were added to detect potential TIR. Flanking sequences containing more than 100 N were removed. When TIRs were found on both sides, the sequence was considered as new complete PBLE, while sequences flanked by a single TIR or no TIR were considered as PGBD-like element (supplementary material S1, Supplementary Material online). Then, in PBLE, direct repeats (DR) and sub-terminal repeats (STIR) were searched by alignment of flanked sequences of the transposase using BlastN (supplementary material S2, Supplementary Material online). At the end of this screening and filtering steps, a total of 157 sequences of PBLE (107 from Repbase and literature plus 50 new sequences) and 406 PGBD or PGBD-like sequences were retrieved (fig. 1 and supplementary material S1, Supplementary Material online).

PBLE vs. PGBD

PBLE transposases were aligned with AliView (version 1.17.1; Larsson 2014) using Muscle with default parameters. Blocks with a conservation level higher than 30% (relative to the consensus given by AliView) with no long indels (>20aa) were retained for phylogenetic analyses (supplementary material S3, Supplementary Material online). The phylogenies (Maximum likelihood, but different methods led to similar topologies), based on the previous blocks, were inferred with MEGA6 (Tamura et al. 2013) after a search for the best evolutionary scenario with ProtTest 2.4 server (AIC, matrix LG + F+G). For the phylogeny of PGBD and all copies (PBLE, PGBD, and PGBD like), the same procedure was followed

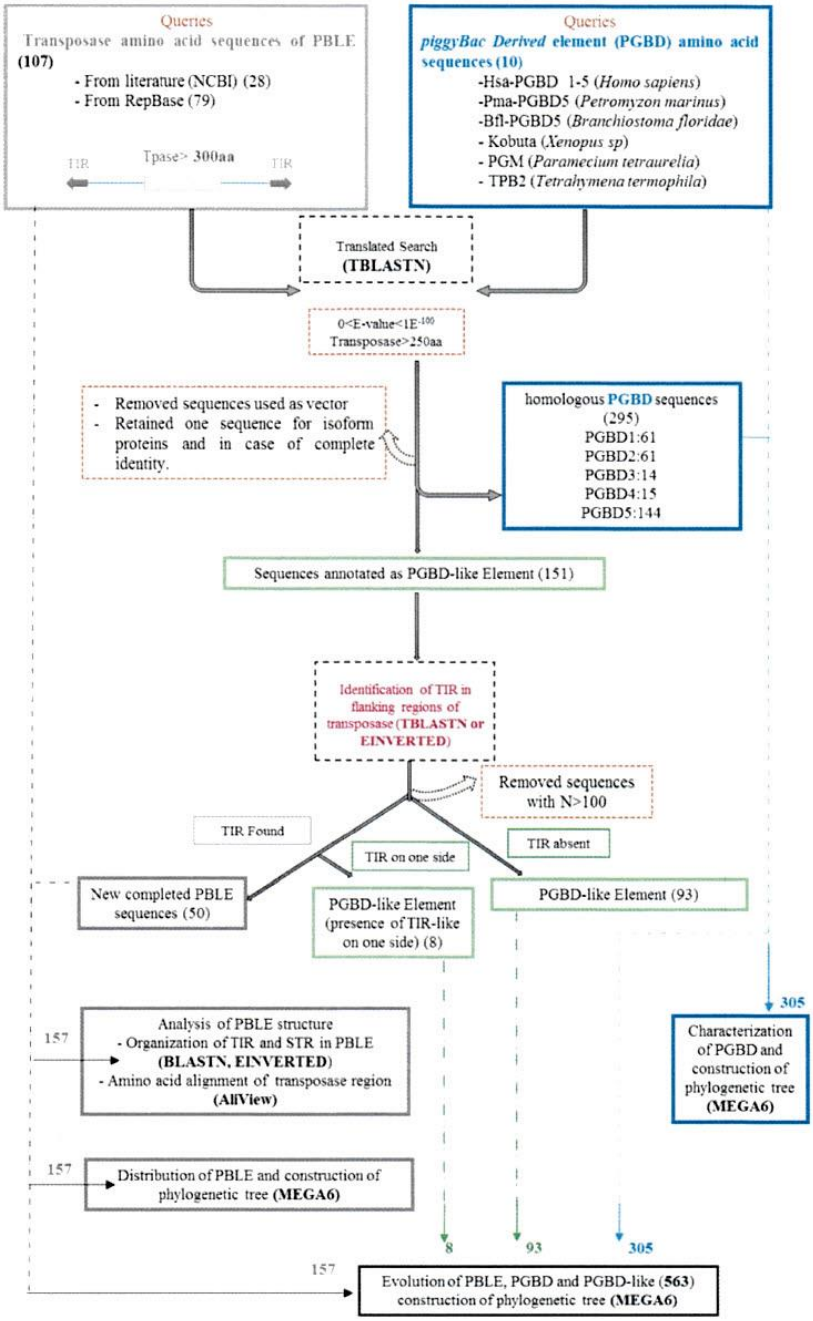


Fig. 1.—Data mining of *piggyBac* element. Search of elements belonging to the *piggyBac* superfamily was done using known copies (literature or database) as BLAST queries. Ten PGBD and 107 PBLE copies were used as queries. After a first run, 295 sequences annotated as PGBD (PGBD1, 2, 3, 4, and 5) and 151 new sequences, annotated as PGBD-like element, belonging to 58 species were retrieved. TIR were looked for in the 5′ and 3′ flanking regions (5 kb in both directions) of these new sequences. Fifty of them were found with two TIR and 101 with one (5′ or 3′) or no TIR. *In fine*, a total of 563 sequences were available. The sequences with more than 100 N were eliminated.

(supplementary materials S4 and S5, Supplementary Material online).

Results

Structures, characteristics and phylogenies of PBLE and PGBD elements were analyzed separately because they correspond to different types of sequences with different evolutionary trajectories. The former are putatively active transposons, while the latter are domesticated. Then, to infer the phylogenetic relationships between all the *piggyBac*-like sequences, the PGBD-like elements were added to these two groups. Beyond the analysis of the molecular evolution of *piggyBac*, our objective was to determine which type of *piggyBac* elements gave rise to the domesticated sequences using the phylogenetic proximities.

PBLE Landscape

Among the 107 sequences used here as queries, 64 (59%) correspond to potentially functional transposases, while the others are defective due to the presence of multiple stop codons and/or frameshifts. The TBlastN allowed us to identify 50 new sequences from different genomes.

Characterization of PBLE: Structures and Distribution among Species

Members of the *piggyBac* superfamily have a TTAA sequence as target site duplication (TSD) and have TIR of 12-19 bp long (Fraser et al. 1983). From the analysis of the 157 PBLE, we show that rarely other insertion sites, including TTTT, ATAG, TTAT, ATAT, and ATAA, can be detected for a unique sequence (supplementary material S2, Supplementary Material online), and that the average length of the TIR is 14 ± 2 bp (without the TIR of 50 bp from *Paracoccidioides brasiliensis*). Moreover, while the first six nucleotides are relatively well conserved with the following motif **C**[**C**/A/T][**C**/A/T][**T**/G/A][**T**/A/G][**T**/G/A], the remaining part of the TIR can be highly divergent from one copy to another (supplementary material S2, Supplementary Material online). The element size varies from 1,721 to 8,451 bp with an average value of 2,813 ± 84 bp.

A detailed analysis of the 5′ and 3′ ends revealed that, besides the presence of TIR, 94 PBLE contain DR and/or sub-TIR (STIR) at their ends. Based on the presence/absence of these repeats, four structural groups (SG) of PBLEs can be defined (fig. 2 and supplementary material S2, Supplementary Material online). The first group (SG1), comprises 63 sequences characterized by TIR ranging between 5 and 19 bp with the exception of a long TIR of 50 bp in the fungi *P. brasiliensis*. The second group (SG2) contains TIR (7–30 bp) and imperfect STIR, varying from 11 to 400 bp; 16 out of 32 sequences showed an overlapping region between TIR and STIR. The third one (SG3) is characterized by TIR (8–20 bp) and DR (12–42 bp); 5 sequences out of 22 show an overlapping region between TIR and DR. The last group (SG4), containing 40 sequences, is more complex since all sequences contains not only TIRs (6-19 bp), but also the DR (11–42 bp) and STIRs (15-34 bp). We note that 36 out of 40 transposons belonging to this group have STIR and DR that are similar. For example, the 5′ and 3′ DR of *piggyBac-5_Ccri* are identical (same sequence and orientation) to the 5′ and 3′ STIR respectively. Such a phenomenon is possible because the common sequence is palindromic (supplementary material S2, Supplementary Material online). For the other members of this group, and as already mentioned for the SG2 and SG3, overlaps between DR/STIR and TIR can be observed. Finally, no consensus sequence has been found for DRs or STIRs.

Another interesting feature is length variability of the four previous groups. Indeed, when the different parts of the element are considered [5′ and 3′ TIR, UTRs and open reading frame (ORF)], the coefficients of variation of UTRs (5′ and 3′) are much higher (from 2 to 8 times) than those of TIR and ORF for all SGs considered, suggesting more selective pressures on the latter (supplementary material S6, Supplementary Material online). Moreover, the SG1 group characterized by the absence of subterminal DR and STIRs, seems to be more variable than the other SG, except for the ORF region.

The specific distribution of the four SGs of PBLE (fig. 2), shows that some species can contain a single SG and others from two to four SG. In particular, the SG1 group is present in many species including protozoa, red algae, fungi and metazoan, while the others SG are less widely distributed. Therefore, the predominance of SG1 over the other groups (63 SG1, 32 SG2, 22 SG3 and 40 SG4, χ^2 $P < 10^{-5}$), its large specific distribution and its higher variability could suggest that the *piggyBac* sequences containing only TIR are the ancestral structure. The underlying hypothesis to this proposition is that the presence of sequences belonging to other SG could be due to evolutionary convergence and/or to horizontal transfers. An argument in favor of the convergence hypothesis is the absence of consensus sequences in the UTR. However, the alternative hypothesis suggesting the presence of the four SGs in the common ancestor of all species here considered, followed by independent loss in several lineages, leading to a patchy distribution and a rapid evolution of DR and STIR sequences cannot be excluded.

The piggyBac-like Transposase Protein Family

In spite of a strong variability of the element lengths, the alignment of amino-acid sequences of the putative transposases encoded by the 157 PBLE shows several highly conserved motifs (supplementary material S3, Supplementary Material online). However, as already mentioned by Sarkar et al. (2003), the N-terminal region (positions 1–130, using *T. ni* transposase as reference), suspected to be a DNA-binding-domain interacting with TIR, is not well conserved and no rational alignment can be provided.

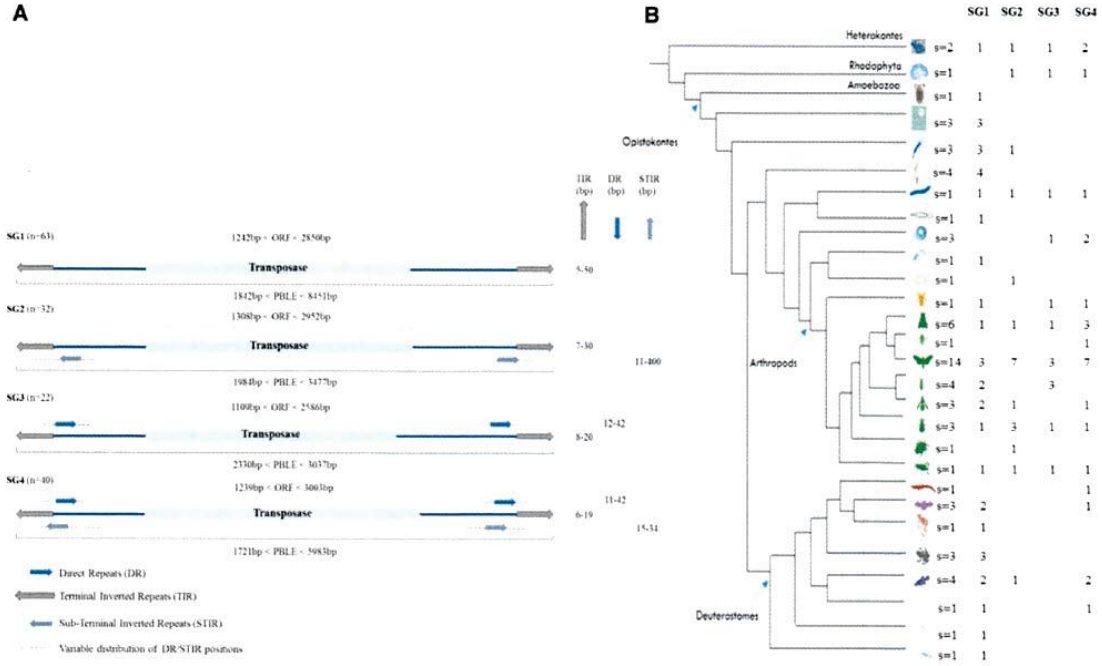


FIG. 2.—Structure of PBLEs and their distribution among species. (*A*) According to the presence/absence of DR and STIRs, four groups of PBLE were proposed (from SG1 to SG4). The number of sequences (*n*) in each group is given in brackets. (*B*) Specific distribution of the four SGs. The number of species (s) is given for each branch of the phylogenetic tree and according to each structure. Involved species (Heterokontes, Rhodophyta, Amoebozoa, and Opistokontes) are detailed, from the top down, as follows: *[Phytophthora infestans, P. ramorum], [Chondrus crispus], [Entamoeba invadens], [Paracoccidioides brasiliensis, Mucor circinelloides, Nosema apis], [Acropora millepora, Hydra magnipapillata, Nematostella vectensis], [Aplysia californica, Biomphalaria glabrata, Crassostrea gigas, Lingula anatina], [Schmidtea mediterranea], [Adineta vaga], [Ancylostoma ceylanicum, Haemonchus contortus, Trichinella spiralis], [Branchipoda crustacean], [Lepeophtheirus salmonis], [Stegodyphus mimosarum], [Aedes aegypti, Anopheles gambiae, Drosophila ananassae, D. biarmipes, D. bipectinata, D. eugracilis], [Mengenilla moldrzyki], [Anticarsia gemmatalis, Agrotis ipsilon, Amyelois transitella, Bombyx mori, Ctenoplusia agnata, Chilo suppressalis, Helicoverpa armigera, Heliconius melpomene, Heliothis virescens, Macdunnoughia crassisigna, Pectinophora gossypiella, Papilio xuthus, Spodoptera frugiperda, Trichoplusia ni], [Cerapachys biroi, Orussus abietinus, Solenopsis invicta, Vollenhovia emeryi], [Athalia rosae, Megachile rotundata, Nasonia vitripennis], [Aphis gossypii, Acyrthosiphon pisum, Diaphorina citri], [Tribolium castaneum], [Locusta migratoria], [Alligator mississippiensis], [Myotis davidii, Myotis lucifugus, Pteropus vampyrus], [Microcebus murinus], [Xenopus borealis, X. laevis, X. tropicalis], [Fundulus heteroclitus, Latimeria chalumnae, Oreochromis niloticus, Salmo salar], [Branchiostoma floridae], [Ciona intestinalis], [Saccoglossus kowalevskii].*

The central part of transposase (positions 130–522, *T. ni* as reference) contains several conserved and clearly delimited blocks (supplementary material S3, Supplementary Material online). Among them are those surrounding the residues of the catalytic domain DDD-D/G in positions 268, 346, 447, and 450. The four first Aspartates are strictly conserved in the putatively active elements and only one sequence with a G is found for the last position in the nematode *Trichinella spiralis*. However, according to Keith et al. (2008), a Glutamate can be tolerated at this position. Therefore, this last residue can be less constrained than the others and may be essential but not involved in the catalytic site itself.

The C-terminal region (positions 559–594) overlapping the *piggyBac* nuclear localization signal (NLS = PVMKKRTY CTYCPSKIRRKAN from position 551 to 571, Keith et al. 2008) is relatively well conserved. However, as mentioned by these authors, if this motif is a functional NLS, the definition of the complete NLS seems to be more difficult. The ZnF motif of the C-terminal region starts in the NLS motif, which includes the first two Cysteines (bold/underlined in the previous motif). Comparison of the different sequences shows that this region may present five to seven conserved Cysteines and sometimes a Histidine. For instance, there are seven cysteines in the *piggyBac* transposase of *T. ni* but five in that of *Aplysia*

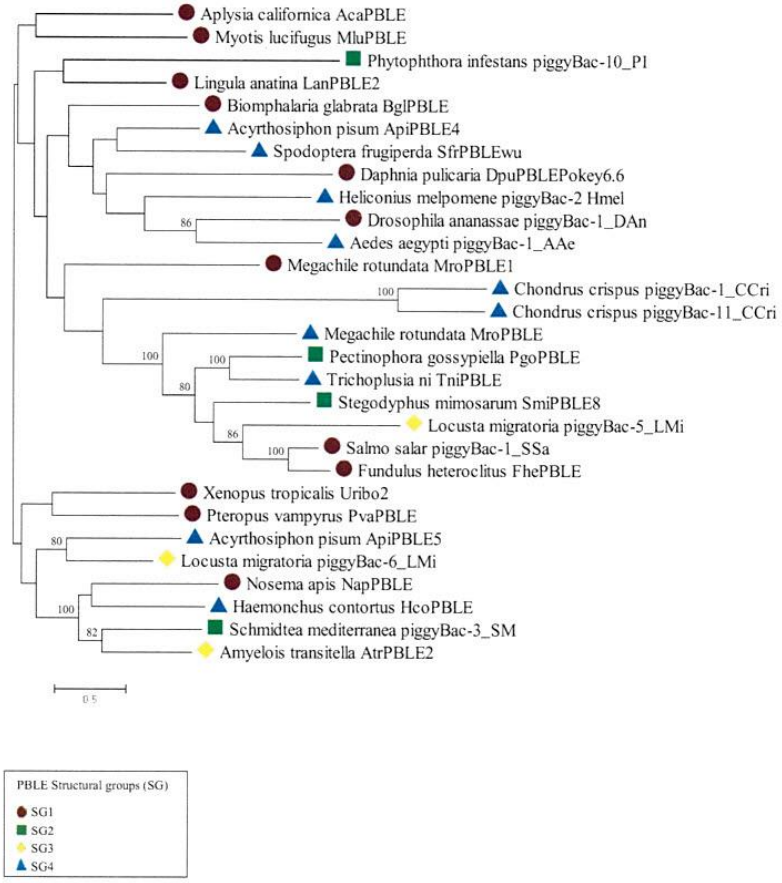


Fig. 3.—Phylogenetic tree of PBLE. This phylogeny is based on amino-acid sequences covering about 237 residues (the alignment is given in supplementary material S3, Supplementary Material online). For simplicity, only 29 sequences are represented, but the complete tree is available in supplementary material S7, Supplementary Material online. After a search of the best evolutionary scenario (ProTest 2.4), this unrooted tree was generated in MEGA6 with the maximum likelihood (ML) method, using LG + F + G matrix. Only bootstrapping values (100 replications) higher than 70% are mentioned on the branch. Red dots, green squares, yellow lozenges and blue triangles refer to the different SGs, that is, SG1, SG2, SG3, and SG4, respectively. Sequence names: the term PBLE (*piggyBac-Like Element*) is used for sequences extracted from literature, or copies newly characterized in this study, while the term *piggyBac* is restricted to sequences extracted from Repbase (real name of these sequences in this database).

californica (sequence named AcaPBLE). Moreover, spacing between these residues are relatively well conserved. Keith et al. (2008) suggested that the ZnF region might not be involved in the DNA binding process but in other processes, including protein-protein interaction as required for a putative dimerization of the transposase.

Phylogenetic Analysis

Despite their high divergence in sequence, most PBLE transposases were found to contain conserved domains for a total of about 237 residues. Thus, it is possible to infer a phylogenetic tree based on these conserved regions (supplementary material S3, Supplementary Material online). The PBLE tree (supplementary material S7, Supplementary Material online) presents short terminal branches reflecting the high similarity between these elements. The simplified phylogeny (fig. 3) shows that there is no congruence between the phylogeny of the transposases and the general structure of the elements. Therefore, no clear evolutionary scenario can be proposed to explain the distribution of the different SG. In addition, the flexibility of the UTR length and sequences, compared to other parts of the elements, and the absence of

consensus sequence between DRs or between STIRs, suggest that the structures of elements belonging to the same SG might be the result of evolutionary convergences.

Analysis of the PGBD

During evolution, functional features of transposases can be exapted to create new genes with specific cellular functions (see for instance the V(D)J system of mammalian derived from the *transib* transposase; Kapitonov and Jurka 2005). In this respect, several copies of *piggyBac* have been described as domesticated sequences (table 1). More precisely, in the human genome, five genes (PGBD1–PGBD5) derived from *piggyBac* transposases (Sarkar et al. 2003). Otherwise, the *KOBUTA* gene of *Xenopus* (Hikosaka et al. 2007), the *PGM* (Baudry et al. 2009) and *TPB2* (Cheng et al. 2010) genes of the ciliates *P. tetraurelia* and *T. thermophila*, respectively, and the *Pma* and *Bfl* sequences of agnathe and cephalochordate (Pavelitz et al. 2013), are also originated from *piggyBac* transposases. The human PGBD5 and the last two sequences (*Pma* and *Bfl*) are orthologous, suggesting an ancient domestication (Pavelitz et al. 2013).

These eight genes were used as query against general database to extract new PGBD sequences. The objective was to get a better idea of the specific distribution of these sequences. This allowed us to retrieve 295 PGBD sequences. All these sequences can be clustered as eight groups. The orthology was verified from a detailed analysis of the 5′ and 3′ flanking regions (as shown in supplementary material S8, Supplementary Material online). All these sequences are found in single copies. They present a coding region in which all or a part of the *piggyBac* transposase is identified. Within each group, the similarity level between the sequences is higher than 85% (see below), and between groups no alignment can be made except for the parts corresponding to the *piggyBac* transposase. Therefore, sequences of the different groups probably correspond to different domestication events.

Based on the five genes described in the human genome, five groups of PGBD sequences (from PGBD1 to PGBD5) can be defined according to the similarity level of their amino-acid sequences (table 1, supplementary materials S4 and S9, Supplementary Material online). PGBD1 members ($n = 62$) are the result of an ancestral fusion between five exons containing LER or SCAN domains (leucine-rich regions) in the N-terminal regions and the transposase of a *piggyBac* element (Sarkar et al. 2003). The size of PGBD1 members varies from 312 (Rno-PGBD1) to 826 (Bbi-PGBD1) amino acids, suggesting one or several indels in the different parts of the transposase during evolution. Nevertheless, they present a high level of similarity (average similarity =85%). In the PGBD2 group ($n = 62$), a single uninterrupted exon corresponding to the *piggyBac* transposase is found. As mentioned by Sarkar et al. (2003) and Pavelitz et al. (2013), we observed two exons in 5′ of the transposase sequence. Based on all PGBD2 sequences extracted from the database, the length of the encoded proteins (uninterrupted exon) varies from 586 (Pal-PGBD2) to 759 (Lve-PGBD2) amino acids. The average similarity between the members of this group is 88%. The PGBD3 ($n = 15$) transposase is inserted into the fifth intron of the *Cockayne Syndrome group B* gene (*CSB*). Indeed, unlike other PGBD, the PGBD3 transposase is flanked by a potential 3′ splicing site in the 5′ region and a polyadenylation signal in the 3′ region. Thus, an alternative splicing of this region leads to a regular CSB product or to the CSB-PGBD3 fusion protein which has been conserved since the common ancestor of human and marmoset lineages, that is, 43 Ma (Newman et al. 2008). Moreover, it has been demonstrated that the CSB-PGBD3 protein regulates gene expression from AP1, TEAD, and CTCF sites but not from MER85 sites (Gray et al. 2012). The average similarity of the PGBD3 sequences retrieved in various primates is very high (98%).

PGBD4 sequences ($n = 16$) present a single ORF encoding for a protein of 585 aa. Only the product of the Ppa-PGBD4 from *Pan paniscus* has a longer size (603 aa). The average percentage of similarity between the members of this group is also very high (98%).

PGBD5 is the largest group ($n = 147$). It contains several introns (six in the *Bfl* gene of cephalochordates, seven in *Pma* of agnathes and six in all other vertebrates). According to Pavelitz et al. (2013) and assuming that all PGBD5 behave similarly, there is no alternative splicing and all introns are spliced. Furthermore, a potential polyadenylation signal is observed in 3′ region of gnathostomates. The length of the PGBD5 encoded protein varies from 343 aa (Eca-PGBD5 in *Equus caballus*) to 732 aa (Eed-PGBD5 in *Elephantulus edwardii*). This size variation seems to be due to a weak conservation of the N-terminal part of the protein (supplementary material S4, Supplementary Material online). Nonetheless, the similarity between these sequences with the N-terminal sequence is close to 76% and 87% without this region.

Two other domesticated elements present introns within their transposase-derived sequence. Indeed, *PGM* contains two introns with a coding region about 1,065 aa while *TPB2* contains 12 introns with an ORF coding for 1,220 aa. Finally, the *KOBUTA* element of *Xenopus* presents a single ORF with no introns encoding a protein of 610 aa.

The analysis of the *Ka/Ks* ratio provides arguments in favor of the domestication of these sequences. Indeed, this ratio calculated within each PGBD group varies from 0.123 ± 0.003 in PGBD2 to 0.432 ± 0.004 in PGBD1 (PGBD3 = 0.143 ± 0.005, PGBD4 = 0.150 ± 0.009, PGBD5 = 0.397 ± 0.002). This suggests the existence of a purifying or stabilizing selection. In this context, the DDD catalytic domain of the *piggyBac* transposase is not conserved among the domesticated sequences, including those retrieved from databases and those described in literature (Sarkar et al. 2003; Newman et al. 2008; Pavelitz et al. 2013), suggesting that this

Table 1

Structural and Putative Functions of Eight Domesticated piggyBac Elements

Gene ID	Name	References	Organism	Length of Coding Region (aa)	Presence of Introns	Catalytic Motif D D D D 1 2 3 4	CRD	Additional Domains or Genes	Functions
PGBD1	*piggyBac-derived 1*	Sarkar et al. 2003	*Homo sapiens*/ Mammals (61)	809[a,b] 312 < aa < 826	–	D D _ D G E	–	Zn_SCAN (290aa)	Unknown
PGBD2	*piggyBac-derived 2*	Sarkar et al. 2003	*Homo sapiens*/ Mammals (61)	592[a] 586 < aa < 759	–	G D G D _	+	–	Unknown
PGBD3	*piggyBac-derived 3*	Sarkar et al. 2003 Newman et al. 2008	*Homo sapiens*/ Primates (14)	593[a,c]	–	D N D D G	+	CSB gene	Involved in Cockayne syndrome
PGBD4	*piggyBac-derived 4*	Sarkar et al. 2003	*Homo sapiens*/ Primates (15)	585[a] 585 < aa < 603	–	D D D D	+	–	Unknown
KOBUTA	*KOBUTA*	Hikosaka et al. 2007	*Xenopus tropicalis, laevis, borealis*	610	–	D D N D N	+	–	Unknown
PGM	*piggyMac*	Baudry et al. 2009	*Paramecium tetraurelia*	1,065	2	D D D N	+	–	Required for programmed genome rearrangements
TPB2		Cheng et al. 2010	*Tetrahymena thermophila*	1,220	12				
PGBD5	*piggyBac-derived 5*	Sarkar et al. 2003 Pavelitz et al. 2013	*Homo sapiens*/ Myomerozoa (146)	554[a] 343 < aa < 732	5–7	_ _ _ D N	–	–	Neural specific

NOTE.—This table summarizes the information from several references including the organism (the number of sequences is in brackets), the length of the coding region, the presence and number of introns, the fate of the *piggyBac* catalytic motif using the four aspartate of PBLE as reference, presence/absence of Cysteine Rich Domain (CRD) in C-terminal region, existence of additional domains or part of genes, and putative functions.

[a]Refers to *Homo sapiens*.

[b]PGBD1 is composed by two parts. The first part, localized in 5', includes five exons encoding a sequence of 290 aa corresponding to a *Zn_Scan* domain. The second part, derived from the piggyBac transposase, encodes a sequence of 519 aa. Therefore, the total PGBD1 is a sequence of 809 aa of unknown function.

[c]PGBD3 is located in the fifth intron of the *CSB* gene, The *CSB-PGBD3* fusion encodes a protein of 1,061 aa including 468 residues of *CSB* and the entire transposase of PGBD3 (593 aa).

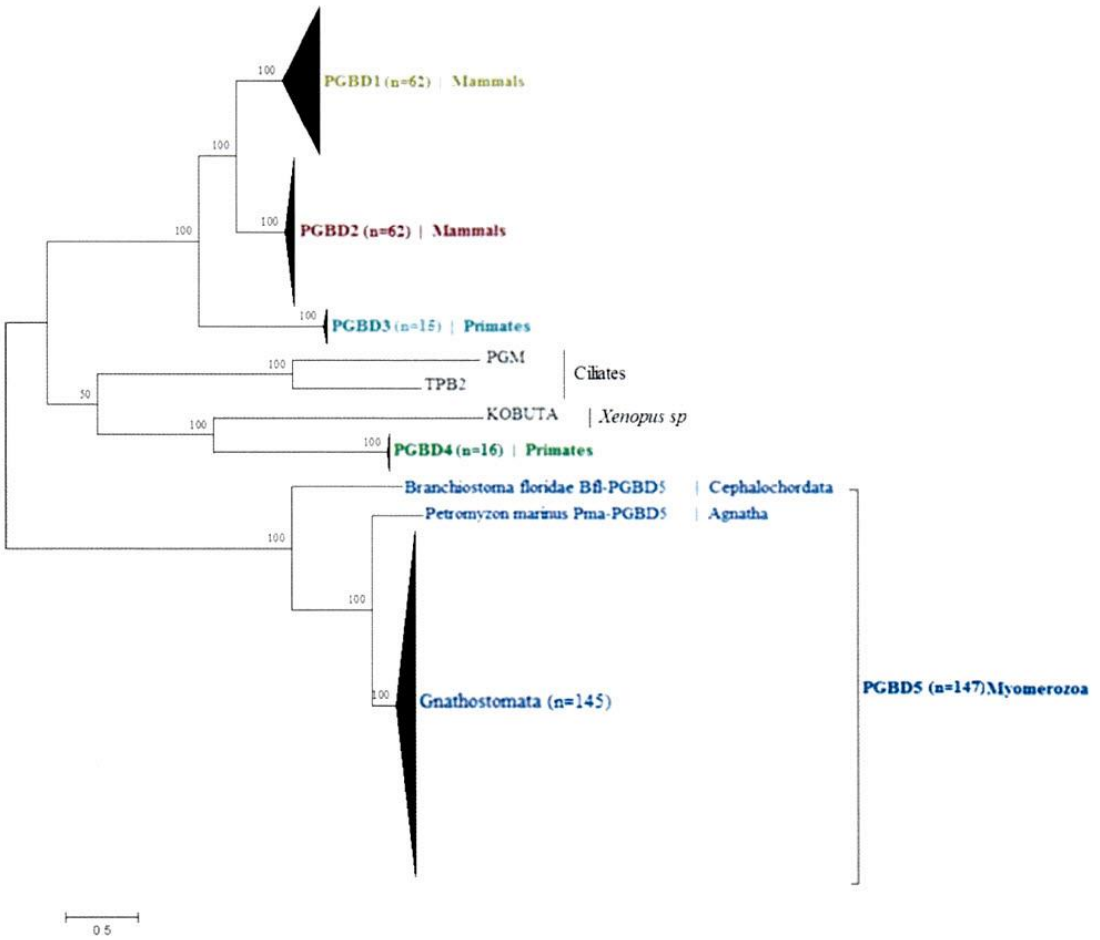


Fig. 4.—Phylogenetic tree of 'domesticated' elements. This phylogenetic tree comprises 10 PGBDs used as queries, including Human PGBD (Hsa-PGBD1, 2, 3, 4, 5), two orthologous to PGBD5 of Humans, namely *Pma* from the agnathe *Petromyzon marinus* and *Bfl* from the cephalochordate *Branchiostoma floridae*, *Tetrahymena thermophila* (*TPB2*) and *Paramecium tetraurelia* (*PGM*) *piggyMac*, *KOBUTA* from *Xenopus sp*, and 295 homologous PGBD sequences including 61 PGBD1, 61 PGBD2, 14 PGBD3, 15 PGBD4 and 144 PGBD5 sequences. Conserved transposase regions including 240 residues (see supplementary material S4, Supplementary Material online) were used to generate a maximum likelihood (ML) tree with LG + F + G matrix (best evolutionary scenario proposed by Protest 2.4). The number of sequences in the PGBDs groups is given in brackets.

characteristic was not selected for during all exaptation processes. This unconserved catalytic domain is particularly observed for the members of the PGBD5, which is probably the oldest domesticated sequence (table 1). In addition, in the PGBD1 and PGBD5 the C-terminal region is truncated, removing the ZnF motif.

The unrooted tree inferred from the 305 PGBDs (fig. 4) showed that the eight clades of PGBD are clearly identified and well supported (all bootstrap values are equal to 100). PGBD1 and PGBD2, exclusively found in mammals, are related to PGBD3 only present in primates, while PGBD4, also detected exclusively in primates, is closely related to *KOBUTA*. *PGM* and *TPB2* of ciliates are grouped together. PGBD5, found in a large spectrum of species, is quite distinct from the other PGBDs. As previously described (Pavelitz et al. 2013), this reflects an early domestication event. In this respect, its absence in echinoderms, hemichordates and urochordates (fig. 5), suggests a domestication event at least in the ancestor of the Myomerozoa lineage.

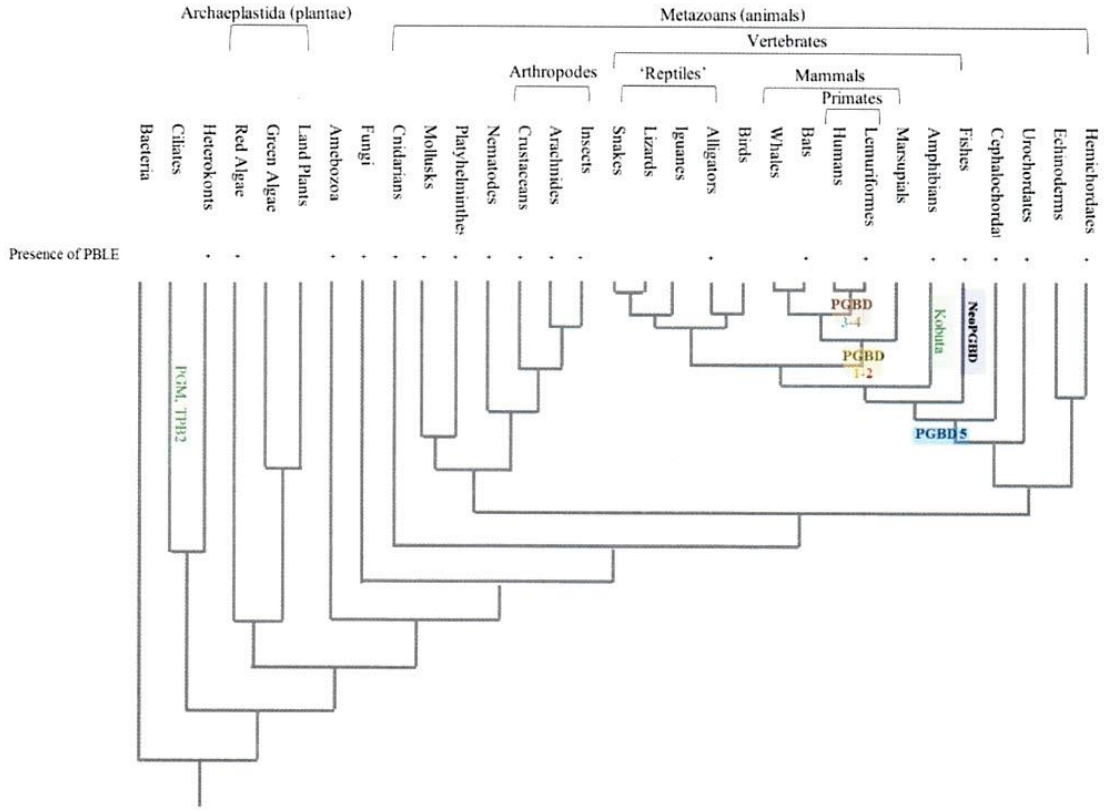


Fig. 5.—Distribution of PGBD sequences in eukaryotes. Each domesticated group of elements is highlighted by specific colors. PGBD1 and PGBD2 are found in mammals, PGBD3 and PGBD4 in primates, PGBD5 present a large spectrum of species belonging to Myomerozoa (including cephalochordates and all vertebrates). NeoPGBD is only specific to teleost fishes (Actinopterygii). *TPB2* and *PGM* are found in ciliates and *KOBUTA* in *Xenopus* sp. The presence of PBLE elements is mentioned by "+". The general phylogeny used here is redrawn from ref. http://www.talkorigins.org/faqs/comdesc/phylo.html#fig1 and has been modified to add some organisms.

Relationship between PGBD/PGBD-like and PBLE

In addition to the 157 PBLE (elements with both ends) and 305 PGBD identified previously, 101 *PGBD-like-elements* including eight sequences containing a single TIR (5′ or 3′) and 93 sequences with no TIR were detected (supplementary material S1, Supplementary Material online). In order to infer the relationship between all *piggyBac-related elements*, a maximum likelihood tree was built from the most conserved blocks. These blocks roughly correspond to the region surrounding the catalytic domain DDD. This provides an alignment of a total of 170 aa after concatenation (supplementary material S5, Supplementary Material online).

In this tree (fig. 6 and supplementary material S10, Supplementary Material online), PGBD1, PGBD2 and PGBD3 remain clustered compared to the tree of figure 4. Interestingly, three sequences—one from the aphid *Acyrthosiphum pisum* (Api-PGBD-like3) and two from the spider *Stegodyphus mimosarum* (SmiPBLE7 and Smi-PGBD-like3)—appear closely related to those of PGBD3; nevertheless, Api-PGBD-like3 and Smi-PGBD-like3, found in multiple copies with indels, appear not to be domesticated. The analysis of flanking regions of these ORFs (5 kb on both side) reveals in 5′ the presence of a potential 3′ splicing site (TTTTCTCTTCATATTTTTTTAG in SmiPBLE7 and Smi-PGBD-like3, TTTTTACTAGTTTTAG in Api-PGBD-like3 and CCTTTTTTCCGTTTTAG in PGBD3) and in 3′, a potential polyadenylation signal (AATAAA). These observations suggest that all these sequences share a common ancestor. Moreover, the 3′ splicing site (TTTTTTCTGTGTTAATATCTAG) and polyadenylation signal are also found in two other lepidopteran sequences (piggyBac-2_Hmel from *Heliconius melpomene* and CsuPBLE

from *Chilo suppressalis*), closely related to the three groups of domesticated elements PGBD1, 2 and 3. According to the phylogenetic analysis (fig. 6), these two sequences diverged before the emergence of the three groups. The most parsimonious explanation is that the splicing site and the polyadenylation signal were present in the common ancestor of all these sequences and were then lost along the branches leading to PGBD1 and PGBD2, while being retained in the clade containing PGBD3. Outside of this clade, these motifs can be present but with a patchy distribution.

The *KOBUTA* element forms a robust cluster with two other elements of *Xenopus* (*Uribo1* and *Uribo2*). The PGBD4 and the PvaPBLE of the chiropteran *Pteropus vampyrus* group together (similarity = 96% and *Ka/Ks* = 0.23±0.01). PGBD5 is always distinct as previously described. Nonetheless, it must be stressed that a PGBD-like sequence of a hemichordata (Sko-PGBD-like5) seems closely related to this group (supplementary material S10, Supplementary Material online). However, this sequence is intronless and presents several deletions and substitutions. While degenerated, this sequence was kept in our analysis since this is the only one close to the PGBD5 group. But, it must be stressed that this sequence is probably no longer functional as PBLE nor domesticated as PGBD5 members.

In the complete tree inferred from PBLE, PGBD like, and PGBD, three distinct groups (G1, G2, and G3) are specific to Actinopterygii species (supplementary material S10, Supplementary Material online). On the one hand, the first two contain complete or partial transposase of PBLE or PGBD-like copies that can be present in several copies in each species. On the other hand, the third group corresponds to sequences found as single copy in 14 teleost fish species (supplementary material S1, Supplementary Material online). All these sequences are annotated PGBD-like4 in database. No ortholog is detected in other vertebrates. This group, here named NeoPGBD, presents a conserved putative ORF with few substitutions or gaps but with no frameshift or non-sense mutation (supplementary material S8, Supplementary Material online). The only exception is for Ali-PGBD-like4 showing a stop codon. The ORF length varies from 649 to 682 aa with an average similarity of 86% and *Ka/Ks* value of 0.094 ± 0.004. This is consistent with strong purifying selection acting on this insertion. On one hand, no promoter, no splicing signal and no binding site of transcription factor is found in the 5′ region, but a highly conserved sequence of 365 bp (identity > 90%) of unknown function is located immediately upstream the ORF. On the other hand, a potential polyadenylation signal AATAAA seems to be conserved in the 3′ region except in Lpe-PGBD-like4. When the analysis of the flanking regions is extended (5 kb on each side), the identity level remains relatively high (65% in 5′ and 70% in 3′), and no gene or associated domain can be detected.

Discussion

A large diversity of transposons belonging to the *piggyBac* superfamily has been documented in eukaryotes, except in plants, from data mining in databanks. This study provides a detailed characterization of these elements, their specific distribution and their possible evolution.

A total of 157 PBLE, including 50 new sequences, were analyzed. The target site is TTAA. Rarely, for only five elements (unique copy), other TSD can be observed (supplementary material S2, Supplementary Material online). This is probably due to mutations, since the 5′ TSD and 3′ TSD sequences are not the same. In only two cases (Yabusame1, AcePBLE), the two TSD are identical, but Yabusame1 does not seem active in spite of a putative intact ORF (Daimon et al. 2010).

TIRs of PBLE appear to be divergent both in length and sequence except for the first highly conserved cytosine. TIRs play an essential role in transposase recognition and cleavage from the target site (Elick et al. 1997; Li et al. 2001). Mutagenesis experiments with the *piggyBac* element confirmed that the 3′ terminal G plays an important role in the selection of the excision site and that the deletion of a single 3′G nucleotide from one of the termini is sufficient to abolish excision (Elick et al. 1997). For other transposons, similar examples also indicate that mutations of the first two base pairs of TIR lead to defective excision processes (Haniford and Kleckner 1994). In this study, we show that the first three residues of TIR were not usually CCC/GGG but **C[C**/A/T] [C/A/T]. Nevertheless, these copies are functional with an excision activity (Xu et al. 2006; Hikosaka et al. 2007; Luo et al. 2011; Mitra et al. 2013). Therefore, the functional impact of the second and third residues of TIR is not so clear since they are not always responsible for defects in the excision process.

Structural architecture of PBLE is also variable and the occurrence of DRs close to TIR was previously reported (Wu and Wang 2014). In the present work, four SGs were identified according to the presence/absence of DR and/or STIR. This suggests that these elements are structurally highly flexible. Two hypotheses can be proposed to explain the specific distribution of the different structures. On one hand, the SG1 (with only two TIRs) is the ancestral structure, while the other ones derive from the ancestral sequence by independent acquisitions of DR and STIR (convergence) and/or by horizontal transfers. On the other hand, the four SGs were already present in the common ancestor of all species in which PBLE copies have been found, and their patchy distribution is due to independent loss in several lineages. However, the absence of consensus sequences between these regions and the incongruence with the tree derived from transposases are in favor of independent acquisitions (convergences) and rapid evolution of these regions. The inconsistencies observed between the transposase trees, the element structures, and with the species phylogeny also suggest that *piggyBac* might be frequently and successfully horizontally transferred. This may

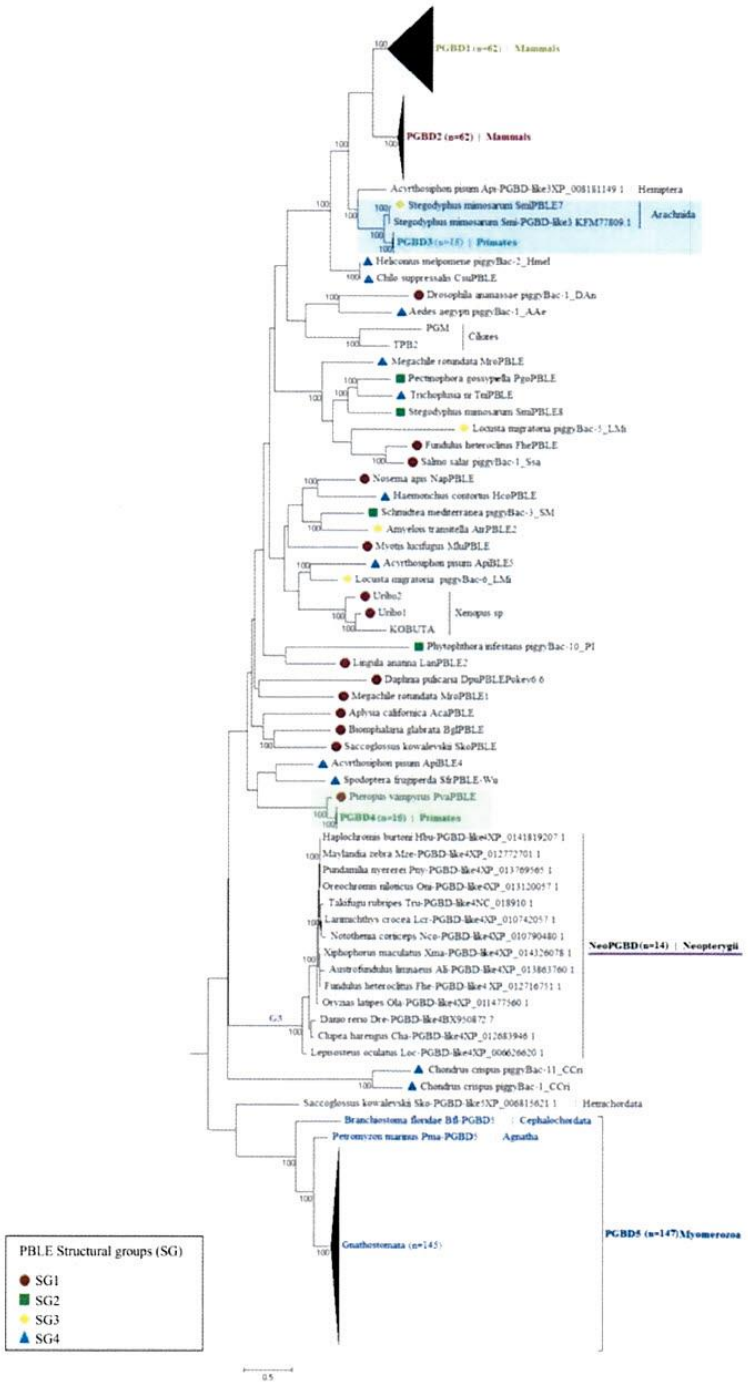


FIG. 6.—Phylogenetic tree of *piggyBac* superfamily. This phylogeny is based on amino acid sequences covering about 170 residues (supplementary material S5, Supplementary Material online). For simplicity, only 355 sequences are represented, including 305 PGBD sequences, 17 PGBD-like sequences

be due to a "host factors independent" activity, and could explain why this element is a powerful vector for genome engineering.

Furthermore, the presence of internal repeats (DR/STIR), sometimes with palindromic motifs, raises several questions. Are these regions involved in the element activity *via* the transposase binding and stabilization of the transposase-TIR/DR/STIR complex? This probably reflects a rapid coevolution between the transposase and the different terminal repeated sequences, but this remains to be functionally demonstrated. For the *mariner-like elements* (MLE), it was shown that conserved palindromic and mirror motifs within TIR, are important features of the transposase-TIR interaction (Bigot et al. 2005).

PBLE transposase includes a $D^{268}D^{346}D^{447}D^{450}$ catalytic domain and a Cysteine Rich Domain (CRD) in C-terminal region (Keith et al. 2008). An additional Histidine can be observed in some sequences between the fifth and the sixth Cysteines (Fraser et al. 1983; Arkhipova and Meselson 2005). However, while Cysteines are found in almost all sequences and are relatively well conserved, the presence and position of the Histidine is highly variable. Thus, Histidine impact on the transposase structure and functionality is not clear.

Like all TEs, *piggyBac* are prone to indels, recombinations and mutations that inactivate many copies, which can be rapidly lost in absence of *trans*-mobilization (Robertson 1993; Le Rouzic et al. 2007; Hua-Van et al. 2011). Nevertheless, molecular domestication events may occur, in which TE-insertions are advantageous, maintained by natural selection and turned into regular genes (Le Rouzic et al. 2007; Sinzelle et al. 2009; Hua-Van et al. 2011). Such sequences can be distinguished from defective copies from a low ratio *Ka/Ks*, indicating that they evolve under purifying selection and do not correspond to pseudogenes (Sarkar et al. 2003; Newman et al. 2008; Pavelitz et al. 2013). Otherwise, in the present study, some evidence suggests a new domestication event of a *piggyBac* element. This new group is exclusively found in Neopterygii (NeoPGBD). This sequence is present in single copy per genome and its ORF is well conserved in several species orders, suggesting an ancient domestication in Neopterygii, that is, at least 250 Ma (Betancur et al. 2013). Similarly, to the previous PGBD groups already described, the *Ka/Ks* ratio is low. Moreover, the high conservation of the region upstream the sequence of these PGBD suggests the existence of selective pressures to maintain a putative function, which remains unknown.

The presence of these exaptations raises the question of their emergence: what is the origin of the PGBD sequences? Answering this question remains difficult, but several hypotheses can be proposed. First, several PBLEs seems to be potentially active, or at least they were within the recent past. Second, the general phylogeny, including PBLE, PGDB and PGDB-like, shows that a close relationship can be observed between PGBD and PBLE or PGBD-like. For instance, PvaPBLE of *Pteropus vampyrus* is closely related to the members of PGBD4. The PBLE (*Uribo1* and *Uribo2*) and *KOBUTA* also form a robust clade. Similarly, with PGBD3 members, SmiPBLE7 and Smi-PGBD-like3 of *Stegodyphus mimosarum* and Api-PGBD-like3 of *Acyrthosiphon pisum*, are grouped together.

For PGBD5, the domestication probably occurred along the branch leading to the Myomerozoa, as described by Pavelitz et al. (2013), based on transposase sequence, intron location, and microsynteny. However, an earlier domestication, followed by a loss of this element in all branches except one leading to Myomerozoa, cannot be excluded. To check this hypothesis, a BlastN search, using PGBD5 and its flanking sequences as query, in urochordates (*Ciona intestinalis, Oikopleura dioica*), echinoderms (*Strongylocentrotus purpuratus, Acanthaster planci*) and hemichordate (*Saccoglossus kowalevskii*), do not allow us to detect PGBD5-like sequences with introns. Indeed, only a few fragments of intronless *piggyBac–like* ORF can be found. Moreover, no sequence similar to those of the flanking region of PGBD5 can be found in hemichordate, echinoderms, and urochordates. From these observations, and assuming that the genome assemblies of these species are correct, an evolutionary scenario can be proposed. The insertion at the origin of PGBD5 occurred in the ancestor of the Myomerozoa. This initial insertion rapidly acquired a new function and was probably under a high selective pressure leading to a selective sweep on the entire region including the flanking sequences.

Therefore, based on the phylogenetic proximities between PBLE and PGBD sequences, and because domesticated sequences are distinct monophyletic groups, we can hypothesize that each PGBD group (including NeoPGBD), as well as *PGM*, *TPB2 and KOBUTA* genes, derive from a single and specific ancestral PBLE sequence, suggesting that the domestication event at the origin of each group occurs once.

The evolutionary trajectory of PGBD is not systematically accompanied by modifications of transposase activity (see, for instance, Sarkar et al. 2003; Pavelitz et al. 2013). In this

Fig. 6.—Continued

and 33 PBLE sequences. Maximum likelihood (ML) method, with LG + F + G matrix, is used to construct this tree. Only bootstrap values (100 replications) higher than 70% are labeled. Groups including PGBD elements and PBLE and/or PGBD-like, are framed in colors. Red dots, green squares, yellow lozenges, and blue triangles refer to the different PBLE SGs, that is, SG1, SG2, SG3, and SG4, respectively. The putative domesticated sequences (NeoPGBD), found only in Actinopterygii, are underlined in purple. Sequence names: the term PBLE is used for sequences extracted from literature, or copies newly characterized in this study, while the term *piggyBac* is restricted to sequences extracted from Repbase (real name of these sequences in this database). *TPB2* is from *Tetrahymena thermophila*, *PGM* from *Paramecium tetraurelia* and *KOBUTA* from *Xenopus sp.*

respect, *piggyMac*, *TPB2* and *KOBUTA* have conserved the DDD catalytic domain and the C-terminal region of the PBLE and are involved in excision mechanisms. This DDD motif is also found in the members of the PGBD4 group, but there is no proof that this is associated to a transposase activity (Mitra et al. 2008). For PGBD3, while the DDD motif is not strictly conserved, the C-terminal transposase domain of the human CSB-PGBD3 fusion protein is able to mobilize the MER85 (Gray et al. 2012). Moreover, the protein encoded by human PGBD5 seems to be involved in stereotypical cut-and-paste DNA transposition in human cells, but in this case, the genomic integration required distinct aspartic acid residues, and specific DNA sequences (including TIR) compared to those of *Uribo2*, *piggyBac*, *piggyMac* and *piggyBat* (Henssen et al. 2015). For the remaining PGBD (1 and 2), the DDD motif is not conserved and no mobilization activity can be suspected. Therefore, a switch to new unknown host functions, since these sequences are under purifying selection, probably occurred.

Several cases of true or putative exaptation of TEs have been reported in all domains of the tree of life (see, for instance, Hoen and Bureau, 2012 for exaptation in plants). However, it seems that all families of TEs are not "equal" in terms of their fate, since some of them tend to be more prone to be domesticated than others. Is it a relevant observation or a sampling effect? Such an observation has been recently made for the members of the *Mutator*-like superfamily (Joly-Lopez et al. 2016). Are the members of the *piggyBac* superfamily prone to such fate? If so, what could be the reason for this? Is it due to specific internal features, a transpositional mechanism and/or particular genomic locations of these elements?

While several features of *piggyBac* can be listed, it remains difficult to know which one(s) are responsible for their domestication success. For instance, full-length copies of *piggyBac* can be found in a large spectrum of species (this work), so this element can potentially move in a large set of organisms and excise precisely without host damage (see the references in Mitra et al. 2008). In addition, according to Newman et al. (2008), *piggyBac* could be a natural "exon trap" as shown for PGBD3 since a potential 3′ splicing site and a polyadenylation signal can be detected. These allow its insertion into *CSB* intron 5, generating an N-terminal fusion protein. Moreover, the analysis of the 5′ flanking regions PGBD1 and PGBD2 reveals that five and two exons, respectively, derived from the host gene (supplementary material S9, Supplementary Material online). Again, *TniPBLE* transposase is able to tolerate N-terminal fusion and to retain a significant transposition activity. In this respect, this element seems more flexible than *Sleeping Beauty, Tol2 and Mos1* transposases (Wu et al. 2006). Based on the "3′ exon trap" hypothesis, Newman et al. (2008) also suggest that *piggyBac* could benefit from the efficient host promoter (see also Gray et al. 2012). All these characteristics may be the reason for its colonizing success and its capacity to be domesticated.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Author Contributions

MB, MM, and PC conceived and designed research. MB performed research. MB, AHV, MM, and PC contributed analysis tools. JDR participated to the sequences analysis. MB, AHV, MM, and PC wrote the article.

Acknowledgments

This work was financially supported by the Centre National de la Recherche Scientifique [UMR 9191], the University Paris-Sud, the Tunisian Ministry of Higher Education and Scientific Research and the University of Tunis El Manar. Authors thank Mireille Bétermier and Julien Bischerour for their helpful comments and Malcolm Eden for the English review of the manuscript.

Literature Cited

Arkhipova IR, Meselson M. 2005. Diverse DNA transposons in rotifers of the class Bdelloidea. Proc Natl Acad Sci U S A. 102 (33):11781–11786.

Baudry C, et al. 2009. *PiggyMac*, a domesticated *piggyBac* transposase involved in programmed genome rearrangements in the ciliate *Paramecium tetraurelia*. Genes Dev. 23:2478–2483.

Betancur R.R, et al. 2013. The tree of life and a new classification of bony fishes. PLOS Curr Tree Life. Edition 1.5. doi: 10.1371/currents.tol.53ba26640df0ccaee75bb165c8c26288.

Bigot Y, Brillet B, Augé-Gouillou C. 2005. Conservation of palindromic and mirror motifs within inverted terminal repeats of *mariner*-like elements. J Mol Biol. 351:108–116.

Britten RJ. 1996. DNA sequence insertion and evolutionary variation in gene regulation. Proc Natl Acad Sci U S A. 93:9374–9377.

Carpes MP, et al. 2009. Molecular analysis of a mutant *Anticarsia gemmatalis* multiple nucleopolyhedrovirus (AgMNPV) shows an interruption of an inhibitor of apoptosis gene (*iap-3*) by a new class-II *piggyBac*-related insect transposon. Insect Mol Biol. 18:747–757.

Cary LC, et al. 1989. Transposon mutagenesis of baculoviruses: analysis of *Trichoplusia ni* transposon *IFP2* insertions within the *FP*-locus of nuclear polyhedrosis viruses. Virology 172:156–169.

Casola C, Hucks D, Feschotte C. 2008. Convergent domestication of *pogo*-like transposases into centromere-binding proteins in fission yeast and mammals. Mol Biol Evol. 25:29–41.

Cheng CY, Vogt A, Mochizuki K, Yao MC. 2010. A domesticated *piggyBac* transposase plays key roles in heterochromatin dynamics and DNA cleavage during programmed DNA deletion in *Tetrahymena thermophila*. Mol Biol Cell. 21:1753–1762.

Cordaux R, Udit S, Batzer MA, Feschotte C. 2006. Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. Proc Natl Acad Sci U S A. 103:8101–8106.

Daimon T, et al. 2010. Recent transposition of *yabusame*, a novel *piggyBac*-like transposable element in the genome of the silkworm, *Bombyx mori*. Genome 53:585–593.

Elick TA, Lobo N, Fraser MJ. 1997. Analysis of cis-acting DNA elements required for *piggyBac* transposable element excision. Mol Gen Genet. 255:605–610.
Feschotte C, Zhang X, Wessler SR. 2002. Miniature inverted-repeat transposable elements (MITEs) and their relationship with established DNA transposons. In: Craig N, Craigie R, Gellert M, Lambowitz A, editors. Mobile DNA II. Washington, D.C. : American Society of Microbiology Press. p. 1147–1158.
Fraser MJ, Smith GE, Summers MD. 1983. Acquisition of host cell DNA sequences by baculoviruses: Relationship between host DNA insertions and FP mutants of *Autographa californica* and *Galleria mellonella* nuclear polyhedrosis viruses. J Virol. 47:287–300.
Fraser MJ, Ciszczon T, Elick T, Bauser C. 1996. Precise excision of TTAA specific lepidopteran transposons *piggyBac* (IFP2) and tagalong (TFP3) from the baculovirus genome in cell lines from two species of Lepidoptera. Insect Mol Biol. 5:141–151.
Gibbs RA, et al. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428:493–521.
Gray LT, Fong KK, Pavelitz T, Weiner AM. 2012. Tethering of the conserved *piggyBac* transposase fusion protein CSB-PGBD3 to chromosomal AP-1 proteins regulates expression of nearby genes in humans. PLoS Genet. 8(9):e1002972.
Haniford D, Kleckner N. 1994. Tn*10* transposition in vivo: temporal separation of cleavages at the two transposon ends and roles of terminal basepairs subsequent to interaction of ends. EMBO J. 13:3401–3411.
Hencken CG, Li X, Craig NL. 2012. Functional characterization of an active *Rag*-like transposase. Nat Struct Mol Biol. 19:834–836.
Henssen AG, et al. 2015. Genomic DNA transposition induced by human PGBD5. Botchan MR, ed. eLife, 4:e10565. doi:10.7554/eLife.10565.
Hikosaka A, Kobayashi T, Saito Y, Kawahara A. 2007. Evolution of the *Xenopus piggyBac* transposon family *TxpB*: domesticated and untamed strategies of transposon subfamilies. Mol Biol Evol. 24:2648–2656.
Hoen DR, Bureau T. 2012. Transposable element exaptation in plant. In Grandbastien MG, Casacuberta JM, editors. Plant transposable elements: impact on genome, structure and function. Topics in Current Genetics. p. 219-251.
Hua-Van A, Le Rouzic A, Boutin TS, Filée J, Capy P. 2011. The struggle for life of the genome's selfish architects. Biol Direct. 6(1):19–47.
Hurst LD. 2002. The Ka/Ks ratio: diagnosing the form of sequence evolution. Trends Genet. 18(9):486–487.
Joly-Lopez Z, Hoen DR, Blanchette M, Bureau TE. 2016. Phylogenetic and genomic analyses resolve the origin of important plant genes derived from transposable elements. Mol Biol Evol. doi:10.1093/molbev/msw067.
Kapitonov VV, Jurka J. 2005. RAG1 core and V(D)J recombination signal sequences were derived from *Transib* transposons. PLoS Biol. 3:e181.
Keith JH, Schaeper CA, Fraser TS, Fraser MJ Jr. 2008. Mutational analysis of highly conserved aspartate residues essential to the catalytic core of the *piggyBac* transposase. BMC Mol Biol. 9:73–92.
Kidwell MG, Lisch D. 1997. Transposable elements as sources of variation in animals and plants. Proc Natl Acad Sci U S A. 94:7704–7711.
Kidwell MG. 2002. Transposable elements and the evolution of genome size in eukaryotes. Genetica 115:49–63.
Lander ES, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.
Larsson A. 2014. AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics 30 (22):3276–3278.
Le Rouzic A, Boutin TS, Capy P. 2007. Long-term evolution of transposable elements. Proc Natl Acad Sci U S A. 104:19375–19380.
Li X, Lobo N, Bauser CA, Fraser MJ. 2001. The minimum internal and external sequence requirements for transposition of the eukaryotic transformation vector piggyBac. Mol Genet Genom. 266:190–198.
Lindblad-Toh K, et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438:803–819.
Liu, et al. 2007. The human SETMAR protein preserves most of the activities of the ancestral Hsmar1 transposase. Mol Cell Biol. 27(3):1125–1132.
Lobo N, Li X, Fraser MJ. 1999. Transposition of the *piggyBac* element in embryos of *Drosophila melanogaster*, *Aedes aegypti* and *Trichoplusia ni*. Mol Gen Genet. 261:803–810.
Luo GH, Wu M, Wang XF, Zhang W, Han ZJ. 2011. A new active *piggyBac*-like element in *Aphis gossypii*. Insect Sci. 18(6):652–662.
Luo GH, et al. 2014. Molecular characterization of the *piggyBac*-like element, a candidate marker for phylogenetic research of *Chilo suppressalis* (Walker) in China. BMC Mol Biol. 15:28–39.
Miller WJ, McDonald JF, Nouaud D, Anxolabéhère D. 1999. Molecular domestication – More than a sporadic episode in evolution. Genetica 107:197–207.
Mitra R, Fain-Thornton J, Craig NL. 2008. *piggyBac* can bypass DNA synthesis during cut and paste transposition. EMBO J. 27:1097–1109.
Mitra R, et al. 2013. Functional characterization of *piggyBat* from the bat *Myotis lucifugus* unveils an active mammalian DNA transposon. Proc Natl Acad Sci U S A. 110:234–239.
Newman JC, Bailey AD, Fan HY, Pavelitz T, Weiner AM. 2008. An abundant evolutionarily conserved CSB-*piggyBac* fusion protein expressed in Cockayne syndrome. PLoS Genet. 4(3):e1000031.
Pace JK, Feschotte C. 2007. The evolutionary history of human DNA transposons: evidence for intense activity in the primate lineage. Genome Res. 17:422–432.
Pagan HJ, Smith JD, Hubley RM, Ray DA. 2010. PiggyBac-ing on a primate genome: novel elements, recent activity and horizontal transfer. Genome Biol Evol. 4:293–303.
Pavelitz T, Gray LT, Padilla SL, Bailey AD, Weiner AM. 2013. PGBD5: a neural-specific intron containing *piggyBac* transposase domesticated over 500 million years ago and conserved from cephalochordates to humans. Mobile DNA 4:23–39.
Penton EH, Sullender BW, Crease TJ. 2002. Pokey, a new DNA transposon in Daphnia (cladocera: crustacea). J Mol Evol. 55:664–673.
Pritham EJ, Feschotte C, Wessler SR. 2005. Unexpected diversity and differential success of DNA transposons in four species of *Entamoeba* protozoans. Mol Biol Evol. 22(9):1751–1763.
Ray DA, et al. 2008. Multiple waves of recent DNA transposon activity in the bat, *Myotis lucifugus*. Genome 18:717–728.
Robertson HM. 1993. The *mariner* transposable element is widespread in insects. Nature 362:241–245.
Sarkar A, et al. 2003. Molecular evolutionary analysis of the widespread *piggyBac* transposon family and related "domesticated" sequences. Mol Genet Genom. 270(2):173–180.
Sinzelle L, Izsvák Z, Ivics Z. 2009. Molecular domestication of transposable elements: from detrimental parasites to useful host genes. Cell Mol Life Sci. 66:1073–1093.
Sun ZC, Wu M, Miller TA, Han ZJ. 2008. *piggyBac*-like elements in cotton bollworm, *Helicoverpa armigera* (Hubner). Insect Mol Biol. 17:9–18.
Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol. 30:2725–2729.
Volff JN. 2006. Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. BioEssays 28:913–922.
Wang J, Ren X, Miller TA, Park Y. 2006. *piggyBac*-like elements PLE in the tobacco budworm, *Heliothis virescens* (Fabricius). Insect Mol Biol. 15:435–443.
Wang JJ, Du YZ, Wang SZ, Brown SJ, Park Y. 2008. Large diversity of the *piggyBac*-like elements in the genome of *Tribolium castaneum*. Insect Biochem Mol Biol. 38(4):490–498.
Wang J, et al. 2009. *piggyBac*-like elements in the pink bollworm, *Pectinophora gossypiella*. Insect Mol Biol. 19:177–184.

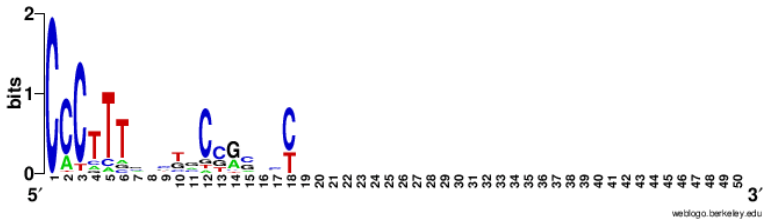
Wicker T, et al. 2007. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 8:973–982.
Wu C, Wang S. 2014. PLE-wu, a new member of *piggyBac* transposon family from insect, is active in mammalian cells. J Biosci Bioeng. 118(4):359–366.
Wu M, Sun ZC, Hu CL, Zhang GF, Han ZJ. 2008. An active *piggyBac*-like element in *Macdunnoughia crassisigna*. Insect Sci. 15:521–528.
Wu M, et al. 2006. piggyBac is a flexible and highly active transposon as compared to sleeping beauty, Tol2, and Mos1 in mammalian cells. Proc Natl Acad Sci U S A. 103(41):15008–15013.
Wu M, et al. 2011. Cloning and characterization of *piggyBac*-like elements in lepidopteran insects. Genetica 139(1):149–154.
Xu HF, et al. 2006. Identification and characterization of *piggyBac*-like elements in the genome of domesticated silkworm, *Bombyx mori*. Mol Genet Genom. 276:31–40.
Yusa K. 2015. *piggyBac* transposon. Microbiol Spectr. 3(2): MDNA3-0028-2014. doi:10.1128/microbiolspec.MDNA3-0028-2014.

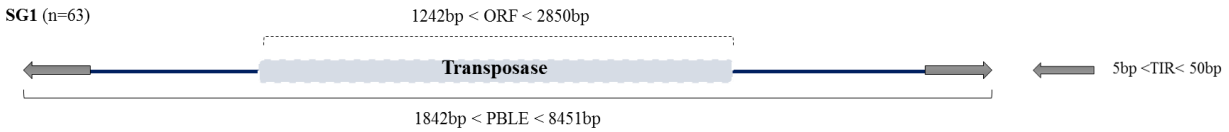
Associate editor: Richard Cordaux

Supplemental data 2. Structure of different PBLEs. Group 1: PBLE sequences with Terminal Inverted Repeats (TIR) only. Group 2: PBLE sequences with TIR and Sub-Terminal Inverted Repeats (STIR). Group 3: PBLE sequences with TIR and Direct Repeats (DR). Group 4: PBLE sequences with TIR, DR and STIR. TIR and STIR are written as 5' to 3'. The underlined region corresponds to the same position of DR and STIR (presence of palindromic sequences in some cases). Sequences in italic correspond to the overlapping region between TIR/STIR or TIR/DR. TIR consensus per group and for all groups was generated using the Web-Logo server (http://weblogo.berkeley.edu/logo.cgi). At each position the nucleotides are stacked one on top of another with the most frequent one on the top. It displays the frequency of bases at each position with height indicating the proportion of occurrence. The vertical scale is in bits with maximum of two bits possible at each position indicating that there can be possibility of four different bases at each position.

Group 1*: 63 PBLE sequences including Terminal Inverted Repeats (TIR).*

Consensus TIR





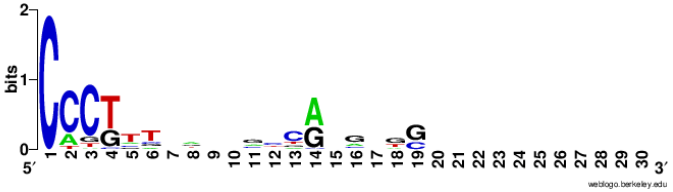
PBLE (bp)		Species	TSD+TIR		References
NapPBLE	3125	*Nosema apis*	TTAA	CACAT	In this study
SmiPBLE3	2581	*Stegodyphus mimosarum*	TTAA	CCCTTTC	In this study
piggyBac-7_PI	3405	*Phytophthora infestans*	TTAA	CACCTTGC	Repbase
SkoPBLE	3104	*Saccoglossus kowalevskii*	TTAA	CCCTTTGC	In this study
LEAPFROG1_EI	1958	*Entamoeba invadens*	TTAA	CCTTAGGAAG	Pritham et al. 2005
piggyBac-4_SM	3161	*Schmidtea mediterranea*	TTAA	CCCTTTGAGG	Repbase
piggyBac-1_NV	2384	*Nematostella vectensis*	TTAA	CCCTTTCACTA	RepBase
CbiPBLE	2208	*Cerapachys biroi*	TTAA	CCTCTTAGgTAC CCTCTTAGcTAC	In this study
piggyBac-3_LMi	2403	*Locusta migratoria*	TTAA	CCCTTTCACTGC	Repbase
LanPBLE2	4030	*Lingula anatina*	TTAA	CCCCATaAGGCC CCCCAT-AGGCC	In this study
piggyBac1_CI	2216	*Ciona intestinalis*	TTAA	CACTTTCGGCGCC	RepBase
BmoPBLE25	2559	*Bombyx mori*	TTAA	CACGTTAAACGCC	Xu et al. 2006
AtrPBLE1	2692	*Amyelois transitella*	TTAA	CACGTTCATCGCC	In this study
AroPBLE1	2358	*Athalia rosae*	TTAA	CACGTTGAACGCC	In this study
LanPBLE3	3133	*Lingula anatina*	TTAA	CTCATTGTCCCCT	In this study
OabPBLE2	2264	*Orussus abietinus*	TTAA	CCACTTCGGTACG	In this study

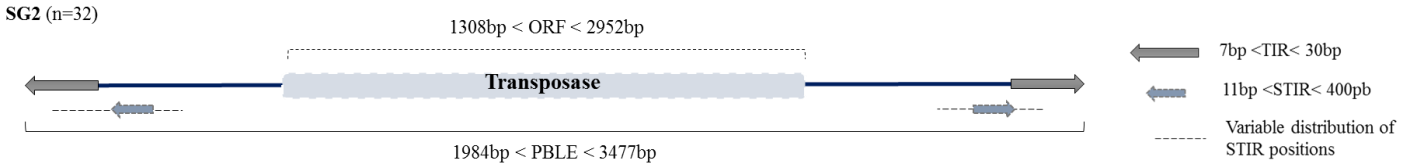
			TTTT		
AvPB2_1p	2333	*Adineta vaga*	TTAA	CCCTAGTACTCCC	Repbase
piggyBac-1_CGi	2350	*Crassostrea gigas*	TTAA	CCCTTAGaCTGCT CCCTTAGgCTGCT	Repbase
piggyBac-2_CGi	4451	*Crassostrea gigas*	TTAA	CCCTTGCTCTGCT	RepBase
piggyBac-1_DAn	2735	*Drosophila ananassae*	TTAA	CCCTTTATATGGC	RepBase
piggyBac-3_HM	2883	*Hydra magnipapillata*	TTAA	CCCTTTCaCTCCC CCCTTTCtCTCCC	RepBase
AmiPBLE1	3668	*Acropora millepora*	TTAA	CCCTTTCCCGTCC	Wang et al. 2010
piggyBac1_Mm	2527	*Microcebus murinus*	TTAA	CCCTTTGCACTCG	Repbase
LanPBLE1 AroPBLE2	3596 2384	*Lingula anatina* *Athalia rosae*	TTAA	CCCTTTGCGGACG	In this study
piggyBac-13_SM	2554	*Schmidtea mediterranea*	TTAA	CACATTGTaATCGG CACATTGTgATCGG	Repbase
piggyBac-7_SM	2588	*Schmidtea mediterranea*	TTAA	CACTAGATTTACCG	Repbase
BglPBLE	2313	*Biomphalaria glabrata*	TTAA	CTCATTAGCTACTA	In this study
piggyBac-3_CGi	2305	*Crassostrea gigas*	TTAA	CTCATTCAGCCCCA	RepBase
Uribo1	6169	*Xenopus sp*	TTAA	CCYTTTNNNTGCCA TGGCANNNAAAYGG	Hikosaka et al. 2007
Uribo2	8451	*Xenopus tropicalis*	TTAA	CCYTTTNNNTGCCA CCRTTTNNNTGCCA	Hikosaka et al. 2007
DpuPBLE-Pokey	6574	*Daphnia pulicaria*	TTAA	CCCTTTtTCGACtG CCCTTTaTCGACcG	Penton et al. 2002
DpuPBLE-Pokey	5058	*Daphnia pulicaria*	TTAA	CCCTTTtTCGACtG CCCTTTaTCGACcG	Penton et al. 2002
PvaPBLE	2176	*Pteropus vampyrus*	TTAA	CCCATTTCCTGaTT CCCATTTCCTGtTT	In this study
piggyBac-3_SSa	3392	*Salmo salar*	TTAA	CCCTCCCGTTGTCC	Repbase
piggyBac-4_Lmi	3161	*Locusta migratoria*	TTAA	CCCTTTgaCTGCTG CCCTTTagCTGCTG	Repbase
SmiPBLE1	2567	*Stegodyphus mimosarum*	TTAA	CCCTTTcCGGACGA CCCTTTgCGGACGA	In this study
SmiPBLE2	2639	*Stegodyphus mimosarum*	TTAA	CCCTTTCGTGaGCC CCCTTTCGTGgGCC	In this study
MroPBLE1	2840	*Megachile rotundata*	TTAA	CCCTTTGCGATCGG	In this study
PxuPBLE1	2544	*Papilio xuthus*	TTAA	CCCTTTGACTGCGG	In this study
SmiPBLE4	2965	*Stegodyphus mimosarum*	TTAA	CCCTTTGAcGCGCG CCCTTTGAaGCGCG	In this study
LanPBLE	2866	*Lingula anatina*	TTAA	CCCTCAAGtGACCG CCCTCAAGaGACCG	In this study
piggyBac2_Mm	2211	*Microcebus murinus*	TTAA	CaCATTgCvTACCdC CgCATTaCgTACCgC	Repbase
MluPBLE	2626	*Myotis lucifugus*	TTAA	CACttgGaTtgCGGG CACcacGgTgtCGGG	Ray et al. 2008
OniPBLE	2522	*Oreochromis niloticus*	**ATAG** **ATAA**	CCTCCTGACTACCAG	In this study
AcaPBLE	2744	*Aplysia californica*	TTAA	CCCTTACAGCCCTGT	In this study
piggyBac-1_Mcir	2026	*Mucor circinelloides*	TTAA	CCCTTAAACTGTAAA	Repbase
piggyBac-8_PI	3744	*Phytophthora infestans*	TTAA	CCCCTCTAGCCCCGC	Repbase
PiggyBac-2_HM	2703	*Hydra magnipapillata*	TTAA	CCCTCTAGTtCCCaAA CCCTCTAGTgCCCcAA	Repbase
BmoPBLE22	4019	*Bombyx mori*	TTAA	CACtTTCGCTGACAGT CACgTTCGCTGACAGT	Xu et al. 2006

PxuPBLE2	1999	*Papilio xuthus*	TTAA	CACGTTAACGACGGCG	In this study
piggyBac-1_SSa	2925	*Salmo salar*	TTAA	CCCTTGTGTGGTGTTC	Repbase
piggyBac-2_BF	4195	*Branchiostoma floridae*	TTAA	CCCTTGTCCTGCgGGC CCCTTGTCCTGCtGGC	Repbase
piggyBac-3_NV	2564	*Nematostella vectensis*	TTAA	CCCTTTCcggaCCAGG CCCTTTCatacCCAGG	Repbase
piggyBac-1_LMi	4725	*Locusta migratoria*	TTAA	CCCTTTGAGcGCTGCA CCCTTTGAGtGCTGCA	Repbase
ApiPBLE10	2421	*Acyrthosiphon pisum*	TTAA	CCTTCCAGcGGGCGCGC CCTTCCAGtGGGCGCGC	In this study
piggyBac-4_PI	3441	*Phytophthora infestans*	TTAA	CACCTTGGTTCCGGACG	RepBase
piggyBac-12_SMp	2445	*Schmidtea mediterranea*	TTAA	CCCTTAACCTCActCTAC CCGTTAACCTCAtaCTAC	Repbase
piggyBac-2_NV	3322	*Nematostella vectensis*	TTAA	CCCTTTGGTGCCgGCCCT CCCTTTGGTGCCtGCCCT	Repbase
MroPBLE3	2633	*Megachile rotundata*	TTAA	CCCTTTCGGTACGaGCGC CCCTTTCGGTACGgGCGC	In this study
MroPBLE2	2124	*Megachile rotundata*	TTAA	CACGTTGAGCGCGGCGCC	In this study
ApiPBLE8	2538	*Acyrthosiphon pisum*	TTAA	CCCGGCGTTAGTCGCACTA	In this study
PiggyBac-1_ParBra	2369	*Paracoccidioides brasiliensis*	TTAA	CCCCTCGTCCGTAAGTTCTTGCTCAATTAGCGTGATTTCAATTAGTGTAT	RepBase

Group 2:*32 PBLE sequences including Terminal Inverted Repeats (TIR) and Sub-terminal Inverted Repeats (STIR).*

Consensus TIR:



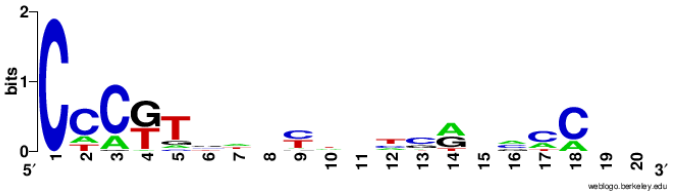


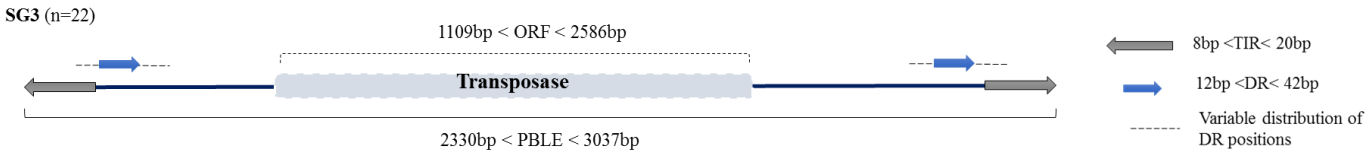
PBLE (bp)		Species	TSD+TIR		STIR	References
PiggyBac-5_HM	2449	*Hydra magnipapillata*	TTAA	CCCTTAA	TGTATCAaATtTGATACAACGGGT TGTATCAgATaTGATACAACGGGT	RepBase
piggyBac-2_SSa	2983	*Salmo salar*	TTAA	CCCTTGTGT	TTGTCtTAAGGGTcAAAAATGACCCGCC TTGTCcTAAGGGTaAAAAATGACCCGCC	Repbase
PiggyBac-6_HM	2384	*Hydra magnipapillata*	TTAA	CACGTTGACCGC	TCGTGTATaCATATGTATACAC TCGTGTATcCATATGTATACAC	Repbase
piggyBac-1_Hmel	2418	*Heliconius melpomene*	TTAA	CCCTTAATCAAA	TGGGAAATTATTTCCCA TGGGAAATTATTTCCCA	RepBase
piggyBac-1_SM	2434	*Schmidtea mediterranea*	TTAA	CCCTTTCT*AAGG*	*AAGG*TGGTGGCAATAT AAGG-GGTGGCAATAT	RepBase
AtrPBLE3	2424	*Amyelois transitella*	TTAA	CACTCTCCCT*GCC*	*GCC*aTGGGATCCGCC GCCgTGGGATCCGCC	In this study
piggyBac-10_PI	3477	*Phytophthora infestans*	TTAA	CCCCcTGGTTCCC CCCCtTGGTTCCC	GTCATAACTTACGcCaTACG GTCATAACTTACGaCgTACG	RepBase
DciPBLE	3036	*Diaphorina citri*	TTAA	CTCTTTgAgGGGC CTC*TTTcAaGGGC*	TTTCAAGGGCGGTGGTT *TTTCAAGGGC*GGTGGTT	In this study
PiggyBac-1_HM	3081	*Hydra magnipapillata*	TTAA	CCCTATATTGCAT	TTTGTTGCTTTAAaGCAACACcATGCA TTTGTTGCTTTAAgGCAACAaTATGCA	RepBase
piggyBac-3_SM	2457	*Schmidtea mediterranea*	TTAA	CACGTTGACTGCCA	CCGGCGGTACA	RepBase
piggyBac-11_SM	1984	*Schmidtea mediterranea*	TTAA	CCcTTTGAGTCCCG CCtTTTGAGTCCCG	GTTTAGAAAAAtTTTG GTTTAGAAAAAgTTTG	RepBase
piggyBac1_AG	3338	*Anopheles gambiae*	TTAA	CCGTCaTGTGTACGA CCGTCtTGTGTACGA	GTACAACCGCCTACCcGGGTAGGCCATACATTTTGTaTGGAaGgATGATGTTTTTCtTGGT TAtAAGgGTATATaTTTTGAGTTAcTTGTaAgATcaGAAaGAAAACcGTCTCATAGGTTcA TAATAATGTTTACT-AACATAtCGnAGtAAAAtTAGAATTTTAAAAATCCTnTnAAtAttT TTTTCAATCAAAAATATGTTGTAACTCCAAcACCCcAACCGGCCTTGCCNCAAGTTTnGAG TaGATGcTT GTACAACCGCCTACCtGGGTAGGCCATACATTTTGTgTGGAgGaATGATGTTTTTCgTGGT TAaAAGcGTATATcTTTTGAGTTAtTTGTtAaATttGAAtGAAAACtGTCTCATAGGTTtA TAATAATGTTTACTaAACATAnCGtAGnAAAAnTAGAATTTTAAAAATCCTaTcAAnAnaT TTTTCAATCAAAAATATGTTGTAACTCCAAaACCCnAACCGGCCTTGCCNCAAGTTT-GAG TgGATGtTT	RepBase
PIGGYB2_SM	2513	*Schmidtea mediterranea*	TTAA	CCTTC*AGTCTGTGCA*	*AGTCTGTGCA*AtTGGGGTCAAatgGACCCCA AGTCTGTGCA-TGGGGTCAAcccGACCCCA	RepBase
AgePBLE-IDT	2531	*Anticarsia gemmatalis*	TTAA	CCCTTTATAAGGCAGA	GgGaaAAA--TATTCCCACTGCCC GtGggAAAaaTATTCCCACTGCCC	Carpes et al.2009
ApiPBLE2	2500	*Acyrthosiphon pisum*	TTAA	CACGTTCACTGCGGA*T*	*T*GACGCCAATTGGCGTCAT TGACGCCAATTGGCGTCAT	In this study
ApiPBLE9	2483	*Acyrthosiphon pisum*	TTAA	CCCTGCGTAG*GTCGCG*	*GTCGCGCG*GAcTACcaTTGTAATCCAT GTCGCGCGGAtTACagTTGTAATCCAT	In this study

TcaPBLE	2270	*Tribolium castaneum*	TTAA	CCcTTTcaCTGCTGcA CCtTTTggCTGCTGaA	CACGTACATATACGT CACGTACATATACGT	Wang et al. 2008
ApiPBLE3	2476	*Acyrthosiphon pisum*	TTAA	CACGTTGACTGCC*ATG*	*ATG*cGGCaGCCGGCGgtCGCA ATGtGGCcGCCGGCGtcCGCA	In this study
AgoPBLE	2634	*Aphis gossypii*	TTAA	CCTTCCAGCGGG*CGCG*	*CGCG*CtGTTGTAAATaaTACAACAGT CGCGC-GTTGTAAATttTACAACAGT	Luo et al. 2011
CagPBLE	2488	*Ctenoplusia agnata*	TTAA	CCCTAG*AAGCCCAATC*	*AAGCCCAATC*-TACG AAGCCCAATCaTACG	Wu et al. 2011
PgoPBLE	2464	*Pectinophora gossypiella*	TTAA	CCCTAG*AtaACTAAAC* CCCTAGActACTAAAC	*ATAACTAAAC*ATTCGTC ATAACTAAACATTCGTC	Wang et al. 2010
Yabusame1	2472	*Bombyx mori*	**TTAT**	CCCGGCGAGC*ATGAGG*	*ATGaGGC*AGGGTATctCATACCC ATGtGGCAGGGTATgaCATACCC	Daimon et al. 2010
YabusameW	3415	*Bombyx mori*	TTAA	CCCGGCGAGC*ATGAGG*	*ATGaGGC*AGGGTATctCATACC ATGtGGCAGGGTATgaCATACC	Daimon et al. 2010
BmoPBLE07	2567	*Bombyx mori*	TTAA	CCCGGCGAGC*ATGAGG*	*ATGaGGC*AGGGTATctCATACCC ATGtGGCAGGGTATgaCATACCC	Xu et al. 2006
HviPBLE1	2319	*Heliothis virescens*	TTAA	CCCTTAATTA*CTCGCG*	*CTCGCG*TGGGGTATATtTAACCCCA CTCGCGTGGGGTATATaTAACCCCA	Wang et al. 2006
ApiPBLE1	2912	*Acyrthosiphon pisum*	TTAA	CAGGTTGACTGCCACGA CAGGTTGACTGCCA*CGA*	CGtGACCCTTATGTGtGTCA *CGa*GACCCTTATGTGgGTCA	In this study
piggyBac-10_CCri	2813	*Chondrus crispus*	TTAA	CTAGTGTCTATTTGACA	CCTCATaTGAGGAT--CAAATCCTCAcATGAGG CCTCATtTGAGGATcgCAAATCCTCAaATGAGG	RepBase
SmiPBLE8	2163	*Stegodyphus mimosarum*	TTAA	CCCGGgtTTgat*CGCGC* CCCGGcaTTacaCGCGC	*CGCGC*TGGGTGTcTCAcACCCAG CGCGCTGGGTGTgTCAtACCCAG	In this study
piggyBac-6_SM	2372	*Schmidtea mediterranea*	TTAA	CCCTATTAacACTAGGTT CCCTATTAttACTAGGTT	TGTGACTcTAAGACACATATTAA TGTGACTtTAAGACACATATTAA	RepBase
NviPBLE	3328	*Nasonia vitripennis*	TTAA	CCCTTTCGTGCACATAAG	GTTACAGCTATGGTGCGGGGAAACTGTTTATATTTCATAGGGTCCCGAGCCGCGCTGAATC CGAATcCGAGaTCAGATTTTGAAAATTTAAGATGGCGGATCCAATATGGCGGACGGGAATG TCAAAAAATCGAGATTTTGAAAAACAAAAAATTTTTTCGTATTCTACGTGAAAAtACAAAT AGAAACTGTTCTTGGTGTTTTTTCGATATCTCTCTTAGTTCTTGAGTAAAAAATTCATAAA CTTTGAAAATAACGCAGACTTATTTACCGTTCGTACACATAAGCGTTGCAGCCACATGAGT AGGAAACTGTTTATATTTCATAGGGTCCCGAGCCGCGCTGAATCCGAATCCGAGGTCAGGT TTTGAAAATTCAAGATGGCGGgTCCAATATGGCG GTTACAGCTATGGTGCGGGGAAACTGTTTATATTTCATAGGGTCCCGAGCCGCGCTGAATC CGAATtCGAGgTCAGATTTTGAAAATTTAAGATGGCGGATCCAATATGGCGGACGGGAATG TCAAAAAATCGAGATTTTGAAAAACAAAAAATTTTTTCGTATTCTACGTGAAAAcACAAAT AGAAACTGTTCTTGGTGTTTTTTCGATATCTCTCTTAGTTCTTGAGTAAAAAATTCATAAA CTTTGAAAATAACGCAGACTTATTTACCGTTCGTACACATAAGCGTTGCAGCCACATGAGT AGGAAACTGTTTATATTTCATAGGGTCCCGAGCCGCGCTGAATCCGAATCCGAGGTCAGGT TTTGAAAATTCAAGATGGCGGaTCCAATATGGCG	In this study
piggyBac-5_LSal	2213	*Lepeophtheirus salmonis*	TTAA	CCCTTTCGGCACaATAgGG CCCTTTCGGGACcATAaGG	GTGTCCcTATTGTCCTGAAgGGTAAAA GTGTCC-TATTGTCCTAAAtGGTAAAA	RepBase
piggyBac-12_LMi	2704	*Locusta migratoria*	TTAA	CACGTTCGCTGCCGTTCACGCTCTGGCGCG	AGGTGCCgC-TCACGCTGTGGCGCGTG AGGTGCCaCtTCACGCTCTGGCGCGTG	RepBase

Group 3:*22 PBLE sequences including Terminal Inverted Repeats (TIR) and Direct Repeats (DR).*

Consensus TIR:

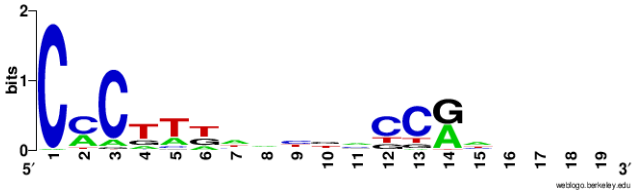


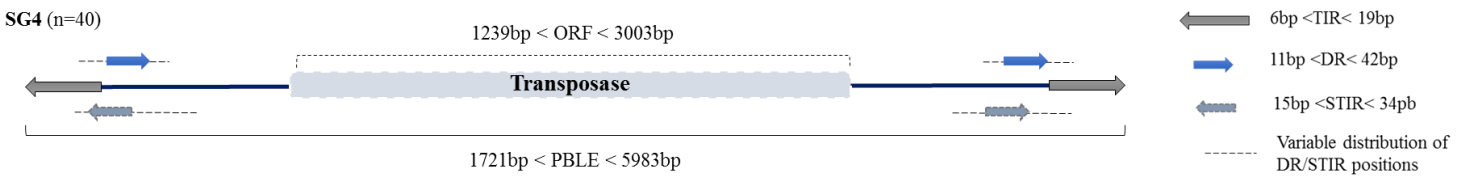


PBLE (bp)		Species	TSD+TIR		DR	References
piggyBac-2_Deu	2796	*Drosophila eugracilis*	TTAA	CCCTCTAA	CCcAAGaGTGCCTTGAGGCAC CCtAAGgGTGCCTTGAGGCAC	Repbase
piggyBac-10_LMilp	2403	*Locusta migratoria*	TTAA	CCCTGCTGT	CGGGTCATTATGACCCGAAA	Repbase
AtrPBLE2	2583	*Amyelois transitella*	**ATAT** TTAA	CACGTTAGaT CACGTTAGcT	AAGTGCCGGTCATTaTGACCC AAGTGCCGGTCATTtTGACCC	In this study
BmoPBLE56	2899	*Bombyx mori*	TTAA	CCCTTttCTTG CAAGctAAGGG	TGTATCTGTAGAGATACA	Xu et al. 2006
piggyBac-14_CCri	2691	*Chondrus crispus*	TTAA	CTAGTGTCTATT	CCTCAttTGAGG---ATTTCGCCTCAtATGAG CCTCAgaTGAGGctcATTTCGCCTCAaATGAG	Repbase
piggyBac-9_CCri	2531	*Chondrus crispus*	TTAA	CTAGTGACCTAA	CACCTCAATTGAGGA---CTAtaTCCTCAATTGAGgGGcgTTTT CACCTCAATTGAGGAaattTAccACCTCAATTGAGtGGAATTTT	Repbase
AcePBLE	2823	*Ancylostoma ceylanicum*	**ATAA**	CACGTTCGCGAC	GTGCTCGACCGGTCGTGCACTTGGTC	In this study
OabPBLE1	2772	*Orussus abietinus*	TTAA	CCAGTTAACC*GTG*	*GTG*TTTGACGACTATAtCCGTCAT GTGTTTGACGACTATAcCCGTCAT	In this study
piggyBac-5_SM	2423	*Schmidtea mediterranea*	TTAA	CCCTTAAATGCAT	TAAATGCATcGTGTtGCCATTTGGCAACAtACC TAAATGCATaTTGTaGCCATTTGGCAACAaACC	Repbase
piggyBac-2_CCri	2499	*Chondrus crispus*	TTAA	CCCTTCCTCtGTA CCCTTCCTCcGTA	TTGTGCCCAtATATGGGCAgAA TTGTGCCCAcATATGGGCAcAA	Repbase
piggyBac-6_LMi	2345	*Locusta migratoria*	TTAA	CACGTTGACTGCCA	CGAGtTAtCTCGTAGTTGGC CGAGaTAaCTCGTAGTTGGC	Repbase
SmiPBLE7	3037	*Stegodyphus mimosarum*	TTAA	CCCATTT*TTAGCCA*	*TTTAGCCA*AtTGTCCTGTACACAGGACAC TTTAGCCAaTGTCCTGTACACAGGACAC	In this study
piggyBac-5_PI	3519	*Phytophthora infestans*	TTAA	CCCTTCTAATACGA	ACGGATAATCCGGAA	Repbase
piggyBac-5_Lmi	2503	*Locusta migratoria*	TTAA	CCCTTGTGTCCA*CA*	AGGTTACCtGGGTAACCCAATG AGGTTACCcGGGTAACCCAATG	Repbase
VemPBLE	2330	*Vollenhovia emeryi*	TTAA	CCCTAGAACgCACA CCCTAGAACtCACA	GTGGGGTCGATATTGACCCCAGTTgTGAGTT GTGGGGTCGATATTGACCCCAGTTaTGAGTT	In this study
BmoPBLE75	2469	*Bombyx mori*	TTAA	CACTAGATTTAC*CAG*	*CAG*TCATtTTGACTG CAGTCATaTTGACTG	Xu et al. 2006
piggyBac-8_CCri	2598	*Chondrus crispus*	TTAA	CTAGTGTCTATTCGAC	CCTCATATGAGG--ActGgtGCCTCATATGAG CCTCATATGAGGaaAacGagGCCTCATATGAG	Repbase
SinPBLE	2398	*Solenopsis invicta*	TTAA	CCCGGCGTCATTGTgGC CCCGGCGTCATTGTtGC	CCCGGCCACAAtGGCTAcGT CCCGGCCACAAcGGCTAaGT	In this study
piggyBac-4_Ccri	2759	*Chondrus crispus*	TTAA	CTAGTGTCTATTTGACA	ACCTCAAATGAGgCTATgTC ACCTCAAATGAGaCTATtTC	Repbase
piggyBac-12_CCri	2650	*Chondrus crispus*	TTAA	CTAGTGTCTAGCCGACAAT	CTCAATTGAGGA	Repbase
ApiPBLE7	2716	*Acyrthosiphon pisum*	TTAA	CCCGTCATGAGTCGCAtCA CCCGCCATGAGTCGCAcCA	GTCAAAAAGACCgTGTGCGAC GTCAAAAAGACCtTGTGCGAC	In this study
piggyBac-1_Sin	2525	*Solenopsis invicta*	TTAA	CCG*GGGTACCCGGGTACCC*A	*GGGTACCCGGGTACCC*	Repbase

Group 4: *40 PBLE sequences including Terminal Inverted Repeats (TIR), Direct Repeats (DR) and Sub-Terminal Inverted Repeats (STIR).*

Consensus TIR:

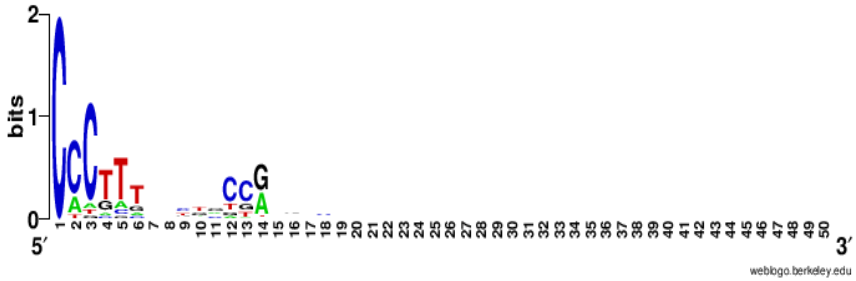


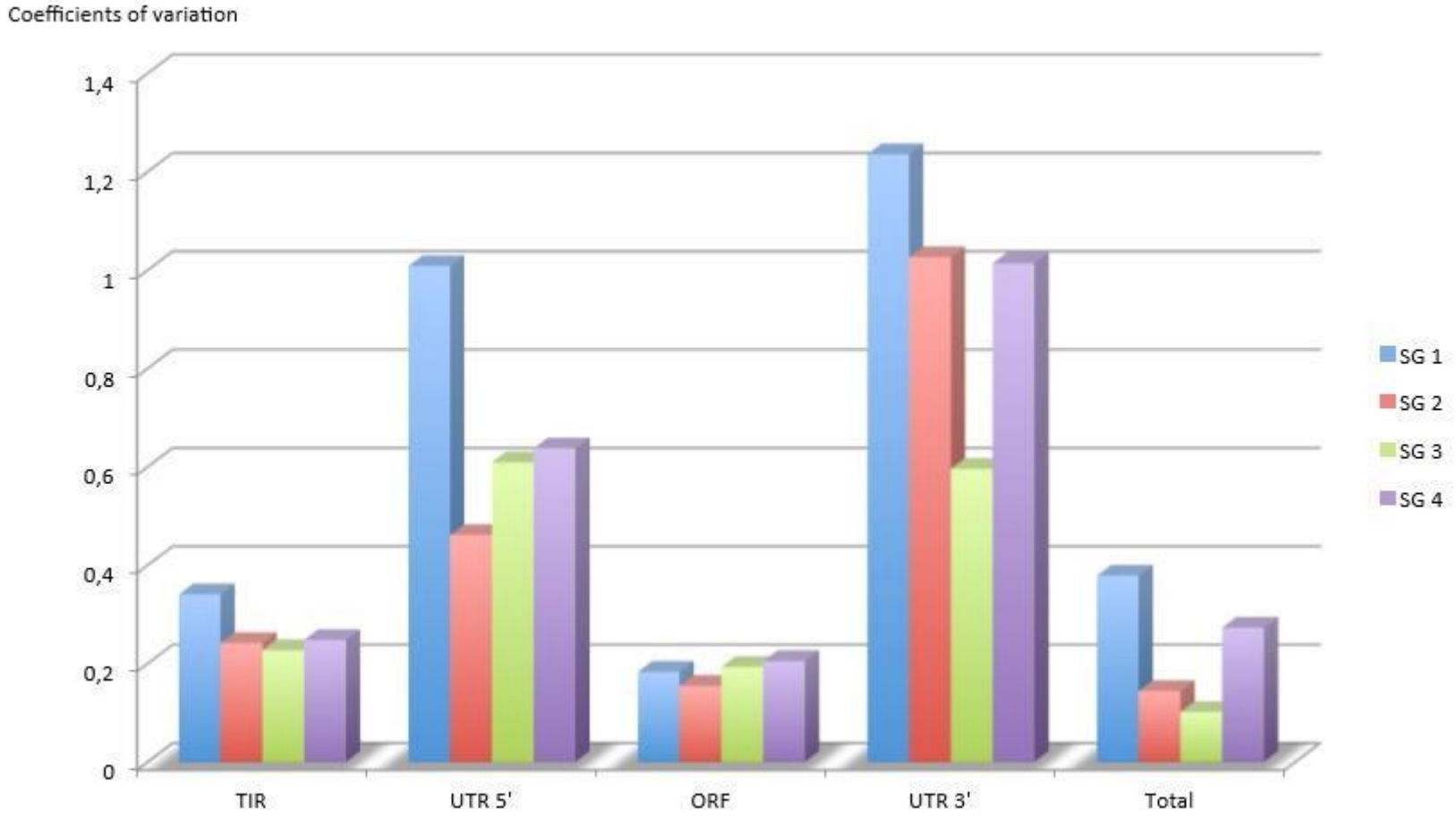


PBLE (bp)		Species	TSD+TIR		DR	STIR	References
SfrPBLE-Wu	2939	*Spodoptera frugiperda*	TTAA	CCCTTT	CAGtTAGAGGGAATATTtTTcCCACCAATaAAAA CAGgTAGAGGGAATATTcTTaCCACCAATtAAAA	TTTTTT--TTGGTGGgAAGAATATTCCCTCTA TTTTTTaaTTGGTGGtAAGAATATTCCCTCTA	Wu et al. 2014
MdaPBLE	2116	*Myotis davidii*	TTAA	CACTAG	CAAgGaAGTCATTTTGACTGCTTTT CAAtGcAGTCATTTTGACTGCTTTT	TATACCTATATTGCCCAAAAGCAGTCA TATACCTATATTGCCCAAAAGCAGTCA	Tang et al. 2015
MmoPBLE	2252	*Mengenilla moldrzyki*	TTAA	CgCTAGA TCTAGtG	AGTCATTTTGACTGCTTT	TATACCTgtATTGCCgaAAAGCAGTCA TATACCTacATTGCCccAAAGCAGTCA	Tang et al. 2015
MroPBLE	2528	*Megachile rotundata*	TTAA	CACTAGA	CAGTCATTTTGACTGCTTT	TATACCTgTATTG-CCtAAAGCAGTCA TATACCTaTATTGcCCcAAAGCAGTCA	Tang et al. 2015
ApiPBLE4	3947	*Acyrthosiphon pisum*	TTAA	CCCTTTA	AATATTTTTCCCA	TTCCCACCTAAGAAAA	In this study
piggyBac-7_PR	5356	*Phytophthora ramorum*	TTAA	CCCTTTCGTG	CTGTgAAAGGGGGCATTTTGCCCCCTTTC CTGTcAAAGGGGGCATTTTGCCCCCTTTC	GGGGCATTTTGcCCCCTTTCAT GGGGCATTTTGaCCCCTTTCAT	RepBase
piggyBac-4_PR	5983	*Phytophthora ramorum*	TTAA	CCCTTTCGCG	CTAAAAGGGCACAAAATTCGCCC	CTAAAAGGgcACaaaaTTCGCCCTTTAAGGtTTT CTAAAAGGatACttttTTCGCCCTTTAAGGgTTT	RepBase
piggyBac-1_CCri	2863	*Chondrus crispus*	TTAA	CTAGTGTCTAG	TGgCCTCATATGAGGcCAA--tTGgCCTCATATGAGGC TGaCCTCATATGAGGtCAAgggTGaCCTCATATGAGGC	TGgCCTCATATGAGGcCAAtTG TGaCCTCATATGAGGtCAAaTG	RepBase
piggyBac-9_SM	2774	*Schmidtea mediterranea*	TTAA	CCCGCAAACGAT	TGGACATTTTTGTCCAAAAT TGGACATTTTTGTCCAAAAT	TTTTGGACAAAAATGTCCA TTTTGGACAAAAATGTCCA	RepBase
SmiPBLE6	2439	*Stegodyphus mimosarum*	TTAA	CCCAT*TTATGCC*	*TTaTGCC*TAcTGTTCCATTTTTGGAACACCA TTtTGCCTAgTGTTCCATTTTTGGAACACCA	GTGTTCCATTTTTGGAACACTAGGCAAAAC GTGTTCCAAAAATGGAACACTAGGCAAAAC	In this study
piggyBac-1_Ami	3224	*Alligator mississippiensis*	TTAA	CCtTCATaCGTTC CCcTCATcCGTTC	TGGGGTAACaacTTACCCCA TGGGGTAACttgTTACCCCA	TGGGGTAACttGTTACCCCA TGGGGTAACaaGTTACCCCA	RepBase
TniPBLE	2475	*Trichoplusia ni*	TTAA	CCCTAGAAAGATA	TGaGTCAAaaTgACGCATGATTATCTTTtACGT TGcGTCAAttTtACGCATGATTATCTTTaACGT	TGCGTAAAATTGACGCATG TGCGTAAAATTGACGCATG	Fraser et al. 1983
McrPBLE	2472	*Macdunnoughia crassisigna*	TTAA	CCCTAGAAAGATA	TGaGTCAAaaTgACGCATGATTATCTTTtACGT TGcGTCAAttTtACGCATGATTATCTTTaACGT	TGCGTAAAATTGACGCATG TGCGTAAAATTGACGCATG	Wu et al. 2008
piggyBac-2_Hmel	2418	*Heliconius melpomene*	TTAA	CCCAGATAAGCCT	CCTACTGTCCTATATTTAGGAC CCTACTGTCCTACATATAGGAC	GTCCTATATTTAGGACAGTAG GTCCTATATGTAGGACAGTAG	RepBase
CsuPBLE	2406	*Chilo suppressalis*	TTAA	CCCAG*ATTAGCCT*	*ATTAgCCT*ACTGTCCTAtATtTAGGAC ATTAtCCTACTGTCCTAcATaTAGGAC	CTATATaTAGGACAGTAGGATAATA CTATATgTAGGACAGTAGGATAATA	Luo et al. 2014
piggyBac-3_PR	3484	*Phytophthora ramorum*	TTAA	CACCTTGACTCCG	TGTCGTaAGTTACGACATT TGTCGTcAGTTACGACATT	ATGTCGTAAGtTACGACAT ATGTCGTAAGcTACGACAT	RepBase
piggyBac-6_CCri	2841	*Chondrus crispus*	TTAA	CGCTAGTGTCTAT	TTGTCCTCATATGAGGACAA	GACCG----TTGTCCTCATATGAGGACAA GACCGgcccTTGTCCTCATATGAGGACAA	RepBase
piggyBac-5_CCri	2720	*Chondrus crispus*	TTAA	CCCCTGCGCTGTC	CTTaGGCCCATATATGGGCCGAA CTTcGGCCCATATATGGGCCGAA	GGCCCATATATGGGCCGAA GGCCCATATATGGGCCGAA	RepBase
piggyBac-2_DBp	2744	*Drosophila bipectinata*	TTAA	CCCTTTATAcGGC CCCTTTATAtGGC	TGtTGcATaTcTGCAACATACCGTTT TGaTGtATtTaTGCAACATACCGTTT	TATGTTGCAgAtATACATCAAA TATGTTGCAtAaATACATCAAA	RepBase
piggyBac-1_Aae	3322	*Aedes aegypti*	TTAA	CCCTTTCcTTCCC CCCTTTCtTTCCC	CTTCCCAtgGtAgCACAGGTGATCCAcCA CTTCCCAatGgAtCACAGGTGATCCAtCA	TGGATCACtgaTGATCCATTGGGAAcAaAAC TGGATCACctgTGATCCATTGGGAAgAtAAC	RepBase
piggyBac-1_LCh	2691	*Latimeria chalumnae*	TTAA	CCCGTTAcTTACC CCCGTTAaTTACC	ATGTTCCAtTTTTGGAACATT ATGTTCCAaTTTTGGAACATT	ATGTTCCAtttTTGGAACATTGGGAA ATGTTCCAaaaTTGGAACATTGGGAA	RepBase
HcoPBLE	2642	*Haemonchus contortus*	TTAA	CACgTTGCGTACCG CACaTTGCGTACC*G*	GCGTGCATCACCATGTGATGCAC GCGTGCATCACCATGTGATGCAC	GCGTGCATCACcatGTGATGCAC *GC*GTGCATCACatgGTGATGCAC	In this study

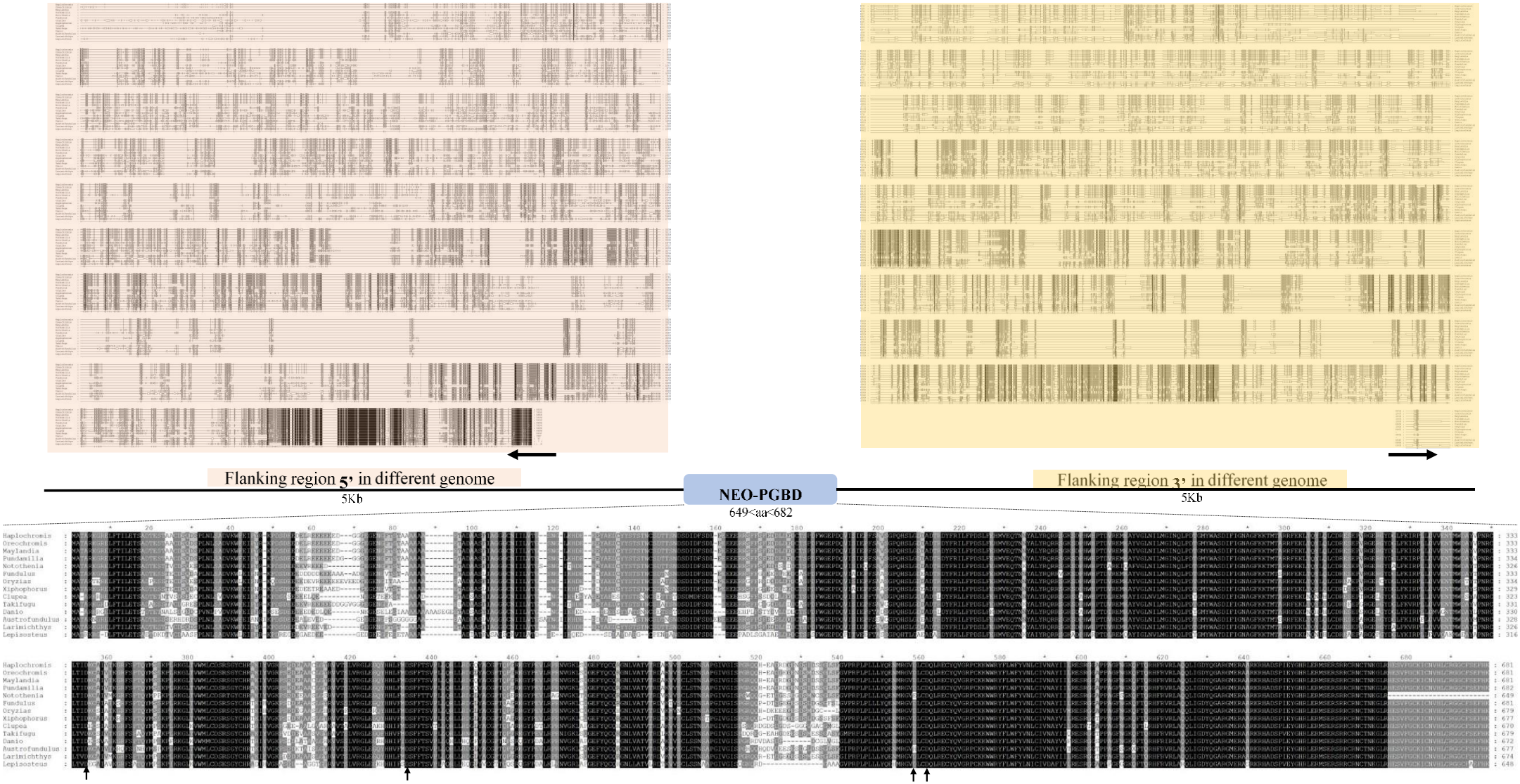
piggyBac-10_SM	2648	*Schmidtea mediterranea*	TTAA	CA*CTATTCCTACCG* CA*CTATTCCTACCG*	GACCGGTCATTTGACCGG	*CTATTCCTACCG*gGaCCGGTCAAATGACCG *CTATTCCTACCG*cGgCCGGTCAAATGACCG	RepBase
piggyBac-2_LMi	2760	*Locusta migratoria*	TTAA	CA*CTAGAAGGACCG*	GGGCCAATTTGGCCCC	*CTAgAAGGACCG*a-GGGGCCAAATTGGCCCC CTAcAAGGACCGgtGGGGCCAAATTGGCCCC	RepBase
piggyBac-8_SM	2409	*Schmidtea mediterranea*	TTAA	CACTAG*ATtTACCA* CACTAGATaTACCA	ACCAGTCATTTaGACTGGT ACCAGTCATTTtGACTGGT	*ATTTaTACCA*gtACCAGTCAAAATGACaGGT ATTTcTACCAa-ACCAGTCAAAATGACtGGT	RepBase
ApiPBLE5	2801	*Acyrthosiphon pisum*	TTAA	CACGTTCACGgCCA CACGTTCACGaCCA	TGTCACATATATGTGACATGT TGTCACACATATGTGACATGT	TGTCACATATGTGTGACA TGTCACATATGTGTGACA	In this study
SmiPBLE5	2550	*Stegodyphus mimosarum*	TTAA	CCCATTAgCGCCCA CCCATTAaCGCCCA	CGCCCAGTGTCCTCAttTGAGGACGctgTAGAAAATtTtGTT CGCCCAGTGTCCTCAaaTGAGGACGgctTAAAAAATaTgGTT	GTCCTCATTTGAGGACgCTG GTCCTCATTTGAGGACaCTG	In this study
FhePBLE	2375	*Fundulus heteroclitus*	TTAA	CCCTTATATTATGTT	AACcTGTGTGTAAAA AACtTGTGTGTAAAA	GATgATgTTGCGGGTCATTTTGtCCCGgAC GATtATcTTGCGGGTCATTTTGaCCCGcAC	In this study
piggyBac-1_BF	3758	*Branchiostoma floridae*	TTAA	CCCTCAAaCA*CCCGA* CCCTCAAgCACCCGA	TGGGGTCCGTTTGGACCCCA TGGGGTCCATTTGGACCCCA	*CCGA*cTGGGGTCCAAAcGGACCCCA CCGAgTGGGGTCCAAAtGGACCCCA	RepBase
HarPBLE	2508	*Helicoverpa armigera*	TTAA	CCCTAG*AAGCCCAATC*	TTgCACGTCATTTTgACGTATaATTGGGCTTT TTaCACGTCATTTTtACGTATgATTGGGCTTT	*AAGCCCAA-TC*TACGTAAATTTGACGT ACGTCATTTTTACGTATGaTTGGGCTT	Sun et al. 2008
AipPBLE	2476	*Agrotis ipsilon*	TTAA	CCCTAG*AAGCCCAATC*	TTgCACGTCATTTTgACGTATaATTGGGCTTT TTaCACGTCATTTTtACGTATgATTGGGCTTT	*AAGCCCAA-TC*TACGTAAATTTGACGT AAGCCCAAtCATACGTAAAAATGACGT	Wu et al. 2011
piggyBac-9_LMi	2489	*Locusta migratoria*	TTAA	CCCTGCG*GCGGGCGCG*	AGaCGGCACTGcGTaCCGCACGCGCCCGgTaAtGTAA AGgCGGCACTGaGTgCCGCACGCGCCCGcTgAcGTAA	*GCGGGCGCGT*gCGGCATACAGTGCCGCC GCGGGCGCGT-GCGGCACTCAGTGCCGCC	RepBase
MroPBLE4	2382	*Megachile rotundata*	TTAA	CCCTTAA*tTAGGCGCAT* CCCTTAAcTAGGCGCAT	GACCCCATGCG	*AATTAGGCGCAT*GGGGTC AATTAGGCGCATGGGGTC	In this study
piggyBac-8_LMi	2477	*Locusta migratoria*	TTAA	CCCT*AGAACGGTCGGGC*	GGGTCTGTCAGACCCAGCCCGACCGTCTACGTAACTAATTTC GGGTCTCTCAGACCCAGCCCGACCGTTCATGAAACTTTTTTC	*GAACGGTCGGGC*TGGGTCTGtGAGACCCA GAACGGTCGGGCTGGGTCTGaGAGACCCA	RepBase
piggyBac-11_PI	3435	*Phytophthora infestans*	TTAA	CACcTTGGtTCCGAACG CACaTTGGcTCCGAACG	ATGtCGTAagTTACGACATACtGCTT ATGgCGTAtcTTACGACATAC-GCTT	TATGTCGTAAGtTACGaCAT TATGTCGTAAGaTACGcCAT	RepBase
piggyBac-3_CCri	2608	*Chondrus crispus*	TTAA	CTAGTGTCTATTCGACA	CCTCATATGAGGA	TgTCCTCATATGAGG TtTCCTCATATGAGG	RepBase
ApiPBLE6	2719	*Acyrthosiphon pisum*	TTAA	CAGGTT*AACcGCTAtGT* CAGGTTAACtGCTAcGT	GTGTACCAAAtATGGTACACGT GTGTACCAAAaATGGTACACGT	*AAcCGCtAtGT*GTACCATTTTTGGTACA AAgCGCcAcGTGTACCATTTTTGGTACA	In this study
piggyBac-1_DBi	1721	*Drosophila biarmipes*	TTAA	CCCTTATG*TTAGTAACC*	GGGTACCCGGGTACCC	*TTAGcAACC*GGGTACCCGGGTACCCA TTAGtAACCGGGTACCCGGGTACCCA	Repbase
piggyBac-11_CCri	2829	*Chondrus crispus*	TTAA	CTAGTGTCtAAGCGTACA CTAGTGTCcAAGCGTACA	GGGcCTCATATGAGGcCA GGGtCTCATATGAGGaCA	GgCCTCATATGAGgCC GtCCTCATATGAGaCC	RepBase
TspPBLE	3006	*Trichinella spiralis*	TTAA	*CCCTTTCATGCCcAATgTT* CCCTTTCATGCCaAATtTT	*CCtTTcATGCCCAATGTT*GCTTTAAAGCA CCgTTaATGCCCAATGTTGCTTTTAAGCA	TGTTGCTTAAAAGCAACAT TGTTGCTTAAAAGCAACAT	In this study

Consensus of all TIR-PBLE

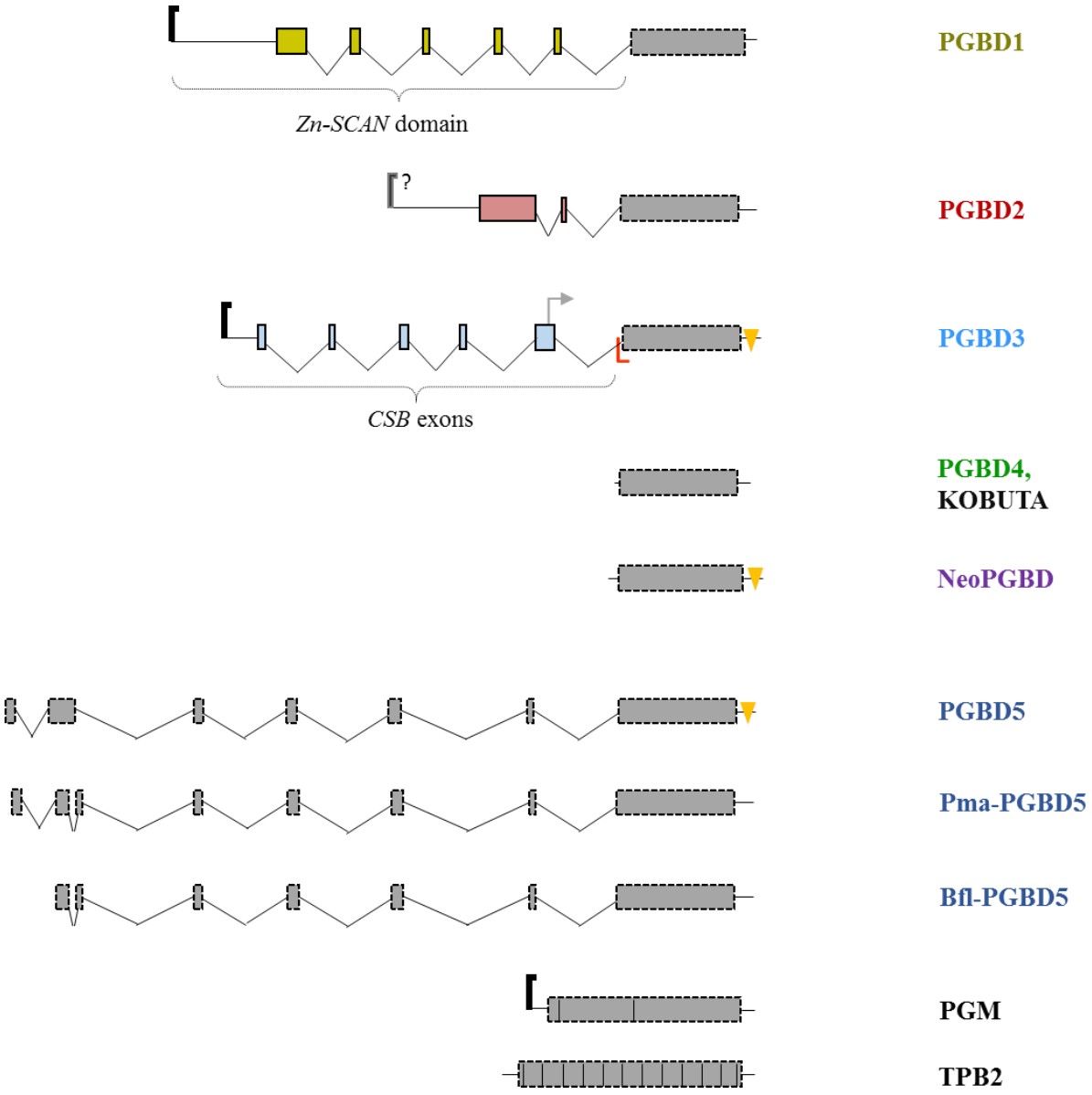




Supplemental data 6. Length variation of PBLE element, UTR and ORF. Each column corresponds to a structural group (SG); SG1: PBLE sequences with Terminal Inverted Repeats (TIR) only. SG2: PBLE sequences with TIR and Sub-Terminal Inverted Repeats (STIR). SG3: PBLE sequences with TIR and Direct Repeats (DR). SG4: PBLE sequences with TIR, DR and STIR. The vertical scale indicates the coefficients of variation from TIR, UTR (Untranslated Terminal region), ORF and total of PBLE element.



Supplemental data 8. Global view of NeoPGBD element. The amino acid transposase of NeoPGBD members (n=14) are aligned. The presumptive DDDD domain is indicated by arrows. The 5Kb alignment of adjacent regions on both sides of this element is represented showing the blocks level of conservation, the 5' region is boxed in pink, and the 3' region in yellow. For all alignments, black and grey boxes are for identical and similar amino acids (from 60%), respectively. Left and right arrows indicate the reading direction 5' and 3'.



Supplemental data 9. Structure of different domesticated elements. This figure gives the general structures of the nine *piggyBac* derived elements, including PGBD1 and PGBD2 found in mammals, PGBD3 and PGBD4 in primates, KOBUTA from *Xenopus sp*, NeoPGBD from Neopterygii, PGBD5 in gnasthostomates with two orthologous sequences, namely Pma from the agnathe *Petromyzon marinus* and Bfl from the cephalochordate *Branchiostoma floridae*, PGM from *Paramecium tetraurelia*, TPB2 from *Tetrahymena thermophila*. The exons derived from the *piggyBac* transposase are represented in grey. Several exons, located in the Nterminal regions of PGBD1, PGBD2 and PGBD3 are indicated in yellow (corresponding to *ZnSCAN* domain), red and blue boxes (corresponding to CSB exon 1 to 5), respectively. Introns are mentioned in thin black lines. PGM has two small introns with 20 and 34 nucleotides (Baudry et al. 2009). Similarly, TPB2 contains twelve introns from 47 to 153 nucleotides according to Cheng et al. (2010). Yellow triangles refer to the polyadenylation signal AATAAA. Only PGBD3 contains a splicing site indicated by a red "L". Thus, a CSB-PGBD3 fusion protein that joins the N-terminal domain of CSB (coding exons 2–5) to the intact PGBD3 transposase is generated. While PGBD1 and PGM contain a host promoter (black flags), PGBD2 includes a possible promoter site and PGBD3 transposase presents a cryptic promoter in CSB exon 5 (grey arrow).

Analyse globale et conclusions

Dans ce chapitre, nous avons entrepris l'identification et la caractérisation des éléments appartenant à la superfamille *piggyBac*, leurs distributions spécifique et leur évolution. Pour cela, une recherche par similarité à partir des 117 séquences protéiques déjà publiées a été effectuée. Après filtration des séquences afin de retenir les éléments les plus complets, nous avons pu récupérer 50 séquences PBLE, 295 séquences domestiquées PGBD et 101 séquences annotées PGBD-like ayant un statut inconnu correspondant soit à des éléments domestiqués ou à des PBLE morts.

Ainsi, durant cette recherche, les PBLE se réfèrent à des séquences (actives ou non) avec deux TIR, deux UTR et une transposase, alors que les PGBD correspondent à des séquences domestiquées, c'est-à-dire des séquences en copie unique au sein du génome, orthologues chez différentes espèces et ayant un rapport *Ka* (mutation non synonyme) /*Ks* (mutation synonyme) indiquant l'existence d'une sélection positive ou purifiante. Enfin, le terme PGBD-like a été limité aux séquences restantes, c'est-à-dire avec un ou sans TIR ou UTR et une transposase qui peut être partiellement tronquée.

Dans un premier temps, cette étude a permis la caractérisation détaillée de 157 éléments PBLE incluant 50 nouvelles séquences. La taille de ces éléments varie de 1721pb à 8451pb, avec une valeur moyenne de 2813 ± 84pb. Alors que, la majorité des PBLE présente un site cible de duplication (TSD) de type TTAA, cinq éléments en copie unique présentent un TSD différent. Ceci est probablement dû à une dégénérescence du site cible de la duplication, postérieure à l'insertion de l'élément. Par ailleurs, les TIR paraissent relativement divergents en longueur (une moyenne de 14pb±2pb, sans inclure les TIR de 50pb trouvés chez le champignon *Paracoccidioides brasiliensis*), et en séquence. Toutefois, il existe une forte conservation de la première cytosine et un C/A/T pour les deux nucléotides suivants. Des expériences de mutagénèse ont confirmé que le premier G en 3'terminal, joue un rôle dans la sélection du site d'excision et que sa substitution est suffisante pour empêcher l'excision de l'élément (Haniford et Kleckner 1994; Elick *et al.* 1997). En revanche, l'impact fonctionnel des deuxième et troisième nucléotides des TIR n'est pas si clair, puisqu'ils ne semblent pas impliqués dans l'abolition d'excision (Xu *et al.* 2006; Hikosaka *et al.* 2007; Luo *et al.* 2011; Mitra *et al.* 2013).

Sur l'ensemble des PBLE, quatre groupes structuraux ont été identifiés en fonction de la présence ou de l'absence de répétitions directes (DR) et/ou de répétions sub-terminales inversées

(STIR). Le groupe structural 1 (SG1) comprend 63 éléments (répartis chez 37 espèces) ayant seulement des répétitions terminales inversées (TIR). SG2 (n=32, chez 21 espèces) correspond à une structure des éléments avec TIR et STIR, SG3 (n=22, chez 14 espèces) avec TIR et DR et SG4 (n=40, chez 26 espèces) présente la structure la plus complexe avec TIR, DR et STIR. Ces observations suggèrent que ces éléments sont structurellement très flexibles. En outre, ces groupes structuraux sont représentés différemment au sein des espèces. SG1 (TIR seulement) est largement répandu des hétérokontes aux primates (37 espèces), alors que les distributions spécifiques de SG2 (TIR et STIR), SG3 (TIR et DR) et SG4 (TIR, STIR et DR) sont plus restreintes.

Deux hypothèses alternatives peuvent être proposées pour expliquer la distribution spécifique des différentes structures :

- SG1 est la structure ancestrale, tandis que les autres structures dérivent de cette séquence par acquisition indépendante de DR et STIR. A ce titre, la présence de structures identiques dans des espèces éloignées peut correspondre à des convergences évolutives, et/ou à des transferts horizontaux.
- les quatre groupes structuraux étaient déjà présents dans l'ancêtre commun de toutes les espèces dans lesquelles des copies de PBLE ont été trouvées, et leur répartition inégale est due à la perte indépendante dans plusieurs lignées.

Cependant, l'absence de congruence entre la phylogénie des éléments et la distribution de leur structures dans la phylogénie, et l'absence de séquences consensus entre les régions UTR, sont en faveur de l'hypothèse d'acquisitions indépendantes (convergences) et d'évolution rapide des séquences de DR et STIR. Les incohérences observées entre la phylogénie des transposons, la distribution des groupes structuraux et la phylogénie des espèces peut également suggérer que *piggyBac* peut être facilement transféré horizontalement. Cela peut être dû à une activité indépendante des facteurs de l'hôte et pourrait expliquer le succès cet élément est un vecteur puissant pour l'ingénierie du génome.

De plus, les répétitions internes (DR/STIR), parfois sous la forme de motifs palindromiques, seraient potentiellement impliquées dans l'activité des éléments *via* la liaison de la transposase et la stabilisation du complexe transposase-TIR/DR/STIR. Toutefois, une vérification expérimentale fonctionnelle doit être réalisée. Si tel est le cas, cela pourrait illustrer une coévolution rapide entre la transposase et les séquences terminales répétées.

Au niveau de la transposase, la région N-terminale est variable en séquence. Ceci dit, le domaine catalytique D^{268}, D^{346}, D^{447}, (D^{450}) est hautement conservé. De même, les cystéines du domaine CRD de la région C-terminale, sont présentes dans presque toutes les séquences et leur espacement est relativement bien conservé alors que la présence et la position de l'histidine est plus aléatoire.

Dans un deuxième temps, neuf groupes d'éléments domestiqués PGBD ont été détectés. Toutes les séquences domestiquées ont été trouvées en une seule copie et présentent une région codante dans laquelle toute ou une partie de la transposase de *piggyBac* a été identifiée. Au sein de chaque groupe, le niveau de similarité entre les séquences est supérieur à 85%. Par ailleurs l'analyse du rapport *Ka/Ks* est faible <1 suggérant l'existence d'une sélection purifiante ou stabilisante pour maintenir une fonction putative. Cependant, aucun alignement ne peut être effectué entre les groupes sur la base de la transposase. De plus, leur phylogénie est bien structurée avec des nœuds robustes, suggérant l'existence de plusieurs événements indépendants de domestication.

Le groupe PGBD1 est présent chez les mammifères et comprend 62 séquences. L'élément provient d'une fusion ancestrale N-terminale, entre cinq exons contenant le domaine LER/SCAN (régions riches en leucine) avec la transposase de *piggyBac* (Sarkar *et al.* 2003). La taille de ces séquences varie de 312 à 826 aa, suggérant l'existence d'un ou de plusieurs indels au cours de l'évolution.

Le groupe PGBD2 présente le même nombre de séquences ainsi que la même distribution que PGBD1. Deux exons sont présents en amont de la séquence transposase dont la taille varie de 586 à 759 aa.

La transposase de PGBD3 est insérée, pour les 15 séquences trouvées, dans le 5ème intron du gène du groupe B du syndrome de Cockayne. En effet, contrairement à d'autres éléments domestiqués, la transposase PGBD3 est flanquée par un site d'épissage 3' potentiel dans la région 5' et par un signal de polyadénylation dans la région 3'. L'épissage alternatif de cette région produit une protéine de fusion CSB-PGBD3 qui a été conservée depuis l'ancêtre commun des lignées humaines et des ouistitis, c'est-à-dire 43 Mya (Newman *et al.* 2008).

Les 16 séquences de PGBD4 contiennent une ORF unique codant pour une protéine d'environ 585 aa et ne sont présentes que chez les primates.

Le groupe PGBD5, le plus large (n=147) et le plus ancien, est présent à partir de l'ancêtre des myomérozoaires, il y a 525 Mya (Pavelitz *et al.* 2013). L'élément contient plusieurs introns (six pour les céphalocordés et les gnathostomes, sept pour les agnathostomes) et la longueur de la protéine codée varie de 343 à 732 aa. Cette variation de taille est probablement due à une faible conservation de la partie N-terminale de la protéine.

Par ailleurs, deux autres éléments domestiqués présentent des introns dans la séquence dérivée de la transposase de *piggyBac*, à savoir PGM qui contient deux introns avec une région codante de 1065 aa et TPB2 avec 12 introns qui présente une ORF codante de 1220 aa.

L'élément KOBUTA, spécifique de *Xenopus sp*, présente une ORF sans introns codant pour une protéine de 610aa.

Enfin un nouveau groupe désigné NeoPGBD, annoté initialement comme PGBD-like 4, a été identifié chez les néoptérygiens (n=14) suggérant une domestication ancienne d'au moins 250 Mya.

La taille de la transposase domestiquée varie de 649 à 682 aa. Par ailleurs, une séquence hautement conservée de 365pb (identité > 90%) de fonction inconnue est située immédiatement en amont de l'ORF et un signal potentiel de polyadénylation AATAAA semble être généralement conservé dans la région 3'. Enfin, aucun gène ou domaine associé n'a été détecté dans les régions flanquantes (5kb de chaque côté) de cette séquence domestiquée.

L'étude de ces éléments montre que la domestication n'est pas systématiquement accompagnée de modifications de l'activité de la transposase (Sarkar *et al.* 2003; Keith *et al.* 2008; Pavelitz *et al.* 2013). Dans ce contexte, *piggyMac*, TPB2 et KOBUTA ont conservé le domaine catalytique DDD et la région C-terminale des PBLE, qui seraient impliqués dans des mécanismes d'excision. Ce motif DDD est présent également chez PGBD4 mais la fonction de cette séquence n'est pas connue (Mitra *et al.* 2008). Pour PGBD3, alors que le motif DDD n'est pas strictement conservé, le domaine C-terminal de la protéine de fusion CSB-PGBD3 est capable de mobiliser les MITE MER85 (Gray *et al.* 2012). En outre, la protéine codée par PGBD5, qui est dépourvue de région riche en cystéines, semble être impliquée dans la transposition de l'ADN selon le mécanisme « couper-coller » dans des cellules humaines. D'après Henssen *et al.* (2015), l'intégration génomique nécessiterait des résidus d'acide aspartique distincts et des séquences d'ADN spécifiques (TIR comprises) comparées à celles de Uribo2 (PBLE chez la grenouille *Xenopus tropicalis*), TniPBLE (trouvé chez le lépidoptère *Trichoplusia ni*), *piggyMac* (chez la paramécie) et *piggyBat* (PBLE trouvé chez la chauve-souris *Myotis lucifugus* MluPBLE). Pour les PGBD1 et PGBD2, le motif DDD n'est pas conservé et aucune activité de type transposase n'est suspectée.

Concernant l'évolution des éléments appartenant à la superfamille des *piggyBac* et l'origine des éléments domestiqués, de quelques éléments peuvent être rappelés.

- plusieurs PBLE semblent être potentiellement actifs, ou l'étaient dans un passé récent.
- la phylogénie générale incluant les 157 PBLE, 305 PGDB et 101 PGDB-like, montre de nombreuses relations étroites entre PGBD et PBLE ou PGBD-like. En effet, les PBLE Uribo1, Uribo2 avec KOBUTA de *Xenopus sp* forment un clade robuste comme l'avait déjà mentionné Hikosaka *et al.* (2007). L'élément PvaPBLE identifié chez la chauve-souris *Pteropus vampyrus* présente un taux de similarité de 96% avec les membres de PGBD4, avec un rapport *Ka/Ks*=0.23±0.01 (PvaPBLE *vs* PGBD4).
- les membres de PGBD3, SmiPBLE7 et Smi-PGBD-like3 trouvés chez l'araignée *Stegodyphus mimosarum* et Api-PGBD-like3 du puceron *Acyrthosiphon pisum* sont regroupés ensemble. De plus, l'analyse des régions adjacentes de ces ORF révèle la présence en 5' d'un site d'épissage 3' et d'un site de polyadénylation AATAAA en 3'. En dehors du clade de PGBD3, ces deux motifs peuvent être présents mais avec une répartition inégale.

- pour PGBD5, quelques fragments de l'ORF ont été identifiés dans le génome de l'hémichordé *Saccoglossus kowalevskii* sans que les introns et/ou les séquences flanquantes aient été trouvés. Cependant, aucune homologie n'a été détectée chez les urochordés (*Ciona intestinalis, Oikopleura dioica*) et chez les échinodermes (*Strongylocentrotus purpuratus, Acanthaster planci*). En supposant que les assemblages génomiques de ces espèces soient corrects, ceci suggère que l'insertion à l'origine de PGBD5 s'est produite chez l'ancêtre des Myomerozoa. Celle-ci a rapidement acquis une nouvelle fonction et était probablement sous une pression sélective élevée conduisant à un balayage sélectif sur toute la région incluant les séquences flanquantes.

Discussion générale & Perspectives

Les éléments transposables (ETs) jouent un rôle fondamental dans la dynamique structurale et fonctionnelle des génomes, pouvant conduire à des changements importants de régulation de l'activité des gènes, notamment *via* des marques épigénétiques, ou encore à l'apparition de nouveaux gènes et de nouvelles fonctions (Britten 1996; Capy *et al.* 2000; Vieira *et al.* 2002; Böhne *et al.* 2008; Sinzelle *et al.* 2009). En raison de leur mobilité et de leur ubiquité, ces séquences peuvent également être utilisées en biotechnologie comme outils de mutagénèse et de transgénèse (Ryder et Russel 2003; Delaurière *et al.* 2009). Ainsi, les transposons des superfamilles *Tc1-mariner-IS630* et *piggyBac* sont fréquemment utilisés en thérapie génique ou dans les stratégies de lutte contre les ravageurs (Wang *et al.* 2000; Handler et al 2002; Jegot 2007; Yusa *et al.* 2015; Voigt *et al.* 2016). Toutefois, l'étude de leurs caractéristiques structurales et fonctionnelles ainsi que leur distribution et leur dynamique au sein des espèces ciblées doit d'abord être élucidée avant de les utiliser comme outils afin de mieux contrôler l'efficacité et le devenir des transformations.

L'objectif principal de cette thèse était de retracer l'histoire évolutive des éléments *mariner* (*MLE*) et *piggyBac* chez des pucerons des céréales (*Rhopalosiphum padi, Sitobion avenae, R. maidis* et *Schizaphis graminum*) et de fournir des informations sur leurs caractéristiques avant d'envisager de les utiliser comme des outils potentiels de transformation chez ces ravageurs des céréales. Par ailleurs, en raison des fortes interactions entre les pucerons et leurs plantes hôtes, il était intéressant de rechercher l'éventuelle existence de transferts horizontaux. En effet, les proximités et les échanges qui existent entre les deux partenaires d'une relation de type hôte/parasite, en font d'excellents modèles pour aborder les mécanismes de tels transferts, en particulier chez les eucaryotes.

Dans le premier chapitre, trois génomes de pucerons (*Acyrthosiphon pisum, Myzus persicae* et *Diuraphis noxia*), disponibles dans les banques de données, ont été exploités pour rechercher les ETs appartenant à différentes sous-familles de *mariner* et déterminer leur distribution. Trois clades ont été identifiés à savoir :

- Le clade de la sous-famille *irritans* DD34D commune aux trois espèces et subdivisée en trois tribus *Macrosiphinimar, Dnomar-like* elements et *Batmar-like* elements. Elle inclut plusieurs éléments complets potentiellement actifs et des MITEs.
- le clade *rosa* DD41D, dont le premier élément *crmar2* découvert chez la mouche méditerranéenne des fruits *Ceratitis capitata*, a été initialement classé avec la famille *mariner* DD34D (Gomulski *et al.* 2001).
- Un nouveau clade *Long Terminal Inverted Repeats* (*LTIR*- like elements) subdivisé en deux tribus DD40-41D, en fonction des motifs de la triade catalytique.

Ces deux derniers clades semblent être plus répandus chez *A. pisum* représentés par quelques éléments complets et des MITEs. La caractérisation et l'analyse phylogénétique de ces copies suggèrent que *rosa* et *LTIR-like* elements partagent un ancêtre commun proche, pouvant ainsi former une nouvelle famille appartenant à la superfamille *Tc1-mariner-IS630* et semblent être plus liés à la famille des *maT* DD37D que de *Tc1* ou *mariner*.

Ce premier chapitre montre que des éléments de la famille *mariner*, ou apparentés à cette famille, sont présents chez les trois espèces étudiées. Même s'il reste à déterminer si les copies complètes sont actives ou non, c'est un point important si l'on souhaite les utiliser pour faire de la transgénèse ou de la mutagenèse insertionnelle.

Dans le deuxième chapitre, nous nous sommes focalisés sur la recherche des éléments de la famille *irritans, in vitro* chez quatre espèces de pucerons, ainsi qu'*in silico* et *in vitro* chez leur plante hôte à savoir les céréales. Deux types de séquences ont été identifiés. Le premièr est commun aux différentes espèces de pucerons. Il a une taille d'environ 950 pb avec une délétion interne du coté 5'. Le deuxième est spécifique à *S. avenae* et à *Hordeum vulgare,* a une taille d'environ 650 pb et il est dépourvu du domaine catalytique. De plus, une séquence *irritans* tronquée du coté 5' a été trouvée chez deux cultivars différents de l'orge (*barke et Roho*) par des approches in *silico* et *in vitro* respectivement. Dans le contig du cultivar *barke*, elle est entourée par de l'ADN génomique de pucerons, le tout flanqué par de l'ADN génomique de l'orge dont la séquence n'est pas caractérisée. Cette étude a donc permis de mettre en évidence un éventuel transfert horizontal d'un élément ancien de la sous-famille *irritans* entre le puceron et le génome de l'hôte. Ceci dit, un point qui n'a pu être résolu est de savoir si la séquence détectée dans le cultivar *barke* est le résultat d'un transfert de l'élément transposable ou d'une région plus importante incluant de l'ADN génomique du puceron, elle même pouvant éventuellement être bordée par des séquences d'un élément transposable de la même famille ou d'une autre famille. En effet, l'étude des régions bordant l'ensemble de l'insertion, ne permet pas de conclure. Il s'agit peut-être d'une insertion ancienne ayant subi des mutations et/ou des réarrangements qui ont effacé les traces des évènements successifs.

Les résultats de ces deux chapitres montrent que malgré l'étroite relation de parenté entre les espèces de pucerons de la sous-famille des Aphidinae, la distribution et la prévalence des éléments *irritans* de la famille *mariner,* mais aussi celles de *rosa* et de *LTIR-like* elements, sont différentes. Cette variabilité reflète vraisemblablement des histoires évolutives indépendantes de ces éléments qui peuvent être le résultat (i) d'invasions verticales liées à un polymorphisme ancestral (Capy *et al.* 1994; Green et Frommer 2001; Prasad *et al.* 2002), suivies de pertes stochastiques dues à l'accumulation de mutations indépendantes et de l'extinction verticale (Le Rouzic *et al.* 2007),

et/ou, (ii) de transferts horizontaux (Kidwell 1992; Wallau *et al.* 2012). Cette dynamique des éléments, différente d'une espèce à une autre (Capy *et al.* 1992; Kidwell 2002; Vieira *et al.* 2002; Wallau *et al.* 2014), voire d'une population à une autre (Barrón *et al.* 2014), confirme que la connaissance des ETs dans une espèce, ne permet pas d'inférer sur celle-ci chez une autre espèce, même si cette dernière est phylogénétiquement proche, d'où l'importance d'avoir une bonne connaissance de l'équipement en ETs et de leurs structures au sein de chaque génome d'intérêt.

Il faut également souligner que la majorité des éléments caractérisés dans ces deux chapitres sont des séquences délétées ou MITEs. Ces délétions sont variables en termes de taille, de localisation et de présence de microhomologies au niveau des sites de cassures (Brunet *et al.* 1999, 2002; Negoua *et al.* 2013). Toutefois, ces éléments pourraient être mobilisés en *trans* par la transposase intacte d'un élément autonome de la même sous-famille ou, par l'intermédiaire d'une transposase provenant d'un *MLE* d'une autre sous-famille proche (Bigot *et al.* 2005). L'invasion de ces éléments peut également dépendre de leur taille (taille restreinte à environ 900 pb) et /ou de leur structure (Brunet *et al.* 1996, Lohe et Hartl 1996). Ceci dit, un MITE ne se limite pas à une simple délétion interne. D'autres caractéristiques sont vraisemblablement indispensables, mais jusqu'à présent, nous ne les connaissons pas. A titre d'exemple, les séquences capables de former des structures secondaires stables pourraient être favorisées (Dufresne *et al.* 2007). D'autres paramètres internes peuvent également influencer l'efficacité de transposition comme cela a été suggéré à partir des essais de création de MITEs synthétiques (Bergemann *et al.* 2008).

Au sein de la famille *mariner*, nos recherches *in silico* n'ont abouti qu'à l'identification d'éléments de la sous-famille *irritans*. Néanmoins des éléments des sous-familles *mellifera* et *mauritiana* ont été détectés *in vitro*, respectivement, chez le puceron du soja *Aphis gossypii* et les pucerons des arbres fruitiers (Mittapalli *et al.* 2011; Kharrat *et al.* 2015). Ces espèces appartiennent à des tribus phylogénétiquement proches. Ces résultats nous ont incités à explorer *in vitro* ces deux sous-familles (*mellifera* et *mauritiana*) chez les pucerons des céréales. L'analyse préliminaire n'a montré que la présence d'éléments de la sous-famille *mauritiana* avec deux types de copies délétées. Le premier est commun à tous les pucerons avec une taille de 917 pb présentant un pourcentage d'identité supérieur à 97%, le deuxième est spécifique aux pucerons des céréales, avec une taille d'environ 720 pb. Une identification et caractérisation plus approfondies seront prochainement développées.

En conclusion, bien que des éléments appartenant à différentes sous-familles peuvent cohabiter dans le génome de l'hôte (Green et Frommer 2001), plusieurs paramètres tels que le taux d'excision ou d'insertion, la vitesse d'accumulation des mutations, les recombinaisons ectopiques

ainsi que la sélection peuvent influencer la prévalence d'un ET par rapport à un autre (Wright et Schoen 1999; Le Rouzic *et al.* 2007; Biémont 2008).

En outre, les deux séquences *irritans*, trouvées dans le génome de deux cultivars de l'orge, suggèrent un éventuel transfert horizontal. L'identification de régions flanquantes correspondant à des séquences connues de l'ADN génomique de l'orge constitue un argument supplémentaire pour appuyer cette hypothèse. De plus si ces régions sont identiques pour ces deux cultivars ceci pourrait suggérer qu'ils dérivent d'un même événement. L'absence des séquences *irritans* dans les autres génomes de céréales hôtes peut être expliquée par (i) l'éloignement phylogénétique entre les ancêtres de ces espèces, ou par (ii) le défaut d'annotation des génomes se traduisant souvent par l'élimination de plusieurs séquences répétées.

La caractérisation des éléments appartenant à cette superfamille *Tc1-mariner-IS630* et l'identification d'une nouvelle famille regroupant les éléments *rosa* et *LTIR-like* elements, nous ont incités à nous poser la question de savoir s'il est nécessaire de remettre à jour l'ancienne classification de cette superfamille établie par Shao et Tu (2001). Etant donné le nombre croissant des génomes séquencés, il serait préférable d'attendre d'avoir une analyse globale exhaustive de ces ETs avant de redéfinir des groupes sur la base des TIRs, des régions UTR et/ou de la triade catalytique, ce qui permettrait d'éviter toutes confusions dans la catégorisation des éléments intra-famille.

L'objectif initial du troisième chapitre était d'appliquer la même procédure de recherche des ETs, précédemment décrite pour les éléments *mariner*, à la superfamille des *piggyBac*. Cependant, le peu de travaux portant sur ces éléments et leur évolution, nous a encouragés à entreprendre d'abord l'identification de tous les éléments *piggyBac*, leur caractérisation, leur distribution et leur évolution. Ainsi, à partir de 117 séquences protéiques, déjà publiées dans les banques de données, correspondant à des éléments complets *piggyBac-like* elements que nous avons désignés PBLE et des éléments domestiqués PGBD, nous avons identifié sur la base des similitudes, 50 PBLE, 295 PGBD et 101 séquences annotées PGBD-like. Tous les PBLE (potentiellement actifs ou non) présentent deux TIR, deux UTR et une transposase. Les PGBD sont des séquences à copie unique au sein du génome, orthologues entre les espèces et sous pression de sélection (positive ou purifiante). Les PGBD-like rassemblent toutes les autres séquences avec un ou sans TIR, UTR et une transposase qui peut être partiellement tronquée.

Dans un premier temps, il est apparu que les PBLE formaient quatre groupes structuraux en fonction de la présence ou de l'absence de STIR et de DR. Par ailleurs, aucune congruence entre la phylogénie des espèces, la phylogénie des transposons et celle des groupes structuraux n'a été

observée. Ceci nous a permis de proposer l'hypothèse que le groupe le plus largement distribué et qui contient uniquement des TIRs, pourrait correspondre à la structure ancestrale, les trois autres groupes proviendraient d'acquisitions indépendantes (convergences) de STIR et de DR, de transferts horizontaux et seraient sujets à une évolution rapide. La présence de répétions internes (DR/STIR), parfois sous la forme de motifs palindromiques, serait, quant à elle, potentiellement impliquée dans la liaison de la transposase et la stabilisation du complexe transposase-DR/STIR. Par ailleurs, le domaine riche en cystéines (CRD) en C-terminale de la transposase, ressemble à une structure de type PHD-like finger domain (Plant HomeoDomain) mais ne présente pas de résidu(s) potentiel(s) de liaison à l'un des deux motifs en doigt de zinc entremêlés (Vogt et Mochizuki 2013). Serait-il un domaine de fixation à l'ADN ? La perte de cette région conduirait-elle à la diminution de l'activité de transposition ? Y aurait-il une corrélation entre les groupes structuraux des PBLE et le domaine C-terminal ? Dans ce dernier cas, il faudrait étudier l'activité de transposition de chaque structure de PBLE en relation avec le polymorphisme de la région CRD. Il serait intéressant d'apporter des preuves expérimentales pour vérifier cette hypothèse. Cela permettrait d'établir une éventuelle coévolution entre la transposase et les séquences terminales répétées.

Dans un deuxième temps, nous avons caractérisé en détails les huit groupes d'éléments domestiqués, précédemment décrits dans la littérature (PGBD1-5, KOBUTA, *piggyMac*, TPB2). De plus, nous avons mis en évidence un nouveau groupe d'éléments domestiqués chez les néoptérygiens. Ce groupe a été nommé NeoPGBD. Toutes les séquences domestiquées sont soumises à une forte sélection purifiante et présentent un domaine catalytique qui n'est pas toujours conservé par rapport à la transposase ancestrale. Ceci dit, il faut noter que la domestication de ces séquences n'est pas systématiquement accompagnée de modifications de l'activité de la transposase (Sarkar *et al.* 2003; Keith *et al.* 2008; Pavelitz *et al.* 2013). En d'autres termes, les caractéristiques de la transposase ancestrale ont parfois été récupérées pour de nouvelles fonctions.

Alors que trois fonctions ont été élucidées pour les quatre groupes PGBD3, PGM, TPB2 et PGBD5 (Newman *et al.* 2008; Baudry *et al.* 2009; Cheng *et al.* 2010; Pavelitz *et al.* 2013), des recherches *in vitro* sur les autres groupes conservés depuis plusieurs millions d'années, seraient nécessaires afin de déterminer leur nouvelle fonction. De plus, il serait intéressant d'étudier l'évolution des régions flanquantes à proximité de ces séquences afin de déterminer l'existence de balayage sélectif et de préciser même grossièrement la date de ces domestications. Enfin, bien que plusieurs caractéristiques des éléments *piggyBac* soient identifiées telles que l'excision précise du transposon (Mitra *et al.* 2008), la présence de site d'épissage potentiel en 5'UTR et de signal de polyAdénylation en 3'UTR (Newman *et al.* 2008) ou la possibilité de fusion N-terminale de la transposase tout en gardant une activité de transposition et de colonisation (Wu *et al.* 2006), il est

difficile de déterminer le ou les mécanismes responsables du succès de leur domestication. Il semble, à première vue, que certains éléments transposables soient plus sujets à la domestication. C'est le cas des éléments *piggyBac* mais également des éléments *Mutator* qui sont également des éléments *de Classe II* (Joly-Lopez et al, 2016). Toutefois, avant de proposer une telle conclusion, il faudrait avoir une vision plus globale sur l'ensemble des éléments transposables.

Dans un autre domaine, il serait également intéressant de rechercher les éléments *piggyBac* endogènes potentiellement fonctionnels chez les pucerons des céréales, afin de déterminer les séquences les plus efficaces qui pourraient être utilisées pour faire de la transgénèse et/ou de la mutagenèse. A ce titre, *piggyBac* a déjà été utilisé, avec succès, pour faire de la transgénèse chez la mouche de l'olive *Bactrocera oleae*, qui est dépourvu d'éléments *piggyBac* (Ant *et al.* 2012; Genc *et al.* 2016). Ceci dit, avant de se lancer dans ce type de transformation, il est important de rechercher si des copies endogènes complètes et/ou des MITEs de l'élément utilisé pour faire la transgénèse, existent dans le génome de l'espèce à transformer. En effet, si tel est le cas, cela conduira vraisemblablement à une forte instabilité génétique. Il est également important de souligner que l'efficacité des ETs comme vecteur de transfert de gènes dépend, non seulement de l'élément lui-même, mais aussi de l'organisme receveur et du type cellulaire (Izsvák *et al.* 2000 ; Wu *et al.* 2006). Il est donc intéressant de ne pas se limiter à l'optimisation d'un seul élément pour son utilisation en biotechnologie, mais plutôt envisager des systèmes de transposons complémentaires tels que *Frog Prince* et *Sleeping Beauty* (Miskey *et al.* 2003) ou des systèmes chimériques de transposons comme *piggyBac* et *mos1* (Maragathavally *et al.* 2006).

Enfin, il est difficile de ne pas formuler des hypothèses quant à l'impact des ETs sur la spécialisation (adaptation) des pucerons en fonction des plantes hôtes. Par exemple, le puceron du cotonnier *Aphis gossypii* et celui des agrumes *Aphis spiraecola* sont des pucerons spécialisées sur des races d'hôtes uniques (Vanlerberghe-Masutti et Chavigny 1998; Charaabi *et al.* 2008; Mezghani-Khemakhem *et al.* 2012) alors que le puceron de la pomme de terre *Macrosiphum euphorbiae* est une espèce généraliste qui est capable de coloniser plusieurs familles botaniques (Raboudi *et al.* 2012). L'activité des ETs représente une source non négligeable de variation d'autant plus qu'ils pourraient être mobilisés suite à des stress environnementaux ou génomiques et être ainsi à l'origine d'un certain nombre d'adaptation (Capy *et al.* 2000; Biémont et Vieira 2006; González *et al.* 2008; Jiang *et al.* 2017). Néanmoins, le génome est capable de réprimer épigénétiquement, notamment par la voie des *si*RNA, l'activité des éléments transposables dans la lignée germinale (Siomi *et al.* 2011) et avoir un impact sur l'expression des gènes ou des séquences régulatrices au voisinage des ETs (Fablet et Vieira 2011). Ainsi la comparaison de la distribution et de la diversité des ETs dans plusieurs génomes de pucerons serait intéressante pour éventuellement mieux comprendre l'adaptation des espèces en fonction des plantes hôtes.

Références Bibliographiques

Abrouk, M., Murat, F., Pont, C., Messing, J., Jackson, S., Faraut, T et al. (2010). Palaeogenomics of plants: synteny-based modelling of extinct ancestors. *Trends in plant science*, *15*(9), 479-487.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, *215*(3), 403-410.

Anderson-Gerfaud, P. J., Deraprahamian, G. & Willcox, G. (1991). Les premières cultures de céréales sauvages et domestiques primitives au Proche-Orient néolithique: résultats préliminaires d'expériences à Jalès (Ardéche). Cahiers Euphrate 5-6, 191-232.

Ant, T., Koukidou, M., Rempoulakis, P., Gong, H. F., Economopoulos, A., Vontas, J., & Alphey, L. (2012). Control of the olive fruit fly using genetics-enhanced sterile insect technique. *BMC biology*, *10*(1), 1.

Anxolabéhère, D., Kidwell, M. G., & Periquet, G. (1988). Molecular characteristics of diverse populations are consistent with the hypothesis of a recent invasion of Drosophila melanogaster by mobile P elements. *Molecular biology and evolution*, *5*(3), 252-269.

Anxolabéhère, D., Nouaud, D., Quesneville, H., & Ronsseray, S. (2007). Transposons: des gènes. *Pour la science*, (351), 82.

Auge-Gouillou, C., Brillet, B., Germon, S., Hamelin, M. H., & Bigot, Y. (2005a). Mariner Mos1 transposase dimerizes prior to ITR binding. *Journal of molecular biology*, *351*(1), 117-130.

Augé-Gouillou, C., Brillet, B., Hamelin, M. H., & Bigot, Y. (2005b). Assembly of the mariner Mos1 synaptic complex. *Molecular and cellular biology*, *25*(7), 2861-2870.

Awazu, S., Matsuoka, T., Inaba, K., Satoh, N., & Sasakura, Y. (2007). High-throughput enhancer trap by remobilization of transposon Minos in Ciona intestinalis. *Genesis*, *45*(5), 307-317.

Bannai, H., Tamada, Y., Maruyama, O., Nakai, K., & Miyano, S. (2002). Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics*, *18*(2), 298-305.

Barrón, M. G., Fiston-Lavier, A. S., Petrov, D. A., & González, J. (2014). Population genomics of transposable elements in Drosophila. *Annual review of genetics*, *48*, 561-581.

Barry, E. G., Witherspoon, D. J., & Lampe, D. J. (2004). A bacterial genetic screen identifies functional coding sequences of the insect mariner transposable element Famar1 amplified from the genome of the earwig, Forficula auricularia. *Genetics*, *166*(2), 823-833.

Baudry, C., Malinsky, S., Restituito, M., Kapusta, A., Rosa, S., Meyer, E., & Bétermier, M. (2009). PiggyMac, a domesticated piggyBac transposase involved in programmed genome rearrangements in the ciliate Paramecium tetraurelia. *Genes & development*, *23*(21), 2478-2483.

Bennetzen, J. L., Ma, J., & Devos, K. M. (2005). Mechanisms of recent genome size variation in flowering plants. *Annals of botany*, *95*(1), 127-132.

Bergemann, M., Lespinet, O., M'Barek, S. B., Daboussi, M. J., & Dufresne, M. (2008). Genome-wide analysis of the Fusarium oxysporum mimp family of MITEs and mobilization of both native and de novo created mimps. *Journal of molecular evolution*, *67*(6), 631-642.

Biémont, C. (2008). Within-species variation in genome size. *Heredity*, *101*, 297-298.

Biémont, C. (2010). A brief history of the status of transposable elements: from junk DNA to major players in evolution. *Genetics*, *186*(4), 1085-1093.

Biémont, C., & Vieira, C. (2006). Genetics: junk DNA as an evolutionary force. *Nature*, *443*(7111), 521-524.

Biemont, C., Tsitrone, A., Vieira, C., & Hoogland, C. (1997). Transposable element distribution in Drosophila. *Genetics*, *147*(4).

Bigot, Y., Brillet, B., & Auge-Gouillou, C. (2005). Conservation of palindromic and mirror motifs within inverted terminal repeats of mariner-like elements. *Journal of molecular biology*, *351*(1), 108-116.

Bingham, P. M., Kidwell, M. G., & Rubin, G. M. (1982). The molecular basis of PM hybrid dysgenesis: the role of the P element, a P-strain-specific transposon family. *Cell*, *29*(3), 995-1004.

Blackman, R. K., Koehler, M. M., Grimaila, R., & Gelbart, W. M. (1989). Identification of a fully-functional hobo transposable element and its use for germ-line transformation of Drosophila. *The EMBO Journal*, *8*(1), 211-217.

Blackman, R. L., & Eastop, V. F. (2000). *Aphids on the world's crops: An identification and information guide*, 2nd edn. Wiley Ltd., Chichester.

Blake, N. K., Lehfeldt, B. R., Lavin, M., & Talbert, L. E. (1999). Phylogenetic reconstruction based on low copy DNA sequence data in an allopolyploid: the B genome of wheat. *Genome*, *42*(2), 351-360.

Böhne, A., Brunet, F., Galiana-Arnoux, D., Schultheis, C., & Volff, J. N. (2008). Transposable elements as drivers of genomic and biological diversity in vertebrates. *Chromosome Research*, *16*(1), 203-215.

Bonin, C. P., & Mann, R. S. (2004). A piggyBac transposon gene trap for the analysis of gene expression and function in Drosophila. *Genetics*, *167*(4), 1801-1811.

Brandström, M., & Ellegren, H. (2007). The genomic landscape of short insertion and deletion polymorphisms in the chicken (Gallus gallus) genome: a high frequency of deletions in tandem duplicates. *Genetics*, *176*(3), 1691-1701.

Brault, V., Herrbach, E., Hauser, S., & Lemaire, O. (2001). Les Luteoviridae: propriétés biologiques et évolution. *Virologie*, *5*(1), 9-21.

Brenchley, R., Spannagl, M., Pfeifer, M., Barker, G. L., D'Amore, R., Allen, A. M. et al. (2012). Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature*, *491*(7426), 705-710.

Brillet, B., Y. Bigot, et C. Augé-Gouillou. (2007). Assembly of the Tc1 and mariner transposition initiation complexes depends on the origins of their transposase DNA binding domains. *Genetica*, *130*(2), 105-120.

Britten, R. J. (1996). DNA sequence insertion and evolutionary variation in gene regulation. *Proceedings of the National Academy of Sciences*, *93*(18), 9374-9377.

Brookfield, J. F., Montgomery, E., & Langley, C. H. (1984). Apparent absence of transposable elements related to the P elements of D. melanogaster in other species of Drosophila. *Nature*, 310, 330-332.

Brown, P. A., & Blackman, R. L. (1988). Karyotype variation in the corn leaf aphid, Rhopalosiphum maidis (Fitch), species complex (Hemiptera: Aphididae) in relation to host-plant and morphology. *Bulletin of Entomological Research*, *78*(02), 351-363.

Brunet, F., Giraud, T., Godin, F., & Capy, P. (2002). Do deletions of Mos1-like elements occur randomly in the Drosophilidae family?. *Journal of molecular evolution*, *54*(2), 227-234.

Brunet, F., Godin, F., Bazin, C., & Capy, P. (1999). Phylogenetic analysis of Mos1-like transposable elements in the Drosophilidae. *Journal of molecular evolution*, *49*(6), 760-768.

Brunet, F., Godin, F., Bazin, C., David, J. R., & Capy, P. (1996). The mariner transposable element in natural populations of Drosophila teissieri. *Journal of molecular evolution*, *42*(6), 669-675.

Bucheton, A., Paro, R., Sang, H. M., Pelisson, A., & Finnegan, D. J. (1984). The molecular basis of IR hybrid dysgenesis in Drosophila melanogaster: identification, cloning, and properties of the I factor. *Cell*, *38*(1), 153-163.

Bui, Q. T., Casse, N., Leignel, V., Nicolas, V., & Chénais, B. (2008). Widespread occurence of mariner transposons in coastal crabs. *Molecular phylogenetics and evolution*, *47*(3), 1181-1189.

Bureau, T. E., & Wessler, S. R. (1994). Stowaway: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. *The Plant Cell*, *6*(6), 907-916.

Burki, F., & Kaessmann, H. (2004). Birth and adaptive evolution of a hominoid gene that supports high neurotransmitter flux. *Nature genetics*, *36*(10), 1061-1063.

Calvi, B. R., Hong, T. J., Findley, S. D., & Gelbart, W. M. (1991). Evidence for a common evolutionary origin of inverted repeat transposons in Drosophila and plants: hobo, Activator, and Tam3. *Cell*, *66*(3), 465-471.

Canova, A., & Quaglia A. (1960). The Soil-borne wheat mosaic. *Informatore Fitopatologico 10*(10), 206-208.

Capy, P., Anxolabéhère, D., & Langin, T. (1994). The strange phylogenies of transposable elements: are horizontal transfers the only explanation?. *Trends in Genetics*, *10*(1), 7-12.

Capy, P., David, J. R., & Hartl, D. L. (1992a). Evolution of the transposable element mariner in the Drosophila melanogaster species group. *Genetica*, *86*(1-3), 37.

Capy, P., Gasperi, G., Biémont, C., & Bazin, C. (2000). Stress and transposable elements: co-evolution or useful parasites?. *Heredity*, *85*(2), 101-106.

Capy, P., Koga, A., David, J. R., & Hartl, D. L. (1992b). Sequence analysis of active mariner elements in natural populations of Drosophila simulans. *Genetics*, *130*(3), 499-506.

Capy, P., Langin, T., Higuet, D., Maurer, P., & Bazin, C. (1997). Do the integrases of LTR-retrotransposons and class II element transposases have a common ancestor?. *Genetica*, *100*(1-3), 63-72.

Carpes, M. P., Nunes, J. F., Sampaio, T. L., Castro, M. E. B., Zanotto, P. M. A., & Ribeiro, B. M. (2009). Molecular analysis of a mutant Anticarsia gemmatalis multiple nucleopolyhedrovirus (AgMNPV) shows an interruption of an inhibitor of apoptosis gene (iap-3) by a new class-II piggyBac-related insect transposon. *Insect molecular biology*, *18*(6), 747-757.

Cary, L. C., Goebel, M., Corsaro, B. G., Wang, H. G., Rosen, E., & Fraser, M. J. (1989). Transposon mutagenesis of baculoviruses: analysis of Trichoplusia ni transposon IFP2 insertions within the FP-locus of nuclear polyhedrosis viruses. *Virology*, *172*(1), 156-169.

Casse, N., Bui, Q. T., Nicolas, V., Renault, S., Bigot, Y., & Laulier, M. (2006). Species sympatry and horizontal transfers of Mariner transposons in marine crustacean genomes. *Molecular phylogenetics and evolution*, *40*(2), 609-619.

Catania, F., Kauer, M. O., Daborn, P. J., Yen, J. L., Ffrench-Constant, R. H., & Schlötterer, C. (2004). World-wide survey of an Accord insertion and its association with DDT resistance in Drosophila melanogaster. *Molecular ecology*, *13*(8), 2491-2504.

Chalupska, D., Lee, H. Y., Faris, J. D., Evrard, A., Chalhoub, B., Haselkorn, R., & Gornicki, P. (2008). Acc homoeoloci and the evolution of wheat genomes. *Proceedings of the National Academy of Sciences*, *105*(28), 9691-9696.

Charaabi, K., Carletto, J., Chavigny, P., Marrakchi, M., Makni, M., & Vanlerberghe-Masutti, F. (2008). Genotypic diversity of the cotton-melon aphid Aphis gossypii (Glover) in Tunisia is structured by host plants. *Bulletin of entomological research*, *98*(04), 333-341.

Charbonnier, E., Ronceux, A., Carpentier, A. S., Soubelet, H., & Barriuso, E. (2016). Pesticides: des impacts aux changements de pratiques. Editions Quae.

Charles, M., Belcram, H., Just, J., Huneau, C., Viollet, A., Couloux, A., et al. (2008). Dynamics and differential proliferation of transposable elements during the evolution of the B and A genomes of wheat. *Genetics*, *180*(2), 1071-1086.

Cheng, C. Y., Vogt, A., Mochizuki, K., & Yao, M. C. (2010). A domesticated piggyBac transposase plays key roles in heterochromatin dynamics and DNA cleavage during programmed DNA deletion in Tetrahymena thermophila. *Molecular biology of the cell*, *21*(10), 1753-1762.

Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature, 437(7055), 69-87.*

Chinwalla, A. T., Cook, L. L., Delehaunty, K. D., Fewell, G. A., Fulton, L. A., Fulton, R. S., et al. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, *420*(6915), 520-562.

Chougule, N. P., & Bonning, B. C. (2012). Toxins for transgenic resistance to hemipteran pests. *Toxins*, *4*(6), 405-429.

Chougule, N. P., Li, H., Liu, S., Linz, L. B., Narva, K. E., Meade, T., & Bonning, B. C. (2013). Retargeting of the Bacillus thuringiensis toxin Cyt2Aa against hemipteran insect pests. *Proceedings of the National Academy of Sciences*, *110*(21), 8465-8470.

Clark, A. G., Eisen, M. B., Smith, D. R., Bergman, C. M., Oliver, B., Markow, T. A., et al. (2007). Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, *450*(7167), 203-218.

Clark, J. B., Maddison, W. P., & Kidwell, M. G. (1994). Phylogenetic analysis supports horizontal transfer of P transposable elements. *Molecular biology and evolution*, *11*(1), 40-50.

Clark, K. J., Geurts, A. M., Bell, J. B., & Hackett, P. B. (2004). Transposon vectors for gene-trap insertional mutagenesis in vertebrates. *Genesis*, *39*(4), 225-233.

Claudianos, C., Brownlie, J., Russell, R., Oakeshott, J., & Whyard, S. (2002). maT-a clade of transposons intermediate between mariner and Tc1. *Molecular biology and evolution*, *19*(12), 2101-2109.

Cordaux, R., Udit, S., Batzer, M. A., & Feschotte, C. (2006). Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. *Proceedings of the National Academy of Sciences*, *103*(21), 8101-8106.

Craig N.L., Gracie R., Gellert M., & Lambowitz A.M. Mobile dna ii second edition. ASM Press, 2002.

Crooks, G. E., Hon, G., Chandonia, J. M., & Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome research*, *14*(6), 1188-1190.

Cunfer, B. M., & Scolari, B. L. (1982). Xanthomonas campestris pv. translucens on Triticale and Other Small Grains. *Phytopathology*, *72*(6), 683-686.

Daimon, T., Mitsuhiro, M., Katsuma, S., Abe, H., Mita, K., & Shimada, T. (2010). Recent transposition of yabusame, a novel piggyBac-like transposable element in the genome of the silkworm, Bombyx mori. *Genome*, *53*(8), 585-593.

Daniels, S. B., Peterson, K. R., Strausbaugh, L. D., Kidwell, M. G., & Chovnick, A. (1990). Evidence for horizontal transmission of the P transposable element between Drosophila species. *Genetics*, *124*(2), 339-355.

Daron, J., Glover, N., Pingault, L., Theil, S., Jamilloux, V., Paux, E., et al. (2014). Organization and evolution of transposable elements along the bread wheat chromosome 3B. *Genome biology*, *15*(12), 1.

Davidson, A. E., Balciunas, D., Mohn, D., Shaffer, J., Hermanson, S., Sivasubbu, S., et al. (2003). Efficient gene delivery and gene expression in zebrafish using the Sleeping Beauty transposon. *Developmental biology*, *263*(2), 191-202.

Dawkins, R. (1990). Le gène égoïste. Edts Armand Collin.

Dawson, A., & Finnegan, D. J. (2003). Excision of the Drosophila mariner transposon Mos1: comparison with bacterial transposition and V (D) J recombination. *Molecular cell*, *11*(1), 225-235.

Delaurière, L., Chénais, B., Hardivillier, Y., Gauvry, L., & Casse, N. (2009). Mariner transposons as genetic tools in vertebrate cells. *Genetica*, *137*(1), 9-17.

Deragon, J. M., & Capy, P. (2000). Impact of transposable elements on the human genome. *Annals of medicine*, *32*(4), 264-273.

Devos, K. M., Brown, J. K., & Bennetzen, J. L. (2002). Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. *Genome research*, *12*(7), 1075-1079.

Dewannieux, M., Esnault, C., & Heidmann, T. (2003). LINE-mediated retrotransposition of marked Alu sequences. *Nature genetics*, *35*(1), 41-48.

Diamond, J. (2002). Evolution, consequences and future of plant and animal domestication. *Nature*, *418*(6898), 700-707.

Djili, K., Daoud, Y., Gaouar, A., & Beldjoudi, Z. (2003). La salinisation secondaire des sols au Sahara. Conséquences sur la durabilité de l 'agriculture dans les nouveaux périmètres de mise en valeur. *Science et changements planétaires/Sécheresse*, *14*(4), 241-246.

Doak, T. G., Doerder, F. P., Jahn, C. L., & Herrick, G. (1994). A proposed superfamily of transposase genes: transposon-like elements in ciliated protozoa and a common" D35E" motif. *Proceedings of the National Academy of Sciences*, *91*(3), 942-946.

Dombroski, B. A., Mathias, S. L., Nanthakumar, E., Scott, A. F., & Kazazian Jr, H. H. (1991). Isolation of an active human transposable element. *Science*, *254*(5039), 1805-1808.

Dowling, D., Pauli, T., Donath, A., Meusemann, K., Podsiadlowski, L., Petersen, M., et al. (2017). Phylogenetic Origin and Diversification of RNAi Pathway Genes in Insects. *Genome Biology and Evolution*, evw281.

Doyle JJ, Doyle JL (1987) A rapid DNA isolation procedure for small quantities of fresh leaves tissue. *Phytochemical Bulletin* 19, 11-15.

Duelli, P. (2001). Lacewings in field crops. Dans McEwen, F. K., T. R. New et A. E. Whittington (dir), *Lacewings in the crop environment* (p. 158-171). New York: Cambridge university press.

Dufresne, M., Hua-Van, A., El Wahab, H. A., M'Barek, S. B., Vasnier, C., Teysset, L., et al. (2007). Transposition of a fungal miniature inverted-repeat transposable element through the action of a Tc1-like transposase. *Genetics*, *175*(1), 441-452.

Dupeyron, M., Leclercq, S., Cerveau, N., Bouchon, D., & Gilbert, C. (2014). Horizontal transfer of transposons between and within crustaceans and insects. *Mobile DNA*, *5*(1), 1.

Duplouy, A., Iturbe-Ormaetxe, I., Beatson, S. A., Szubert, J. M., Brownlie, J. C., McMeniman, C. J., et al. (2013). Draft genome sequence of the male-killing Wolbachia strain w Bol1 reveals recent horizontal gene transfers from diverse sources. *BMC genomics*, *14*(1), 1.

Dupuy, A. J., Akagi, K., Largaespada, D. A., Copeland, N. G., & Jenkins, N. A. (2005). Mammalian mutagenesis using a highly mobile somatic Sleeping Beauty transposon system. *Nature*, *436*(7048), 221-226.

Dvorak, J., Akhunov, E. D., Akhunov, A. R., Deal, K. R., & Luo, M. C. (2006). Molecular characterization of a diagnostic DNA marker for domesticated tetraploid wheat provides evidence for gene flow from wild tetraploid wheat to hexaploid wheat. *Molecular Biology and Evolution*, *23*(7), 1386-1396.

El Baidouri, M., Carpentier, M. C., Cooke, R., Gao, D., Lasserre, E., Llauro, C., et al. (2014). Widespread and frequent horizontal transfers of transposable elements in plants. *Genome research*, *24*(5), 831-838.

Elick, T. A., Lobo, N., & Fraser Jr, M. J. (1997). Analysis of the cis-acting DNA elements required for piggyBac transposable element excision. *Molecular and General Genetics*, *255*(6), 605-610.

Emmons, S. W., Yesner, L., & Katzenberg, D. (1983). Evidence for a transposon in Caenorhabditis elegans. *Cell*, *32*(1), 55-65.

Essner, J. J., McIvor, R. S., & Hackett, P. B. (2005). Awakening gene therapy with Sleeping Beauty transposons. *Current opinion in pharmacology*, *5*(5), 513-519.

Evgen'ev, M. B., & Arkhipova, I. R. (2005). Penelope-like elements–a new class of retroelements: distribution, function and possible evolutionary significance. *Cytogenetic and genome research*, *110*(1-4), 510-521.

Fablet, M., & Vieira, C. (2011). Evolvability, epigenetics and transposable elements. *Biomolecular concepts*, *2*(5), 333-341.

Feldman, M., Lupton, F. G. H., & Miller, T. E. (1995). Wheats. In 'Evolution of crop plants'.(Eds J Smartt and NW Simmonds) pp. 184–192.

Feschotte, C. (2004). Merlin, a new superfamily of DNA transposons identified in diverse animal genomes and related to bacterial IS1016 insertion sequences. *Molecular biology and evolution*, *21*(9), 1769-1780.

Feschotte, C., & Mouches, C. (2000). Evidence that a family of miniature inverted-repeat transposable elements (MITEs) from the Arabidopsis thaliana genome has arisen from a pogo-like DNA transposon. *Molecular Biology and Evolution*, *17*(5), 730-737.

Feschotte, C., & Pritham, E. J. (2007). DNA transposons and the evolution of eukaryotic genomes. *Annual review of genetics*, *41*, 331-368.

Feschotte, C., & Wessler, S. R. (2001). Treasures in the attic: rolling circle transposons discovered in eukaryotic genomes. *Proceedings of the National Academy of Sciences*, *98*(16), 8923-8924.

Feschotte, C., Jiang, N., & Wessler, S. R. (2002). Plant transposable elements: where genetics meets genomics. *Nature Reviews Genetics*, *3*(5), 329-341.

Feschotte, C., Swamy, L., & Wessler, S. R. (2003). Genome-wide analysis of mariner-like transposable elements in rice reveals complex relationships with stowaway miniature inverted repeat transposable elements (MITEs). *Genetics*, *163*(2), 747-758.

Filée, J., Rouault, J. D., Harry, M., & Hua-Van, A. (2015). Mariner transposons are sailing in the genome of the blood-sucking bug Rhodnius prolixus. *BMC genomics*, *16*(1), 1.

Finnegan, D. J. (1989). Eukaryotic transposable elements and genome evolution. *Trends in genetics*, *5*, 103-107.

Finnegan, D. J. (1992). Transposable elements. *Current opinion in genetics & development*, *2*(6), 861-867.

Foottit, R. G., Maw, H. E. L., Von Dohlen, C. D., & Hebert, P. D. N. (2008). Species identification of aphids (Insecta: Hemiptera: Aphididae) through DNA barcodes. *Molecular Ecology Resources*, *8*(6), 1189-1201.

Fraser, M. J., Brusca, J. S., Smith, G. E., & Summers, M. D. (1985). Transposon-mediated mutagenesis of a baculovirus. *Virology*, *145*(2), 356-361.

Fraser, M. J., Clszczon, T., Elick, T., & Bauser, C. (1996). Precise excision of TTAA-specific lepidopteran transposons piggyBac (IFP2) and tagalong (TFP3) from the baculovirus genome in cell lines from two species of Lepidoptera. *Insect molecular biology*, *5*(2), 141-151.

Fraser, M. J., Smith, G. E., & Summers, M. D. (1983). Acquisition of host cell DNA sequences by baculoviruses: relationship between host DNA insertions and FP mutants of Autographa californica and Galleria mellonella nuclear polyhedrosis viruses. *Journal of virology*, *47*(2), 287-300.

Gale, M. D., & Devos, K. M. (1998). Comparative genetics in the grasses. *Proceedings of the National Academy of Sciences*, *95*(5), 1971-1974.

Gao, Y., Mathee, K., Narasimhan, G., & Wang, X. (1999). Motif detection in protein sequences. In *String Processing and Information Retrieval Symposium, 1999 and International Workshop on Groupware* (pp. 63-72). IEEE.

Genç, H., Schetelig, M. F., Nirmala, X., & Handler, A. M. (2016). Germline transformation of the olive fruit fly, Bactrocera oleae (Rossi)(Diptera: Tephritidae), with a piggyBac transposon vector. *Turkish Journal of Biology*, 40.

Gibbs, R. A., Rogers, J., Katze, M. G., Bumgarner, R., Weinstock, G. M., Mardis, E. R., et al. (2007). Evolutionary and biomedical insights from the rhesus macaque genome. *Science*, *316*(5822), 222-234.

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*(7293), 1347-1350.

Goff, S. A., Ricke, D., Lan, T. H., Presting, G., Wang, R., Dunn, M., et al. (2002). A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). *Science*, *296*(5565), 92-100.

Gomulski, L. M., Torti, C., Bonizzoni, M., Moralli, D., Raimondi, E., Capy, P., et al. (2001). A new basal subfamily of mariner elements in Ceratitis rosa and other tephritid flies. *Journal of molecular evolution*, *53*(6), 597-606.

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*(10), e251.

Goodwin, T. J., & Poulter, R. T. (2001). The DIRS1 group of retrotransposons. *Molecular biology and evolution*, *18*(11), 2067-2082.

Goodwin, T. J., Butler, M. I., & Poulter, R. T. (2003). Cryptons: a group of tyrosine-recombinase-encoding DNA transposons from pathogenic fungi. *Microbiology*, *149*(11), 3099-3109.

Grabher, C., Henrich, T., Sasado, T., Arenz, A., Wittbrodt, J., & Furutani-Seiki, M. (2003). Transposon-mediated enhancer trapping in medaka. *Gene*, *322*, 57-66.

Gray, L. T., Fong, K. K., Pavelitz, T., & Weiner, A. M. (2012). Tethering of the conserved piggyBac transposase fusion protein CSB-PGBD3 to chromosomal AP-1 proteins regulates expression of nearby genes in humans. *PLoS Genet*, *8*(9), e1002972.

Gray, Y. H. (2000). It takes two transposons to tango: transposable-element-mediated chromosomal rearrangements. *Trends in Genetics*, *16*(10), 461-468.

Green, C. L., & Frommer, M. (2001). The genome of the Queensland fruit fly Bactrocera tryoni contains multiple representatives of the mariner family of transposable elements. *Insect molecular biology*, *10*(4), 371-386.

Haffter, P., Granato, M., Brand, M., Mullins, M. C., Hammerschmidt, M., Kane, D. A., et al. (1996). The identification of genes with unique and essential functions in the development of the zebrafish, Danio rerio. *Development*, *123*(1), 1-36.

Han, F., Fedak, G., Guo, W., & Liu, B. (2005). Rapid and repeatable elimination of a parental genome-specific DNA repeat (pGc1R-1a) in newly synthesized wheat allopolyploids. *Genetics*, *170*(3), 1239-1245.

Handler, A. M. (2002). Use of the piggyBac transposon for germ-line transformation of insects. *Insect biochemistry and molecular biology*, *32*(10), 1211-1220.

Handler, A. M., Gomez, S. P., & O'Brochta, D. A. (1993). A functional analysis of the P-element gene-transfer vector in insects. *Archives of insect biochemistry and physiology*, *22*(3-4), 373-384.

Haniford, D., & Kleckner, N. (1994). Tn 10 transposition in vivo: temporal separation of cleavages at the two transposon ends and roles of terminal basepairs subsequent to interaction of ends. *The EMBO journal*, *13*(14), 3401.

Hellsten, U., Harland, R. M., Gilchrist, M. J., Hendrix, D., Jurka, J., Kapitonov, V., et al. (2010). The genome of the Western clawed frog Xenopus tropicalis. *Science*, *328*(5978), 633-636.

Henikoff, S. (1992). Detection of Caenorhabditis transposon homologs in diverse organisms. *The New biologist*, *4*(4), 382-388.

Henssen, A. G., Henaff, E., Jiang, E., Eisenberg, A. R., Carson, J. R., Villasante, C. M., et al. (2015). Genomic DNA transposition induced by human PGBD5. *Elife*, *4*, e10565.

Hikosaka, A., Kobayashi, T., Saito, Y., & Kawahara, A. (2007). Evolution of the Xenopus piggyBac transposon family TxpB: domesticated and untamed strategies of transposon subfamilies. *Molecular biology and evolution*, *24*(12), 2648-2656.

Hill, A. S., Foot, N. J., Chaplin, T. L., & Young, B. D. (2000). The most frequent constitutional translocation in humans, the t (11; 22)(q23; q11) is due to a highly specific Alu-mediated recombination. *Human molecular genetics*, *9*(10), 1525-1532.

Honeybee Genome Sequencing Consortium. (2006). Insights into social insects from the genome of the honeybee Apis mellifera. *Nature*, *443*(7114), 931.

Huang, H.C, Harper, A.M, Kokko, E.G., & Howard R.J. (1981). Aphids transmission of verticillum albo-atrum to alfalfa. *Plant Pathology*, *5*, 141-147.

Huang, S., Sirikhachornkit, A., Su, X., Faris, J., Gill, B., Haselkorn, R., & Gornicki, P. (2002). Genes encoding plastid acetyl-CoA carboxylase and 3-phosphoglycerate kinase of the Triticum/Aegilops complex and the evolutionary history of polyploid wheat. *Proceedings of the National Academy of Sciences*, *99*(12), 8133-8138.

Hua-Van, A., Le Rouzic, A., Maisonhaute, C., & Capy, P. (2005). Abundance, distribution and dynamics of retrotransposable elements and transposons: similarities and differences. *Cytogenetic and genome research*, *110*(1-4), 426-440.

Hurst, L. D. (2002). The Ka/Ks ratio: diagnosing the form of sequence evolution. *TRENDS in Genetics*, *18*(9), 486-487.

Hutchinson, J. B. (1929). The application of the" Method of Maximum Likelihood" to the estimation of linkage. *Genetics*, *14*(6), 519.

International Aphid Genomics Consortium. (2010). Genome sequence of the pea aphid Acyrthosiphon pisum. *PLoS Biol*, *8*(2), e1000313.

International Barley Genome Sequencing Consortium. (2012). A physical, genetic and functional sequence assembly of the barley genome. *Nature*, *491*(7426), 711-716.

International Wheat Genome Sequencing Consortium. (2014). A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. *Science*, *345*(6194), 1251788.

Iperti, G. (1999). Biodiversity of predaceous coccinellidae in relation to bioindication and economic importance. *Agriculture, ecosystems & environment*, *74*(1), 323-342.

Ivics, Z., & Izsvak, Z. (2006). Transposons for gene therapy!. *Current gene therapy*, *6*(5), 593-607.

Ivics, Z., Hackett, P. B., Plasterk, R. H., & Izsvák, Z. (1997). Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. *Cell*, *91*(4), 501-510.

Ivics, Z., Kaufman, C. D., Zayed, H., Miskey, C., Walisko, O., & Izsvák, Z. (2004). The Sleeping Beauty transposable element: evolution, regulation and genetic applications. *Current issues in molecular biology*, *6*, 43-56.

Ivics, Z., Li, M. A., Mátés, L., Boeke, J. D., Nagy, A., Bradley, A., & Izsvák, Z. (2009). Transposon-mediated genome manipulation in vertebrates. *Nature methods*, *6*(6), 415-422.

Izsvák, Z., Ivics, Z., & Plasterk, R. H. (2000). Sleeping Beauty, a wide host-range transposon vector for genetic transformation in vertebrates. *Journal of molecular biology*, *302*(1), 93-102.

Izsvák, Z., Ivics, Z., Shimoda, N., Mohn, D., Okamoto, H., & Hackett, P. B. (1999). Short inverted-repeat transposable elements in teleost fish and implications for a mechanism of their amplification. *Journal of molecular evolution*, *48*(1), 13-21.

Jacobs, G., Dechyeva, D., Menzel, G., Dombrowski, C., & Schmidt, T. (2004). Molecular characterization of Vulmar1, a complete mariner transposon of sugar beet and diversity of mariner-and En/Spm-like sequences in the genus Beta. *Genome*, *47*(6), 1192-1201.

Jacobson, J. W., Medhora, M. M., & Hartl, D. L. (1986). Molecular structure of a somatically unstable transposable element in Drosophila. *Proceedings of the National Academy of Sciences*, *83*(22), 8684-8688.

Jarvik, T., & Lark, K. G. (1998). Characterization of Soymar1, a mariner element in soybean. *Genetics*, *149*(3), 1569-1574.

Jegot, G. (2007). *Mise au point d'un système dérivé du transposon Mos1 comme vecteur non viral de transfert de gène en cellules eucaryotes* (Doctoral dissertation, Tours).

Jehle, J. A., Fritsch, E., Nickel, A., Huber, J., & Backhaus, H. (1995). TCl4. 7: a novel lepidopteran transposon found in Cydia pomonella granulosis virus. *Virology*, *207*(2), 369-379.

Jia, J., Zhao, S., Kong, X., Li, Y., Zhao, G., He, W., et al. (2013). Aegilops tauschii draft genome sequence reveals a gene repertoire for wheat adaptation. *Nature*, *496*(7443), 91-95.

Jiang, X., Tang, H., Ye, Z., & Lynch, M. (2017). Insertion polymorphisms of mobile genetic elements in sexual and asexual populations of Daphnia pulex. *Genome Biology and Evolution*, *9*(2), 362-374.

Joly-Lopez, Z., Hoen, D. R., Blanchette, M., & Bureau, T. E. (2016). Phylogenetic and genomic analyses resolve the origin of important plant genes derived from transposable elements. *Molecular biology and evolution*, msw067.

Jurka, J., & Kapitonov, V. V. (2001). PIFs meet Tourists and Harbingers: a superfamily reunion. *Proceedings of the National Academy of Sciences*, *98*(22), 12315-12316.

Jurka, J., Bao, W., Kojima, K. K., Kohany, O., & Yurka, M. G. (2012). Distinct groups of repetitive families preserved in mammals correspond to different periods of regulatory innovations in vertebrates. *Biology direct*, *7*(1), 1.

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., & Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and genome research*, *110*(1-4), 462-467.

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

Kapitonov, V. V., & Jurka, J. (2001). Rolling-circle transposons in eukaryotes. *Proceedings of the National Academy of Sciences*, *98*(15), 8714-8719.

Kapitonov, V. V., & Jurka, J. (2003). Molecular paleontology of transposable elements in the Drosophila melanogaster genome. *Proceedings of the National Academy of Sciences*, *100*(11), 6569-6574.

Kapitonov, V. V., & Jurka, J. (2006). Self-synthesizing DNA transposons in eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(12), 4540-4545.

Kapitonov, V. V., & Jurka, J. (2008). A universal classification of eukaryotic transposable elements implemented in Repbase. *Nature Reviews Genetics*, *9*(5), 411-412.

Kaul, S., Koo, H. L., Jenkins, J., Rizzo, M., Rooney, T., Tallon, L. J., et al. (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, *408*(6814), 796-815.

Keith, J. H., Schaeper, C. A., Fraser, T. S., & Fraser, M. J. (2008). Mutational analysis of highly conserved aspartate residues essential to the catalytic core of the piggyBac transposase. *BMC molecular biology*, *9*(1), 1.

Keller, B., & Feuillet, C. (2000). Colinearity and gene density in grass genomes. *Trends in plant science*, *5*(6), 246-251.

Kellogg, E. A. (2015). VI. Subfamily Pooideae Benth.(1861). In *Flowering Plants. Monocots* (pp. 199-265). Springer International Publishing.

Keravala, A., Liu, D., Lechman, E. R., Wolfe, D., Nash, J. A., Lampe, D. J., & Robbins, P. D. (2006). Hyperactive Himar1 transposase mediates transposition in cell culture and enhances gene expression in vivo. *Human gene therapy*,*17*(10), 1006-1018.

Kerry, B. R., & Crump, D. H. (1998). The dynamics of the decline of the cereal cyst nematode, Heterodera avenae, in four soils under intensive cereal production. *Fundamental and Applied Nematology*, *21*(5), 617-625.

Kharrat, I., Mezghani, M., Casse, N., Denis, F., Caruso, A., Makni, H., et al. (2015). Characterization of mariner-like transposons of the mauritiana Subfamily in seven tree aphid species. *Genetica*, *143*(1), 63-72.

Kidwell, M. G. (2002). Transposable elements and the evolution of genome size in eukaryotes. *Genetica*, *115*(1), 49-63.

Kidwell, M. G., & Lisch, D. (1997). Transposable elements as sources of variation in animals and plants. *Proceedings of the National Academy of Sciences*, *94*(15), 7704-7711.

Kidwell, M. G., & Lisch, D. R. (2001). Perspective: transposable elements, parasitic DNA, and genome evolution. *Evolution*, *55*(1), 1-24.Klassen, W., & Curtis, C. F. (2005). History of the sterile insect technique. In Sterile Insect Technique (pp. 3-36). Springer Netherlands.

Kidwell, M. G., Kidwell, J. F., & Sved, J. A. (1977). Hybrid dysgenesis in Drosophila melanogaster: a syndrome of aberrant traits including mutation, sterility and male recombination. *Genetics*, *86*(4), 813-833.Kidwell, M. G. (1992). Horizontal transfer. *Current opinion in genetics & development*, *2*(6), 868-873.

Klinakis, A. G., Zagoraiou, L., Vassilatis, D. K., & Savakis, C. (2000). Genome-wide insertional mutagenesis in human cells by the Drosophila mobile element minos. *EMBO reports*, *1*(5), 416-421.

Koukidou, M., Klinakis, A., Reboulakis, C., Zagoraiou, L., Tavernarakis, N., Livadaras, I., et al. Savakis, C. (2006). Germ line transformation of the olive fly Bactrocera oleae using a versatile transgenesis marker. *Insect Molecular Biology*, *15*(1), 95-103.

Kramerov, D. A., & Vassetzky, N. S. (2005). Short retroposons in eukaryotic genomes. *International review of cytology*, *247*, 165-221.

Kunze, R., & Weil, C. F. (2002). The hAT and CACTA superfamilies of plant transposons. In *Mobile DNA II* (pp. 565-610). American Society of Microbiology.

Kurokawa, T., Uji, S., & Suzuki, T. (2005). Identification of cDNA coding for a homologue to mammalian leptin from pufferfish, Takifugu rubripes. *Peptides*, *26*(5), 745-750.

Lacroix M. (2002). Maladie des céréales et de la luzerne. Diagnostic, dépistage, prévention. Ministére de l'argiculture des pêcheries et de l'alimentation.Québec.

Lampe, D. J., Churchill, M. E., & Robertson, H. M. (1996). A purified mariner transposase is sufficient to mediate transposition in vitro. *The EMBO journal*, *15*(19), 5470.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860-921.

Lapierre, H., & Signoret, P. A. (2004). *Viruses and virus diseases of Poaceae (Gramineae)*. Editions Quae.

Laven, H. (1967). Eradication of Culex pipiens fatigans through cyioplasmic incompatibility. *Nature, London*, *216*(5113), 383-384.

Le Rouzic, A., & Capy, P. (2005). The first steps of transposable elements invasion: parasitic strategy vs. genetic drift. *Genetics*, *169*(2), 1033-1043.

Le Rouzic, A., Boutin, T. S., & Capy, P. (2007). Long-term evolution of transposable elements. *Proceedings of the National Academy of Sciences*, *104*(49), 19375-19380.

Leroy, H., Castagnone-Sereno, P., Renault, S., Augé-Gouillou, C., Bigot, Y., & Abad, P. (2003). Characterization of Mcmar1, a mariner-like element with large inverted terminal repeats (ITRs) from the phytoparasitic nematode Meloidogyne chitwoodi. *Gene*, *304*, 35-41.

Lidholm, D. A., Gudmundsson, G. H., & Boman, H. G. (1991). A highly repetitive, mariner-like element in the genome of Hyalophora cecropia. *Journal of Biological Chemistry*, *266*(18), 11518-11521.

Lin, N. S., & Langenberg, W. G. (1984). Distribution of Barley stripe mosaic virus protein in infected wheat root and shoot tips. *Journal of general virology*, *65*(12), 2217-2224.

Lisch, D. (2002). Mutator transposons. *Trends in plant science*, *7*(11), 498-504.

Lobo, N., Li, X., & Fraser Jr, M. J. (1999). Transposition of the piggyBac element in embryos of Drosophila melanogaster, Aedes aegypti and Trichoplusia ni. *Molecular and General Genetics MGG*, *261*(4-5), 803-810.

Lohe, A. R., & Hartl, D. L. (1996). Autoregulation of mariner transposase activity by overproduction and dominant-negative complementation. *Molecular biology and evolution*, *13*(4), 549-555.

Lohe, A. R., De Aguiar, D., & Hartl, D. L. (1997). Mutations in the mariner transposase: the D, D (35) E consensus sequence is nonfunctional. *Proceedings of the National Academy of Sciences*, *94*(4), 1293-1297.

Lohe, A. R., Moriyama, E. N., Lidholm, D. A., & Hartl, D. L. (1995). Horizontal transmission, vertical inactivation, and stochastic loss of mariner-like transposable elements. *Molecular biology and evolution*, *12*(1), 62-72.

Loreto, E. L. S., Carareto, C. M. A., & Capy, P. (2008). Revisiting horizontal transfer of transposable elements in Drosophila. *Heredity*, *100*(6), 545-554.

Love, A. (1982). Generic evolution of the wheatgrasses. Biol. Zentralbl. 101: 199-212. 1984. Conspectus of the Triticeae. *Feddes Repert*, *95*, 425-521.

Luo, G. H., Li, X. H., Han, Z. J., Guo, H. F., Yang, Q., Wu, M., et al. (2014). Molecular characterization of the piggyBac-like element, a candidate marker for phylogenetic research of Chilo suppressalis (Walker) in China. *BMC molecular biology*, *15*(1), 1.

Luo, G. H., Wu, M., Wang, X. F., Zhang, W., & Han, Z. J. (2011). A new active piggyBac-like element in Aphis gossypii. *Insect Science*, *18*(6), 652-662.

Macas, J., Koblí ková, A., & Neumann, P. (2005). Characterization of Stowaway MITEs in pea (Pisum sativum L.) and identification of their potential master elements. *Genome*, *48*(5), 831-839.

Maragathavally, K. J., Kaminski, J. M., & Coates, C. J. (2006). Chimeric Mos1 and piggyBac transposases result in site-directed integration. *The FASEB journal*, *20*(11), 1880-1882.

Martienssen, R., & Baron, A. (1994). Coordinate suppression of mutations caused by Robertson's mutator transposons in maize. *Genetics*, *136*(3), 1157-1170.

Maruyama, K., & Hartl, D. L. (1991). Evidence for interspecific transfer of the transposable element mariner between Drosophila and Zaprionus. *Journal of molecular evolution*, *33*(6), 514-524.

Matsutani, S., Ohtsubo, H., Maeda, Y., & Ohtsubo, E. (1987). Isolation and characterization of IS elements repeated in the bacterial chromosome. *Journal of molecular biology*, *196*(3), 445-455.

Mazaheri, M., Kianian, P., Mergoum, M., Valentini, G. L., Seetan, R., Pirseyedi, S. M., et al. (2014). Transposable element junctions in marker development and genomic characterization of barley. *The Plant Genome*, *7*(1). doi:10.3835/plantgenome2013.10.0036

McClintock, B. (1950). The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences*, *36*(6), 344-355.

McClintock, B. (1951). Chromosome organization and genic expression. In *Cold Spring Harbor Symposia on Quantitative Biology* (16, pp. 13-47). Cold Spring Harbor Laboratory Press.

McDonald, J. F. (1995). Transposable elements: possible catalysts of organismic evolution. *Trends in ecology & evolution*, *10*(3), 123-126.

McVey, M., & Lee, S. E. (2008). MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. *Trends in Genetics*, *24*(11), 529-538.

Medhora, M. M., MacPeek, A. H., & Hartl, D. L. (1988). Excision of the Drosophila transposable element mariner: identification and characterization of the Mos factor. *The EMBO journal*, *7*(7), 2185.

Medhora, M., Maruyama, K., & Hartl, D. L. (1991). Molecular and functional analysis of the mariner mutator element Mos1 in Drosophila. *Genetics*, *128*(2), 311-318.

Mezghani-Khemakhem, M., Bouktila, D., Kharrat, I., Makni, M., & Makni, H. (2012). Genetic variability of green citrus aphid populations from Tunisia, assessed by RAPD markers and mitochondrial DNA sequences. *Entomological science*, *15*(2), 171-179.

Miller RH, Pike KS (2002) Insects in wheat-based systems. In: Curtis BC, Rajaram S, Go´mez Macpherson H (eds). *Bread wheat: improvement and production, plant production and protection series* no. 30, FAO, Rome, pp. 367–393.

Miller, W. J., McDonald, J. F., Nouaud, D., & Anxolabéhère, D. (2000). Molecular domestication—more than a sporadic episode in evolution. In *Transposable Elements and Genome Evolution* (pp. 197-207). Springer Netherlands.

Miskey, C., Izsvák, Z., Kawakami, K., & Ivics, Z. (2005). DNA transposons in vertebrate functional genomics. *Cellular and molecular life sciences*, *62*(6), 629-641.

Miskey, C., Izsvák, Z., Plasterk, R. H., & Ivics, Z. (2003). The Frog Prince: a reconstructed transposon from Rana pipiens with high transpositional activity in vertebrate cells. *Nucleic acids research*, *31*(23), 6873-6881.

Miskey, C., Papp, B., Mátés, L., Sinzelle, L., Keller, H., Izsvák, Z., & Ivics, Z. (2007). The ancient mariner sails again: transposition of the human Hsmar1 element by a reconstructed transposase and activities of the SETMAR protein on transposon ends. *Molecular and cellular biology*, *27*(12), 4589-4600.

Mitra, R., Fain-Thornton, J., & Craig, N. L. (2008). piggyBac can bypass DNA synthesis during cut and paste transposition. *The EMBO journal*, *27*(7), 1097-1109.

Mitra, R., Li, X., Kapusta, A., Mayhew, D., Mitra, R. D., Feschotte, C., & Craig, N. L. (2013). Functional characterization of piggyBat from the bat Myotis lucifugus unveils an active mammalian DNA transposon. *Proceedings of the National Academy of Sciences*, *110*(1), 234-239.

Mittapalli, O., Rivera-Vega, L., Bhandary, B., Bautista, M. A., Mamidala, P., Michel, A. P., ... & Mian, M. A. R. (2011). Cloning and characterization of mariner-like elements in the soybean aphid, Aphis glycines Matsumura. *Bulletin of entomological research*, *101*(06), 697-704.

Modolo L (2014). Analyse bioinformatique des évènements de transferts horizontaux entre espèces de drosophiles et lien avec la régulation des éléments transposables. Biologie moléculaire. Université Claude Bernard - Lyon I, 2014. Français. <NNT : 2014LYO10258>. <tel-01167124>.

Moore, G. W., Barnabas, J., & Goodman, M. (1973). A method for constructing maximum parsimony ancestral amino acid sequences on a given network. *Journal of Theoretical Biology*, *38*(3), 459-485.

Moore, G., Devos, K. M., Wang, Z., & Gale, M. D. (1995). Cereal genome evolution: grasses, line up and form a circle. *Current biology*, *5*(7), 737-739.

Moran, J. V., DeBerardinis, R. J., & Kazazian, H. H. (1999). Exon shuffling by L1 retrotransposition. *Science*, *283*(5407), 1530-1534.

Mullis, K. B. (1990). The unusual origin of the polymerase chain reaction. *Scientific American*, *262*(4), 56-61.

Muñoz-López, M., Siddique, A., Bischerour, J., Lorite, P., Chalmers, R., & Palomeque, T. (2008). Transposition of Mboumar-9: identification of a new naturally active mariner-family transposon. *Journal of molecular biology*, *382*(3), 567-572.

Narasimhan, G., Bu, C., Gao, Y., Wang, X., Xu, N., & Mathee, K. (2002). Mining protein sequences for motifs. *Journal of Computational Biology*, *9*(5), 707-720.

Negoua, A., Rouault, J. D., Chakir, M., & Capy, P. (2013). Internal deletions of transposable elements: the case of Lemi elements. *Genetica*, *141*(7-9), 369-379.

Newman, J. C., Bailey, A. D., Fan, H. Y., Pavelitz, T., & Weiner, A. M. (2008). An abundant evolutionarily conserved CSB-PiggyBac fusion protein expressed in Cockayne syndrome. *PLoS Genet*, *4*(3), e1000031.

Nicholas, K. B., Nicholas, H. B. J., & Deerfield, D. W. (1997). GeneDoc: analysis and visualization of genetic variation. *EMBNEW.NEWS*, *4*(1).

Nicholson, S. J., Nickerson, M. L., Dean, M., Song, Y., Hoyt, P. R., Rhee, H., et al. (2015). The genome of Diuraphis noxia, a global aphid pest of small grains. *BMC genomics*, *16*(1), 1.

Nicol, J. M., & Rivoal, R. (2008). Global knowledge and its application for the integrated control and management of nematodes on wheat. In *Integrated management and biocontrol of vegetable and grain crops nematodes* (pp. 251-294). Springer Netherlands.

Ochman, H., Gerber, A. S., & Hartl, D. L. (1988). Genetic applications of an inverse polymerase chain reaction. *Genetics*, *120*(3), 621-623.

Okada, N., Hamada, M., Ogiwara, I., & Ohshima, K. (1997). SINEs and LINEs share common 3′ sequences: a review. *Gene*, *205*(1), 229-243.

Orgel, L. E., & Crick, F. H. (1980). Selfish DNA: the ultimate parasite. *Nature*, *284*(5757), 604.

Oswald, J. W., & Houston, B. R. (1951). A new virus disease of cereals, transmissible by aphids. *Plant Disease Reporter*, *11*, 471-475.

Pace, J. K., Gilbert, C., Clark, M. S., & Feschotte, C. (2008). Repeated horizontal transfer of a DNA transposon in mammals and other tetrapods.*Proceedings of the National Academy of Sciences*, *105*(44), 17023-17028.

Palomeque, T., Carrillo, J. A., Muñoz-López, M., & Lorite, P. (2006). Detection of a mariner-like element and a miniature inverted-repeat transposable element (MITE) associated with the heterochromatin from ants of the genus Messor and their possible involvement for satellite DNA evolution. *Gene*, *371*(2), 194-205.

Pasichnyk, L. A. (1999). Properties of bacteria of pathovars of Pseudomonas syringae affecting cereals. *Mikrobiolohichnyi zhurnal (Kiev, Ukraine: 1993)*, *62*(5), 18-22.

Pavelitz, T., Gray, L. T., Padilla, S. L., Bailey, A. D., & Weiner, A. M. (2013). PGBD5: a neural-specific intron-containing piggyBac transposase domesticated over 500 million years ago and conserved from cephalochordates to humans. *Mobile DNA*, *4*(1), 1.

Penton, E. H., Sullender, B. W., & Crease, T. J. (2002). Pokey, a new DNA transposon in Daphnia (Cladocera: Crustacea). *Journal of molecular evolution*, *55*(6), 664-673.

Petit, A., Rouleux-Bonnin, F., Lambelé, M., Pollet, N., & Bigot, Y. (2007). Properties of the various Botmar1 transcripts in imagoes of the bumble bee, Bombus terrestris (Hymenoptera: Apidae). *Gene*, *390*(1), 52-66.

Piégu, B., Guizard, S., Spears, T., Cruaud, C., Couloux, A., Bideshi, D. K., et al. (2013). Complete genome sequence of invertebrate iridescent virus 22 isolated from a blackfly larva. *Journal of General Virology*, *94*(9), 2112-2116.

Piriyaponga, J., Marino-Ramirez, L., and Jordan, I.K., Origin and Evolution of Human microRNAs from Transposable Elements, *Genetics*, 2007, vol. 176, pp. 1323–1337.

Plasterk, R. H., Izsvák, Z., & Ivics, Z. (1999). Resident aliens: the Tc1/mariner superfamily of transposable elements. *Trends in Genetics*, *15*(8), 326-332.

Poulter, R. T. M., & Goodwin, T. J. D. (2005). DIRS-1 and the other tyrosine recombinase retrotransposons. *Cytogenetic and genome research*, *110*(1-4), 575-588.

Prasad, M. D., Nurminsky, D. L., & Nagaraju, J. (2002). Characterization and molecular phylogenetic analysis of mariner elements from wild and domesticated species of silkmoths. *Molecular Phylogenetics and Evolution*, *25*(1), 210-217.

Prasad, V., Strömberg, C. A., Alimohammadian, H., & Sahni, A. (2005). Dinosaur coprolites and the early evolution of grasses and grazers. *Science*, *310*(5751), 1177-1180.

Pritham, E. J. (2009). Transposable elements and factors influencing their success in eukaryotes. *Journal of Heredity*, *100*(5), 648-655.

Puchta, H. (2005). The repair of double-strand breaks in plants: mechanisms and consequences for genome evolution. *Journal of experimental botany*, *56*(409), 1-14.

Raboudi, F., Chavigny, P., Makni, H., Vanlerberghe, F. M., & Makni, M. (2012). Spatial and Temporal Genetic Variation in Tunisian Field Populations of Macrosiphum euphorbiae (Thomas). *Environmental entomology*, *41*(2), 420-425.

Rad, R., Rad, L., Wang, W., Cadinanos, J., Vassiliou, G., Rice, S., et al. (2010). PiggyBac transposon mutagenesis: a tool for cancer gene discovery in mice. *Science*, *330*(6007), 1104-1107.

Rholl, D. A., Trunck, L. A., & Schweizer, H. P. (2008). In vivo Himar1 transposon mutagenesis of Burkholderia pseudomallei. *Applied and environmental microbiology*, *74*(24), 7529-7535.

Richards, S., Gibbs, R. A., Gerardo, N. M., Moran, N., Nakabachi, A., Stern, D., et al. (2010). Genome sequence of the pea aphid Acyrthosiphon pisum. *PLos Biology*, *8*(2).

Richardson, J. M., Dawson, A., O'hagan, N., Taylor, P., Finnegan, D. J., & Walkinshaw, M. D. (2006). Mechanism of Mos1 transposition: insights from structural analysis. *The EMBO journal*, *25*(6), 1324-1334.

Robertson, H. M. (1993). The mariner transposable element is widespread in insects. *Nature*, *362* (6417), 241-245.

Robertson, H. M. (1997). Multiple mariner transposons in flatworms and hydras are related to those of insects. *Journal of Heredity*, *88*(3), 195-201.

Robertson, H. M. (2002). Evolution of DNA transposons in eukaryotes. In *Mobile DNA ii* (pp. 1093-1110). American Society of Microbiology.

Robertson, H. M., & Lampe, D. J. (1995). Recent horizontal transfer of a mariner transposable element among and between Diptera and Neuroptera. *Molecular Biology and Evolution*, *12*(5), 850-862.

Robertson, H. M., & MacLeod, E. G. (1993). Five major subfamilies of mariner transposable elements in insects, including the Mediterranean fruit fly, and related arthropods. *Insect molecular biology*, *2*(3), 125-139.

Robinson, A. S., Franz, G., & Atkinson, P. W. (2004). Insect transgenesis and its potential role in agriculture and human health. *Insect biochemistry and molecular biology*, *34*(2), 113-120.

Rodić, N., & Burns, K. H. (2013). Long interspersed element–1 (LINE-1): passenger or driver in human neoplasms?. *PLoS Genet*, *9*(3), e1003402.

Rogers, D. W., Baldini, F., Battaglia, F., Panico, M., Dell, A., Morris, H. R., & Catteruccia, F. (2009). Transglutaminase-mediated semen coagulation controls sperm storage in the malaria mosquito. *PLoS Biology*, *7*(12), e1000272.

Rouault, J. D., Casse, N., Chénais, B., Hua-Van, A., Filée, J., & Capy, P. (2009). Automatic classification within families of transposable elements: application to the mariner Family. *Gene*, *448*(2), 227-232.

Routh, A., Domitrovic, T., & Johnson, J. E. (2012). Host RNAs, including transposons, are encapsidated by a eukaryotic single-stranded RNA virus. *Proceedings of the National Academy of Sciences*, *109*(6), 1907-1912.

Roy-Engel, A. M., Carroll, M. L., Vogel, E., Garber, R. K., Nguyen, S. V., Salem, A. H., et al. (2001). Alu insertion polymorphisms for the study of human genomic diversity. *Genetics*, *159*(1), 279-290.

Rozhkov, N. V., Aravin, A. A., Zelentsova, E. S., Schostak, N. G., Sachidanandam, R., McCombie, W. R., et al. (2010). Small RNA-based silencing strategies for transposons in the process of invading Drosophila species. *Rna*, *16*(8), 1634-1645.

Rubin, E., & Levy, A. A. (1997). Abortive gap repair: underlying mechanism for Ds element formation. *Molecular and Cellular Biology*, *17*(11), 6294-6302.

Rubin, G. M., & Spradling, A. C. (1982). Genetic transformation of Drosophila with transposable element vectors. *Science*, *218*(4570), 348-353.

Ryan J.D, Morgham A.T, Richardson P.E, Johnson R.C, Mort A.J & Eikenbary R.D. 1990 : Greenbugs and wheat: a model system for the study of phytotoxic Homoptera. In : Campbell R.K. and Eikenbary R.D, eds. Aphid-plant genotype interactions. Elseiver, Amsterdam, 171-186.

Ryder, E., & Russell, S. (2003). Transposable elements as tools for genomics and genetics in Drosophila. *Briefings in functional genomics & proteomics*, *2*(1), 57-71.

Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, *4*(4), 406-425.

Sánchez-Gracia, A., Maside, X., & Charlesworth, B. (2005). High rate of horizontal transfer of transposable elements in Drosophila. *Trends in Genetics*, *21*(4), 200-203.

SanMiguel, P., & Bennetzen, J. L. (1998). Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. *Annals of Botany*, *82*(suppl 1), 37-44.

SanMiguel, P., Tikhonov, A., Jin, Y. K., & Motchoulskaia, N. (1996). Nested retrotransposons in the intergenic regions of the maize genome. *Science*, *274*(5288), 765.

Sarakatsanou, A., Diamantidis, A. D., Papanastasiou, S. A., Bourtzis, K., & Papadopoulos, N. T. (2011). Effects of Wolbachia on fitness of the Mediterranean fruit fly (Diptera: Tephritidae). *Journal of Applied Entomology*, *135* (7), 554-563.

Sarkar, A., Sim, C., Hong, Y. S., Hogan, J. R., Fraser, M. J., Robertson, H. M., & Collins, F. H. (2003). Molecular evolutionary analysis of the widespread piggyBac transposon family and related" domesticated" sequences. *Molecular Genetics and Genomics*, *270*(2), 173-180.

Sassetti, C. M., Boyd, D. H., & Rubin, E. J. (2003). Genes required for mycobacterial growth defined by high density mutagenesis. *Molecular microbiology*, *48*(1), 77-84.

Schlenke, T. A., & Begun, D. J. (2004). Strong selective sweep associated with a transposon insertion in Drosophila simulans. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(6), 1626-1631.

Schmidt, A. L., & Anderson, L. M. (2006). Repetitive DNA elements as mediators of genomic change in response to environmental cues. *Biological Reviews*, *81*(4), 531-543.

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. *Science*, *326*(5956), 1112-1115.

Shao, H., & Tu, Z. (2001). Expanding the diversity of the IS630-Tc1-mariner superfamily: discovery of a unique DD37E transposon and reclassification of the DD37D and DD39D transposons. *Genetics*, *159*(3), 1103-1115.

Shapiro, J. A. (1969). Mutations caused by the insertion of genetic material into the galactose operon of Escherichia coli. *Journal of molecular biology*, *40*(1), 93-105.

Sijen, T., & Plasterk, R. H. (2003). Transposon silencing in the Caenorhabditis elegans germ line by natural RNAi. *Nature*, *426*(6964), 310-314.

Sinzelle, L., Izsvak, Z., & Ivics, Z. (2009). Molecular domestication of transposable elements: from detrimental parasites to useful host genes. *Cellular and molecular life sciences*, *66*(6), 1073-1093.

Siomi, M. C., Sato, K., Pezic, D., & Aravin, A. A. (2011). PIWI-interacting small RNAs: the vanguard of genome defence. *Nature reviews Molecular cell biology*, *12*(4), 246-258.

Slama, A., Salem, M. B., & Zid, E. (2005). Les céréales en Tunisie: production, effet de la sécheresse et mécanismes de résistance. *Science et changements planétaires/Sécheresse*, *16*(3), 225-229.

Slotkin, R. K., & Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. *Nature Reviews Genetics*, *8*(4), 272-285.

Sneath, P. H., & Sokal, R. R. (1973). *Numerical taxonomy. The principles and practice of numerical classification. H. Freeman and Co., San Francisco*.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., et al. (1998). Comprehensive identification of cell cycle–regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Molecular biology of the cell*, *9*(12), 3273-3297.

Spradling, A. C., & Rubin, G. M. (1982). Transposition of cloned P elements into Drosophila germ line chromosomes. *Science*, *218*(4570), 341-347.

Spradling, A. C., Stern, D. M., Kiss, I., Roote, J., Laverty, T., & Rubin, G. M. (1995). Gene disruptions using P transposable elements: an integral component of the Drosophila genome project. *Proceedings of the National Academy of Sciences*, *92*(24), 10824-10830.

St John, J. A., Braun, E. L., Isberg, S. R., Miles, L. G., Chong, A. Y., Gongora, J., et al. (2012). Sequencing three crocodilian genomes to illuminate the evolution of archosaurs and amniotes. *Genome biology*, *13*(1), 1.

Sun, Z. C., Wu, M., Miller, T. A., & Han, Z. J. (2008). piggyBac-like elements in cotton bollworm, Helicoverpa armigera (Hübner). *Insect molecular biology*, *17*(1), 9-18.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Molecular biology and evolution*, *30*(12), 2725-2729.

Tang, Z., Zhang, H. H., Huang, K., Zhang, X. G., Han, M. J., & Zhang, Z. (2015). Repeated horizontal transfers of four DNA transposons in invertebrates and bats. *Mobile DNA*, *6*(1), 1.

Tarchini, R., Biddle, P., Wineland, R., Tingey, S., & Rafalski, A. (2000). The complete sequence of 340 kb of DNA around the rice Adh1–Adh2 region reveals interrupted colinearity with maize chromosome 4. *The Plant Cell*, *12*(3), 381-391.

Tellier, M., Bouuaert, C. C., & Chalmers, R. (2015). Mariner and the ITm superfamily of transposons. *Microbiology spectrum*, *3*(2).

Tenaillon, M. I., Hollister, J. D., & Gaut, B. S. (2010). A triptych of the evolution of plant transposable elements. *Trends in plant science*, *15*(8), 471-478.

Tóth, K. F., Pezic, D., Stuwe, E., & Webster, A. (2016). The piRNA pathway guards the germline genome against transposable elements. In *Non-coding RNA and the Reproductive System* (pp. 51-77). Springer Netherlands.

Traugott, M., Bell, J. R., Broad, G. R., Powell, W., Van Veen, F. J. F., Vollhardt, I. M. G., & Symondson, W. O. C. (2008). Endoparasitism in cereal aphids: molecular analysis of a whole parasitoid community. *Molecular Ecology*, *17*(17), 3928-3938.

Ünsal, K., & Morgan, G. T. (1995). A Novel Group of Families of Short Interspersed Repetitive Elements (SINEs) in Xenopus: Evidence of a Specific Target Site for DNA-mediated Transposition of Inverted-repeat SINEs. *Journal of molecular biology*, *248*(4), 812-823.

Van Emden, H. F., & Harrington, R. (Eds.). (2007). *Aphids as crop pest*s. Cabi.

Vanlerberghe-Masutti, F., & Chavigny, P. (1998). Host-based genetic differentiation in the aphid Aphis gossypii Glover, evidenced from RAPD fingerprints. *Molecular ecology*, *7*(7), 905-914.

Vicient, C. M., Suoniemi, A., Anamthawat-Jónsson, K., Tanskanen, J., Beharav, A., Nevo, E., & Schulman, A. H. (1999). Retrotransposon BARE-1 and its role in genome evolution in the genus Hordeum. *The Plant Cell*, *11*(9), 1769-1784.

Vieira, C., Nardon, C., Arpin, C., Lepetit, D., & Biémont, C. (2002). Evolution of genome size in Drosophila. Is the invader's genome being invaded by transposable elements?. *Molecular biology and evolution*, *19*(7), 1154-1161.

Vogel, J. P., Garvin, D. F., Mockler, T. C., Schmutz, J., Rokhsar, D., Bevan, M. W., et al. (2010). Genome sequencing and analysis of the model grass Brachypodium distachyon. *Nature*, *463*(7282), 763-768.

Vogt, A., & Mochizuki, K. (2013). A domesticated PiggyBac transposase interacts with heterochromatin and catalyzes reproducible DNA elimination in Tetrahymena. *PLoS Genet*, *9*(12), e1004032.

Voigt, F., Wiedemann, L., Zuliani, C., Querques, I., Sebe, A., Mátés, L., et al. (2016). Sleeping Beauty transposase structure allows rational design of hyperactive variants for genetic engineering. *Nature communications*, *7*.

Volff, J. N. (2006). Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. *Bioessays*, *28*(9), 913-922.

Volkl, W., Mackauer, M., Pell, J.K. et Brodeur, J. (2007). Predators, parasitoids and pathogens. Dans van Emden, H. et R. Harrington (dir.), *Aphids as crop pest* (pp. 187-234). Harpenden, UK: CABI.

Vollhardt, I. M., Tscharntke, T., Wäckers, F. L., Bianchi, F. J., & Thies, C. (2008). Diversity of cereal aphid parasitoids in simple and complex landscapes. *Agriculture, ecosystems & environment*, *126*(3), 289-292.

Wallau, G. L., Capy, P., Loreto, E., & Hua-Van, A. (2014). Genomic landscape and evolutionary dynamics of mariner transposable elements within the Drosophila genus. *BMC genomics*, *15*(1), 1.

Wallau, G. L., Ortiz, M. F., & Loreto, E. L. S. (2012). Horizontal transposon transfer in eukarya: detection, bias, and perspectives. *Genome biology and evolution*, *4*(8), 689-699.

Wang, J., Du, Y., Wang, S., Brown, S. J., & Park, Y. (2008). Large diversity of the piggyBac-like elements in the genome of Tribolium castaneum. *Insect biochemistry and molecular biology*, *38*(4), 490-498.

Wang, J., Miller, E. D., Simmons, G. S., Miller, T. A., Tabashnik, B. E., & Park, Y. (2010). PiggyBac-like elements in the pink bollworm, Pectinophora gossypiella. *Insect molecular biology*, *19*(2), 177-184.

Wang, J., Ren, X., Miller, T. A., & Park, Y. (2006). piggyBac-like elements in the tobacco budworm, Heliothis virescens (Fabricius). *Insect molecular biology*, *15*(4), 435-443.

Wang, W., Swevers, L., & Iatrou, K. (2000). Mariner (Mos1) transposase and genomic integration of foreign gene sequences in Bombyx mori cells. *Insect molecular biology*, *9*(2), 145-155.

Warmus H, Brown P (1989) Retroviruses. In Berg DE, Howe M (eds) *Mobile DNA*. American Society for microbiology, Washington DC, USA, pp. 53-108.

Warren, W. C., Hillier, L. W., Graves, J. A. M., Birney, E., Ponting, C. P., Grützner, F., et al. (2008). Genome analysis of the platypus reveals unique signatures of evolution. *Nature*, *453*(7192), 175-183.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., et al. (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, *8*(12), 973-982.

Williams, I. S., & Dixon, A. F. (2007). 3 Life Cycles and Polymorphism. *Aphids as crop pests*, 69-86.

Witherspoon, D. J., & Robertson, H. M. (2003). Neutral evolution of ten types of mariner transposons in the genomes of Caenorhabditis elegans and Caenorhabditis briggsae. *Journal of molecular evolution*, *56*(6), 751-769.

Wright, S. I., & Schoen, D. J. (2000). Transposon dynamics and the breeding system. In *Transposable Elements and Genome Evolution* (pp. 139-148). Springer Netherlands.

Wu, C., & Wang, S. (2014). PLE-wu, a new member of piggyBac transposon family from insect, is active in mammalian cells. *Journal of bioscience and bioengineering*, *118*(4), 359-366.

Wu, M., Sun, Z. C., Hu, C. L., Zhang, G. F., & Han, Z. J. (2008). An active piggyBac-like element in Macdunnoughia crassisigna. *Insect Science*, *15*(6), 521-528.

Wu, M., Sun, Z., Luo, G., Hu, C., Zhang, W., & Han, Z. (2011). Cloning and characterization of piggyBac-like elements in lepidopteran insects. *Genetica*, *139*(1), 149-154.

Wu, S. C. Y., Meir, Y. J. J., Coates, C. J., Handler, A. M., Pelczar, P., Moisyadi, S., & Kaminski, J. M. (2006). piggyBac is a flexible and highly active transposon as compared to sleeping beauty, Tol2, and Mos1 in mammalian cells. *Proceedings of the National Academy of Sciences*, *103*(41), 15008-15013.

Xu, H. F., Xia, Q. Y., Liu, C., Cheng, T. C., Zhao, P., Duan, J., et al. (2006). Identification and characterization of piggyBac-like elements in the genome of domesticated silkworm, Bombyx mori. *Molecular Genetics and Genomics*, *276*(1), 31-40.

Yeadon, P. J., & Catcheside, D. E. (1995). Guest: a 98 by inverted repeat transposable element in Neurospora crassa. *Molecular and General Genetics MGG*, *247*(1), 105-109.

Yusa, K. (2015). piggyBac Transposon. *Microbiology spectrum*, *3*(2). DOI:10.1128/microbiolspec. MDNA3-0028-2014

Zagoraiou, L., Drabek, D., Alexaki, S., Guy, J. A., Klinakis, A. G., Langeveld, A., et al. (2001). In vivo transposition of Minos, a Drosophila mobile element, in mammalian tissues. *Proceedings of the National Academy of Sciences*, *98*(20), 11474-11478.

Zerbib, D., Polard, P., Escoubas, J. M., Galas, D., & Chandler, M. (1990). The regulatory role of the IS 1-encoded InsA protein in transposition. *Molecular microbiology*, *4*(3), 471-477.

Zhang, D., Zheng, X., Xi, Z., Bourtzis, K., & Gilles, J. R. (2015). Combining the sterile insect technique with the incompatible insect technique: I-impact of Wolbachia infection on the fitness of triple-and double-infected strains of Aedes albopictus. *PloS one*, *10*(4), e0121126.

Zhou, M. B., Zhong, H., & Tang, D. Q. (2011). Isolation and characterization of seventy-nine full-length mariner-like transposase genes in the Bambusoideae subfamily. *Journal of plant research*, *124*(5), 607-617.

Title: Plasticity of the genomes of cereal aphids and their host plant: *in silico* and *in vitro* analyses of *Tc1-mariner-IS630* and *piggyBac* superfamilies of transposable elements

Keywords: Cereal aphids, host plants (cereals), *mariner*, *piggyBac*, Transposable elements.

Abstract:

Cereal farming plays an important role in world agriculture and contributes to the food security of the populations. To improve the production of cereals (barley, wheat, oats ...), it is necessary to fight against their pests, especially aphids, able to transmit several viruses. The analysis of aphid's genomes such as *Rhopalosiphum padi*, *R. maidis*, *Sitobion avenae* and *Schizaphis graminum*, their evolution and their relationships with their host plants could contribute to define strategies against pest populations. In this context, this work focused on the analysis of transposable elements belonging to *Tc1-mariner-IS630* and *piggyBac* superfamilies. Indeed, TEs are involved in genomic plasticity and evolution of species, and are also used in biotechnology to develop gene transfer tools.

In the first chapter, we investigate three available genomes of aphids, namely *Acyrthosiphon pisum*, *Myzus persicae* and *Diuraphis noxia*, to search for elements of the *mariner* family or close to it. Based on sequence similarities, we were able to characterize 183 elements distributed in three clades. The first one, common to the three species, corresponds to the clade of *irritans* subfamily DD34D, and is subdivided into three tribes *Macrosiphinimar*, *Batmar-like* elements and *Dnomar-like* elements. The second one includes the *rosa* element DD41D belonging to a group close to the *mariner* family. The third one includes sequences with long Terminal Inverted Repeats and is subdivided into two DD40-41D tribes. These two latter clades, more common in *A. pisum*, likely derive from a common ancestor and would form a new family.

In the second chapter, the results of *in silico* research were exploited, to identify *in vitro*, elements of the *irritans* subfamily in cereal aphids and in their host plants as well. Two types of deleted elements (MITEs) have been identified in aphids, one common to all species with a percentage of identity higher than 98% (*Aphidmar*) and the other one specific to *S. avenae* (*Samar2*). In addition, the genomes of cereals (barley, wheat, brachypodium, aegilops) were investigated using as queries sequences of *irritans* subfamily found in aphids. A single contig identified in *Hordeum vulgare* (cultivar *barke*) contains a 320 bp truncated element flanked by genomic DNA of aphids. The presence of this sequence was checked in several barley cultivars by an *in vitro* approach. Two types of sequences were found. The first one similar to that found in barley from the *in silico* approach, the second one corresponding to *Samar2* element, lacking seven nucleotides at the breaking points of the initial deletion. This suggests a possible horizontal transfer between cereal aphids and barley.

In the last chapter, the abundance of genomic data and the scarcity of in-depth research covering all members of the *piggyBac* superfamily led us to determine *in silico* their characteristic, their distribution and their evolution. A total of 117 proteic sequences of the PBLE (autonomous elements) and PGBD (domesticated elements) have been used as queries. Four structural groups of PBLE have been identified depending on the presence or absence of sub-terminal repeats (direct / inverted). However, there is no relationship between the structural groups and the phylogeny of these PBLE elements. PGBD are clearly structured into nine main groups including a new group of domesticated elements found in Neopterygii. The catalytic domain of the ancestral transposase is not always preserved, but all these domesticated elements are subjected to a strong purifying selection. The general phylogeny of PBLEs and PGBD suggests multiple and independent domestication events of PGBD from different PBLE ancestors.

Titre : Plasticité des génomes des pucerons des céréales et de leur plante hôte : recherche *in silico* et *in vitro* des éléments transposables des superfamilles *Tc1-mariner-IS630* et *piggyBac*

Mots clés : Pucerons des céréales, plantes hôtes (céréales), *mariner*, *piggyBac*, Éléments transposables.

Résumé :

La céréaliculture occupe une place importante dans l'agriculture mondiale et contribue à la sécurité alimentaire des populations. Pour assurer la production des céréales (orge, blé, avoine…), il est nécessaire de lutter contre ses ravageurs, essentiellement les pucerons qui sont capables de transmettre plusieurs virus. L'analyse des génomes des pucerons tels que *Rhopalosiphum padi*, *R. maidis*, *Sitobion avenae*, *Schizaphis graminum*, de leur évolution et de leur relation avec les plantes hôtes (céréales) pourrait contribuer à la mise en place de moyens de lutte pour contrôler les populations de ces ravageurs. Dans ce contexte, cette étude s'est focalisée sur la recherche des éléments transposables des deux superfamilles *Tc1-mariner-IS630* et des *piggyBac*. Les ETs, considérés comme des moteurs de la plasticité génomique et de l'évolution des espèces, sont utilisés en biotechnologie pour développer des outils de transfert de gènes.

Dans un premier temps, nous avons recherché des éléments de la famille *mariner*, ou apparentés à cette famille, dans les génomes séquencés de trois espèces de pucerons : *Acyrthosiphon pisum*, *Myzus persicae* et *Diuraphis noxia*. Sur la base de similitude de séquences, nous avons pu caractériser 183 éléments répartis en trois clades. Le premier, commun aux trois espèces, correspond au clade de la sous-famille *irritans* DD34D. Il est subdivisé en trois tribus *Macrosiphinimar*, *Batmar-like elements* et *Dnomar-like elements*. Le deuxième comprend l'élément *rosa* DD41D qui appartient à une famille phylogénétiquement proche de *mariner*. Le troisième comprend des séquences avec de longues répétitions terminales inversées et inclut deux tribus DD40-41D. Ces deux derniers clades, plus répandus chez *A. pisum*, dérivent vraisemblablement d'un ancêtre commun et formeraient une nouvelle famille.

Dans un deuxième temps, nous avons exploité les résultats de la recherche *in silico* pour identifier *in vitro* des éléments de la sous-famille *irritans* chez les pucerons des céréales et chez leur plante hôte. Deux types d'éléments délétées (MITEs) ont été identifiés chez les pucerons, l'un commun à toutes les espèces avec un pourcentage d'identité supérieur à 98% (*Aphidmar*) et l'autre spécifique à *S. avenae* (*Samar2*). Par ailleurs, les génomes des céréales (orge, blé, brachypodium, égilope) ont été analysés en utilisant comme requêtes des séquences d'éléments de la sous-famille *irritans* trouvés chez les aphides. Un seul contig de l'orge cultivar *barke* comprend un élément tronqué de 320 pb, flanqué par de l'ADN génomique de pucerons. La vérification *in vitro* de la présence de cette séquence chez plusieurs cultivars d'orge révèle deux types de séquences. Le premier est similaire à celui trouvé *in silico* chez l'orge, le second correspond à l'élément *Samar2* délété de 7 nucléotides au niveau du point de cassure de la délétion initiale. Ceci suggère l'existence d'un transfert horizontal entre pucerons des céréales et l'orge.

Enfin, l'abondance de données génomiques et la rareté des travaux approfondis portant sur les membres de la superfamille *piggyBac*, nous ont amenés à analyser *in silico* leurs caractéristiques, leur distribution et leur évolution. Un total de 117 séquences protéiques de PBLE (éléments autonomes) et de PGBD (éléments domestiqués), ont été utilisées comme requêtes. Quatre groupes structuraux de PBLE ont été définis en fonction de la présence ou absence de répétitions sub-terminales (directes/inversées). Toutefois, il n'existe aucune relation entre ces quatre groupes et la phylogénie des PBLE. Les PGBD, soumis à une forte sélection purifiante, sont clairement structurés en neuf groupes dont un correspondant à un nouvel ensemble d'éléments domestiqués trouvé chez les Néopterygiens. L'analyse fine des PGBD révèle que le domaine catalytique de la transposase ancestrale n'est pas toujours conservé. La phylogénie générale des PBLE et des PGBD suggère des événements multiples de domestication des PGBD à partir de différents ancêtres PBLE.