# Etude de transferts horizontaux de matériel génétique entre virus et animaux

Vincent Loiseau

▶ **To cite this version:**

Vincent Loiseau. Etude de transferts horizontaux de matériel génétique entre virus et animaux. Evolution [q-bio.PE]. Université Paris-Saclay, 2020. Français. NNT : 2020UPASL018 . tel-02983223

**HAL Id: tel-02983223**
**https://theses.hal.science/tel-02983223**

Submitted on 29 Oct 2020

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

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

Thèse de doctorat

NNT : 2020UPASL018

# Etude de transferts horizontaux de matériel génétique entre virus et animaux

**Thèse de doctorat de l'université Paris-Saclay**

École doctorale n° 577 : Structure et dynamique des systèmes vivants (SDSV)

Spécialité de doctorat : Sciences de la Vie et de la Santé

Unité de recherche : Université Paris-Saclay, CNRS, IRD, UMR Évolution, Génomes, Comportement et Écologie, 91198, Gif-sur-Yvette, France.

Référent : Faculté des sciences d'Orsay

**Thèse présentée et soutenue à Gif-sur-Yvette, le 21 septembre 2020, par**

**Vincent LOISEAU**

## Composition du Jury

| | |
|---|---|
| **Pierre CAPY**<br>Professeur, Université Paris-Saclay (UMR 9191) | Président |
| **Etienne DANCHIN**<br>Directeur de Recherche, Université de Nice-Sophia Antipolis (UMR 7254) | Rapporteur & Examinateur |
| **Marie FABLET**<br>Maîtresse de Conférence, HDR, Université Claude Bernard Lyon 1 (UMR 5558) | Rapporteur & Examinatrice |
| **Laura EME**<br>Chargée de Recherche, Université Paris-Saclay (UMR 8079) | Examinatrice |
| **Elisabeth HERNIOU**<br>Directrice de Recherche, Université de Tours (UMR 7261) | Examinatrice |
| **Clément GILBERT**<br>Chargé de Recherche, HDR, Université Paris-Saclay (UMR 9191) | Directeur de thèse |
| **Richard CORDAUX**<br>Directeur de Recherche, Université de Poitiers (UMR 7267) | Co-directeur de thèse |

# *Remerciements*

*J'aimerais par ces quelques lignes, apporter ma gratitude à tous les gens qui ont d'une manière ou d'une autre contribué au bon déroulement de cette thèse.*

*En premier plan, je tiens à remercier chaleureusement mon directeur de thèse, Clément Gilbert, de m'avoir accompagné tout au long de ces trois années. Je tiens à saluer la diplomatie, la bienveillance, la bonne humeur et l'efficacité dont tu as fait preuve. Merci pour toutes les discussions intéressantes que nous avons pu avoir et pour tous les moments partagés.*

*Je tiens également à remercier mon co-directeur de thèse, Richard Cordaux, avec qui j'ai pu avoir des contacts réguliers tout au long de ma thèse et qui a su apporter un regard pertinent à chaque rencontre. Je vous remercie, toi et Clément, de m'avoir encadré lors de ces trois ans, mais aussi à Poitiers lors de mon année en tant qu'ingénieur d'étude, qui fut le tout début de ce travail de thèse.*

*Ma gratitude va également à Jean Peccoud. Jean fut mon co-encadrant de stage de Master 2 avec Clément. On a eu l'occasion d'échanger à de nombreuses reprises par la suite, notamment à propos d'analyse de données sous R. Tu as participé à chacun de mes comités de thèse, au plus grand profit des analyses que j'ai pu faire par la suite.*

*Mes remerciements vont également à Elisabeth Herniou. Nous avons pu interagir à plusieurs reprises, car tu as gentiment accepté d'être ma tutrice de thèse. Cette collaboration m'a permis d'aller à l'IRBI à Tours pour faire des infections virales (à cette occasion, je tiens aussi à remercier Yannis Moreau de son aide et sa bonne humeur).*

*Je tiens à remercier tous les membres de mon jury de thèse d'avoir accepté d'évaluer mes travaux. Merci à Marie Fablet et Etienne Danchin d'avoir accepté d'être rapporteurs, et de même à Laura Eme, Elisabeth Herniou et Pierre Capy d'avoir accepté d'être examinateurs.*

*Mes pensées vont aussi à tous les membres du laboratoire EGCE. C'est un lieu où il fait bon travailler, l'ambiance est agréable au même titre que les personnes que j'ai pu côtoyer. Je*

*remercie en particulier l'équipe de 11h45 avec qui j'ai eu l'occasion de partager bien des repas. Merci à JB de son aide en informatique, merci à Nicolas d'être toujours de bon conseil, merci à Jonathan, à Jean-Michel, à Aurélie, à Arnaud, à Jean-Luc, à Didier pour les discussions intéressantes que nous avons pu avoir, merci à Sylvie pour sa gentillesse, merci aussi à David et Emilie sans qui je n'aurais pas pu faire beaucoup de manipes. Ils ont toujours pu m'aider pour tout aspect technique, et leur bonne humeur contribue à la bonne ambiance du laboratoire.*

*Je remercie bien sûr toutes les autres personnes avec qui j'ai pu discuter, en particulier Laure, Claire ou encore Hélène. Merci aussi à Béatrice et Jean pour l'aide apportée lors de nos infections de mouches.*

*Merci aussi aux secrétaires Hélène et Sylvie pour leur aide administrative et leur bienveillance à mon égard.*

*Je tiens aussi à remercier tous les « jeunes » du laboratoire avec qui j'ai partagé plein de bons moments. Merci à Julia, Perrine, Inoussa, Romain, Joseph, Marie, Arnaud, Florian, Hanna, Marwa, Cécile, et merci aux stagiaires : Sandra, Clémence, Emilie, Sylvain, Jérémy et Marie-Jeanne.*

*Un remerciement particulier à mes collègues de bureau. Tout d'abord Andréas, dont la bonne humeur et les discussions m'ont apporté du réconfort pendant plus de deux ans. Merci ensuite à Siddarth. La barrière de la langue n'est pas si gênante et j'ai découvert une personne gentille et intéressante.*

*Un grand merci aussi à l'équipe EES du laboratoire EBI de Poitiers. Merci à toutes les personnes sympathiques que j'ai eu l'occasion de rencontrer comme Bouziane, Amine et Alexandre, mes collègues de bureau, ou encore Didier, Pierre, Maryline, Christine, Sophie, Nicolas, Isabelle et les autres. Merci aussi à Thomas, Sylvine, Margot, Victorien, Benjamin et Charlotte.*

*Mon attention va évidemment à mes proches qui m'ont soutenu tout au long de cette aventure doctorale. Merci à eux d'avoir été présents et de leur soutien inconditionnel. Merci enfin à ma petite famille. Tout d'abord merci Aurélie de m'avoir suivi et soutenu coûte que coûte tout au*

*long de mes pérégrinations, que ce soit à Poitiers ou à Gif. Tu m'as toujours fait confiance et m'as constamment aidé à avancer. Merci enfin à Léon, petit cadeau extraordinaire qui m'aide à me poser les bonnes questions. Je ne saurais trop vous remercier tous les deux d'être à mes côtés au quotidien.*

*MERCI.*

# *Table des matières*

# Introduction

# *Introduction*

## I. Préambule

La classification du vivant ainsi que la compréhension des relations entre les organismes ont fait l'objet de nombreux travaux, dont les premiers datent de l'Antiquité avec des philosophes comme Platon ou Aristote (dans son œuvre 'Histoire des animaux'). Faisant suite à un intérêt croissant pour les sciences observé à la fin du Moyen Âge et à la Renaissance, une rupture importante dans cette longue lignée de recherches peut être attribuée à René Descartes, qui pose les bases d'une approche plus rationnelle des sciences fondamentales au 17ème siècle (Descartes 1637). Au cours des 17, 18 et 19ème siècles, la classification du vivant de Carl von Linné avec l'attribution d'un nom binomial à chaque espèce, la découverte et l'étude approfondie de reste fossiles et les théories de l'évolution qui ont émergées (i.e. les théories catastrophistes de Cuvier ou transformiste de Lamarck …), soutenues par nombre d'observations du vivant, ont permis d'asseoir progressivement les fondements de la biologie de l'évolution. La théorie de l'évolution de Charles Darwin, proposée pendant cette période (Darwin 1859), s'avère toujours être le paradigme central de la biologie de l'évolution aujourd'hui. Les travaux de Mendel sur les lois de l'hérédité (Mendel 1866) redécouverts au début du 20ème siècle ont largement enrichi et amélioré la compréhension du fonctionnement de la vie. Par la suite, la découverte du support de l'information génétique par Watson et Crick au milieu du siècle dernier est venue renforcer et étayer les travaux précédemment cités et a fait évoluer la vision que l'on avait du monde vivant (Watson et Crick 1953). C'est ainsi qu'on a réalisé que l'information génétique des êtres vivants était portée par l'ADN, qui constitue le génome des organismes, c'est-à-dire l'ensemble des gènes, codant les différentes fonctions vitales au bon fonctionnement d'un organisme. Forte de toutes ces avancées, la théorie synthétique de l'évolution considérant à la fois l'évolution darwinienne des espèces ainsi que les découvertes liées à la génétique des populations et l'hérédité a pu voir le jour au milieu du 20ème siècle. Depuis, de nombreuses études ont été réalisées, ce qui a permis d'étoffer cette théorie et de préciser davantage les lois régissant le vivant et son évolution (comme la théorie de l'évolution moléculaire neutre de Motoo Kimura) (Kimura 1983).

## II. Transferts horizontaux chez les procaryotes

Entre autres découvertes faites en biologie au 20ème siècle, certains scientifiques, par l'étude des bactéries, ont envisagé la possibilité d'un transfert d'information entre organismes, par des voies non canoniques, c'est-à-dire autrement que par le transfert vertical des gènes. Ce transfert vertical est le moyen le plus connu de transmission de l'information génétique. Deux individus d'une même espèce ou d'espèces non isolées reproductivement, capables de se reproduire, vont engendrer une descendance qui portera une partie des gènes de chacun des deux parents. C'est la base de notre compréhension de l'hérédité des caractères : la transmission verticale de l'ADN. Or, en mettant en contact deux bactéries appartenant à des espèces différentes, et en observant la transmission d'une caractéristique de l'une à l'autre bactérie (donc sans reproduction), la question de la transmission de cette information se posait. C'est à la suite de ce type d'expérience réalisée par Griffith en 1928 ou par Akiba et al. (1960) que la question du transfert horizontal d'ADN s'est posée. Le transfert horizontal pourrait ainsi se définir comme étant la transmission de matériel génétique entre deux individus appartenant ou non à la même espèce, sans reproduction.

Aujourd'hui, notamment grâce à l'avancée des technologies de séquençage, ces transferts horizontaux (TH) sont de plus en plus étudiés et leur place dans l'évolution du vivant apparaît comme étant prépondérante. Les TH chez les bactéries semblent être la norme (de la Cruz et Davies 2000; Wiedenbeck et Cohan 2011), avec de nombreux cas détectés. De même les mécanismes régissant ces TH ont été mis en lumière (chez les procaryotes). Ces mécanismes, illustrés en Figure 0.1, sont la conjugaison (transmission d'ADN entre bactéries via des pili), la transformation (acquisition d'ADN nu depuis l'environnement extérieur) ou bien la transduction (acquisition de matériel génétique médié par les phages) (Thomas et Nielsen 2005; Sun 2018). Il est à noter que les phages impliqués dans la transduction chez les bactéries sont des virus infectant ces bactéries.

De même que chez les bactéries, de nombreux TH ont été décrits chez les archées. De nombreux gènes acquis horizontalement sont impliqués dans le métabolisme et la biogenèse de l'enveloppe cellulaire (Wagner et al. 2017). De ce fait, les TH auraient également joué un rôle important dans l'adaptation des archées à leur environnement. Les mécanismes d'échange d'ADN bactériens ont aussi été mis en évidence chez les archées. De plus, d'autres mécanismes liés aux vésicules extracellulaires ou encore un système d'échange d'ADN spécifique aux archées ont été découverts (Wagner et al. 2017).

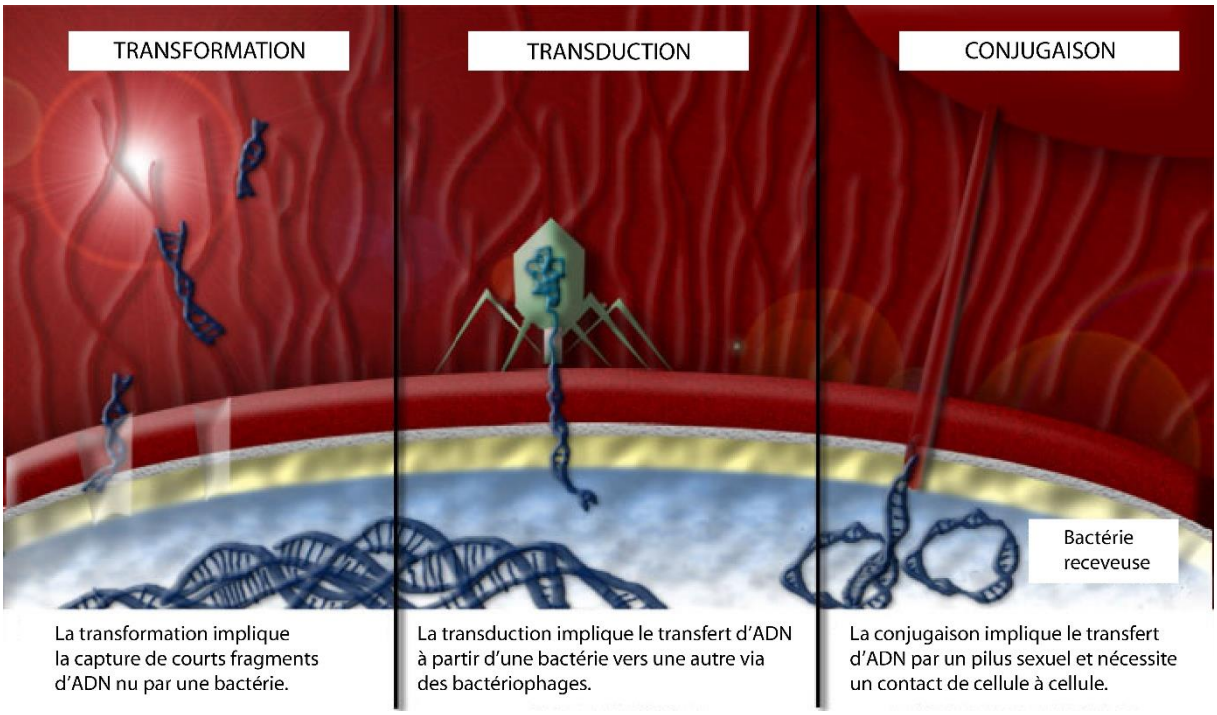



**Figure 0.1: Mécanismes permettant les transferts horizontaux d'ADN chez les bactéries.** Figure modifiée à partir du site internet : https://amrls.cvm.msu.edu/microbiology/molecular-basis-for-antimicrobial-resistance/acquired-resistance/acquisition-of-antimicrobial-resistance-via-horizontal-gene-transfer.

## III. Transferts horizontaux chez les eucaryotes unicellulaires

Sans faire un état des lieux exhaustif des TH découverts chez l'ensemble des eucaryotes unicellulaires jusqu'à aujourd'hui, quelques exemples typiques de TH chez cette catégorie d'eucaryotes méritent d'être exposés.

Dans l'exemple suivant, Nosenko et Bhattacharya (2007) ont identifié, par analyse phylogénomique de données d'expression, 16 gènes ayant subi des TH chez les chromalvéolates, un groupe d'algues unicellulaires. Ils ont montré que ces gènes avaient subi de vieux TH précédant la divergence entre les genres *Karena* et *Karlodinium*. Ces gènes acquis horizontalement à partir de procaryotes et d'autres eucaryotes ont probablement fortement contribué à l'évolution de ces protistes puisqu'ils sont impliqués dans la biogenèse de la membrane plasmique, ainsi que dans diverses fonctions liées au métabolisme énergétique. Bien que la possibilité d'acquisition verticale de ces gènes il y a fort longtemps et leur perte éparpillée dans l'arbre des chromalvéolates ne peut être exclue complètement, les auteurs avancent que l'hypothèse la plus parcimonieuse implique différents évènements de TH.

Autre exemple, une étude phylogénomique suggère que les TH de procaryotes à eucaryotes unicellulaires du genre *Blastocystis* ont contribué grandement à l'adaptation de son

métabolisme (Eme et al. 2017). *Blastocystis* est un genre de parasite unicellulaire appartenant aux straménopiles. On le retrouve dans l'intestin humain, au milieu de nombreux autres eucaryotes unicellulaires et procaryotes. Cet écosystème, de par son importante diversité, serait favorable aux TH entre organismes (Langille, Meehan, Beiko, 2012). Dans cette étude (Eme et al. 2017), les auteurs ont identifié des TH de gènes provenant de procaryotes et d'eucaryotes intestinaux vers le protiste *Blastocystis*. Jusqu'à 2.5% du génome de *Blastocystis* proviendrait de ces TH. Ces gènes sont impliqués dans diverses voies métaboliques, particulièrement dans l'adaptation à l'environnement intestinal (Figure 0.2).

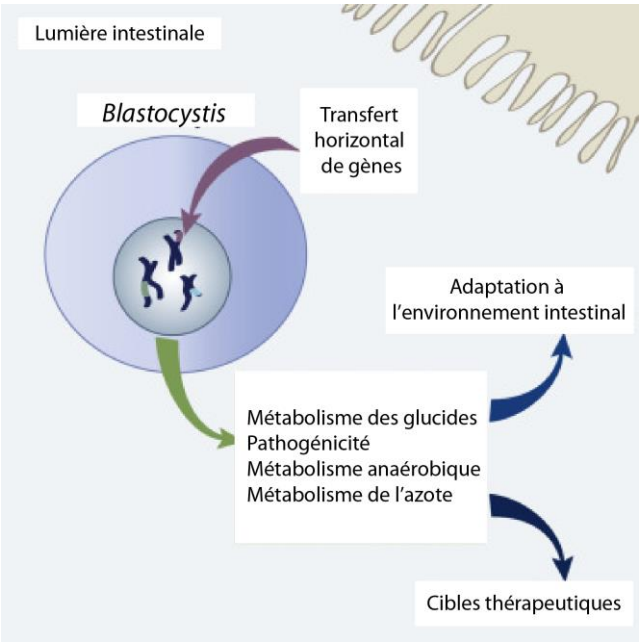



**Figure 0.2: Schéma récapitulatif des TH de gènes acquis par *Blastocystis*, un protiste présent dans l'intestin humain.** Jusqu'à 2.5% de son génome correspond à du matériel génétique acquis par TH provenant de donneurs eucaryotes et procaryotes intestinaux. Figure modifiée issue de Eme et al. (2017).

Une étude récente s'est attachée à détecter les TH de gènes chez les diatomées, un embranchement de microalgues unicellulaires photosynthétiques appartenant au groupe des straménopiles, comme *Blastocystis* (Vancaester et al. 2020). Les auteurs ont recherché à quel point les gènes acquis horizontalement ont contribué au succès écologique des diatomées. Ils ont ainsi recherché des TH de gènes bactériens dans neuf génomes de diatomées disponibles à ce jour. Il ressort de cette analyse que 3 à 5% des gènes des diatomées ont été acquis horizontalement. Fait marquant, plus de 90% de ces gènes sont exprimés et sont impliqués dans l'interaction avec l'environnement et différentes voies métaboliques. Ils ont également montré que la production de vitamine B12 chez les diatomées provient d'un gène acquis via TH. Ce gène, produit par des bactéries et acquis par les diatomées, leur aurait conféré un avantage

sélectif important dans un environnement dépourvu en vitamine B12. Cette étude systématique, la seule ayant été réalisée chez les straménopiles à ce jour, a permis de mettre en avant l'impact majeur joué par les TH sur leur évolution et leur adaptation au milieu.
S'il est aisé de penser que les TH ont pu jouer un rôle important dans l'évolution d'organismes unicellulaires, il n'en est pas de même chez les organismes pluricellulaires.

## IV. Transferts horizontaux chez les eucaryotes multicellulaires

La démocratisation des nouvelles technologies de séquençage depuis ces dix dernières années s'est accompagnée d'un vif regain d'intérêt pour les TH. Malgré une différenciation entre lignée somatique et lignée germinale, rendant l'occurrence des TH d'ADN moins probable, de plus en plus de TH sont détectés en analysant les génomes d'eucaryotes multicellulaires. À mesure que les exemples spectaculaires fleurissent dans la littérature, les TH sont davantage reconnus comme un processus évolutif important non seulement chez les organismes unicellulaires, mais également chez les eucaryotes pluricellulaires, plantes et métazoaires compris.
L'un des cas les plus emblématiques et les mieux connus de TH transrègnes concerne la transformation génétique de plante médiée par les bactéries du genre *Agrobacterium*. *Agrobacterium* est une bactérie pathogène de plantes pouvant engendrer une croissance néoplastique, c'est-à-dire une division cellulaire incontrôlée chez les plantes hôtes, créant ainsi des galles ou bien des racines poussant sans arrêt. Ces pathologies sont causées par le transfert de segments d'ADN d'*Agrobacterium* dans le génome de la cellule infectée (Quispe-Huamanquispe, Gheysen, et Kreuze 2017). La plupart des gènes bactériens nécessaires au transfert d'ADN se trouvent dans un grand plasmide induisant une tumeur ou une racine (plasmide Ti / Ri) qui contient également la partie du plasmide qui est transférée (ADN transféré ou ADN-T). Au cours de l'infection par la bactérie, les composés phénoliques d'origine végétale déclenchent l'expression des gènes de virulence de la bactérie, et les protéines codées assurent la médiation du transfert d'ADN-T vers la cellule végétale hôte. Le destin final de l'ADN-T dans la cellule hôte dépend de diverses interactions entre *Agrobacterium* et les protéines végétales. Plusieurs voies cellulaires hôtes sont utilisées pour garantir que l'ADN-T est importé dans le noyau et intégré dans le génome hôte (Lacroix et Citovsky 2016). L'expression des gènes d'ADN-T dans la plante peut modifier la physiologie pour stimuler la division cellulaire et la croissance des racines. Les gènes iaaM et iaaH codent des enzymes pour la biosynthèse de l'auxine qui est essentielle pour le développement de la galle de la couronne (Y. Zhang et al.

2015). Plusieurs gènes rol (loci racinaires) sont impliqués dans la formation des racines tandis que la fonction de plusieurs gènes d'ADN-T tels que C-prot est encore inconnue (Otten et al. 1999). Les opines sont également codées sur les ADN-T, elles sont utilisées comme sources de carbone et d'azote par les bactéries envahissantes et leur présence peut altérer l'environnement racinaire biologique, en particulier les populations bactériennes associées aux racines (Oger et al. 1997).

La capacité d'*Agrobacterium* à transformer les plantes a été exploitée pendant des décennies comme moyen d'introduire des gènes étrangers d'intérêt dans les plantes cultivées (Tzfira et Citovsky 2006; Gelvin 2009). Cependant, le TH de gène médié par *Agrobacterium* n'est pas limité à la production de cultures génétiquement modifiées. Des preuves du transfert naturel des gènes d'ADN-T d'*Agrobacterium* dans les génomes des plantes et de leur maintien ultérieur dans la lignée germinale ont été documentées chez *Nicotiana*, *Linaria* et plus récemment chez des espèces d'*Ipomoea* (White et al. 1983; Intrieri et Buiatti 2001; Matveeva et al. 2012; Pavlova, Matveeva, et Lutova 2014; Kyndt et al. 2015). Dans ces exemples, les gènes transférés sont fixés et sont exprimés dans la lignée de la plante hôte, suggérant qu'ils pourraient avoir un rôle fonctionnel. Cet exemple de TH est emblématique de l'importance des transferts chez les eucaryotes pluricellulaires.

Toujours concernant les plantes, une étude plus récente a mis en évidence le TH de 23 fragments génomiques impliquant 57 gènes ayant subi un transfert entre la graminée *Alloteropsis semialata* et les 146 autres génomes de graminées disponibles à ce jour (Dunning et al. 2019), augmentant le nombre de TH détectés impliquant cette espèce. Contrairement à l'exemple d'*Agrobacterium*, ces TH ont eu lieu de plante à plante, même si un intermédiaire inconnu (un vecteur) peut avoir été impliqué dans ces TH.

Les TH de matériel génétique ont également été étudié chez les nématodes, dont le mode de vie parasitaire de certaines lignées est en partie dû à l'acquisition de gènes par TH provenant de bactéries et autres eucaryotes. Il a par exemple été mis en évidence que le parasitisme de plantes par des nématodes impliquait l'acquisition d'un gène par TH codant pour une enzyme responsable de la dégradation de la paroi pectocellulosique des plantes (Haegeman, Jones, et Danchin 2011). D'autres cas de TH chez les nématodes ont également été révélés, comme chez le ver nématode à galle *Meloidogyne incognita*, dont 3,34% des gènes auraient été acquis par TH à partir de bactéries et différents groupes d'eucaryotes, suivis de divers réarrangements chromosomiques (Paganini et al. 2012); ou encore chez un autre nématode parasite de plantes *Globodera rostochiensis*, dont 3,5% des gènes auraient été acquis horizontalement (Eves-van den Akker et al. 2016). La majorité de ces TH impliquent des gènes liés au mode de vie

parasitaire de ces nématodes. L'origine de ces derniers pourrait ainsi reposer sur leur capacité d'acquisition de ce type de gènes par TH (Haegeman, Jones, et Danchin 2011).

Concernant les TH chez les eucaryotes, les relations hôtes-parasites semblent revêtir une importance particulière (i.e. les exemples chez les nématodes parasites de plantes cités précédemment). Dans ce contexte, il convient d'évoquer le cas de l'alphaprotéobactérie *Wolbachia pipientis*, l'endosymbiote le plus répandu chez les métazoaires. *Wolbachia* infecte principalement des invertébrés, en particulier des insectes, dont environ 65% des espèces peuvent être porteuses de cette bactérie (Hilgenboecker et al. 2008). Bien que *Wolbachia* soit une bactérie à transmission verticale, sa présence au sein de cellules hôtes et son tropisme pour les gamètes femelles font que tout TH de matériel génétique de *Wolbachia* à l'hôte pourrait avoir un impact évolutif. C'est ainsi que naturellement des chercheurs se sont intéressés aux TH de gènes de cette bactérie dans les génomes hôtes, et de tels TH ont été trouvés. On peut citer comme exemple l'étude de Kondo et al. (2002) dans laquelle les auteurs ont mis en évidence le TH de séquence d'ADN de *Wolbachia* chez la bruche chinoise (*Callosobruchus chinensis*). En 2007, Dunning-Hotopp et al. (2007) ont révélé que la quasi-totalité du génome d'une souche de *Wolbachia* (> 1 Mb) était intégrée dans un chromosome du diptère *Drosophila ananassae*. On trouve également des *Wolbachia* chez certaines espèces de cloportes, comme l'armadille vulgaire (*Armadillidium vulgare*). Il a été montré chez cette espèce un TH de gène de la bactérie *Wolbachia* au génome de l'hôte ayant un impact phénotypique avéré. Ce fragment serait responsable, malgré l'absence de *Wolbachia*, de la féminisation des mâles génétiques en femelles phénotypiques (Legrand et Juchault 1984; Leclercq et al. 2016; Cordaux et Gilbert 2017). Ce cas représente à ce jour le seul cas d'insert fonctionnel avéré du génome de *Wolbachia* dans un génome hôte.

Un autre exemple frappant de l'importance des TH chez les eucaryotes concerne les fragments d'ADN de mitochondrie (un organite cellulaire ayant son propre génome et dérivant d'une bactérie ancestrale) acquis par le génome nucléaire. Ces copies d'ADN mitochondriales dans le génome nucléaire sont appelées NUMTs. On retrouve ce type de transfert chez de nombreux eucaryotes. Des auteurs ont d'ailleurs montré chez l'humain que ce processus était toujours à l'œuvre aujourd'hui, puisque des loci polymorphiques concernant des NUMTs ont pu être détectés (Hazkani-Covo, Zeller, et Martin 2010). De plus, la présence de cinq NUMTs a pu être corrélée avec des maladies génétiques. La quantité de NUMTs dans les génomes eucaryotes semble corrélée à la taille du génome, ce qui laisse penser que l'intégration de ces fragments d'ADN mitochondrial dans le génome nucléaire se ferait par assemblage d'extrémités non-homologues lors de cassures double-brin. Or, le taux d'insertion de ces NUMTs serait limité

par la fréquence des cassures double-brin de l'ADN, dépendante de la taille du génome nucléaire (Hazkani-Covo, Zeller, et Martin 2010).

## V. Virus endogènes dans les génomes de métazoaires

Un autre cas de TH concerne la présence de virus endogènes intégrés dans les génomes de leurs hôtes. Ces virus provenant d'hôtes donneurs ont pu s'intégrer dans le génome d'hôtes receveurs, ce qui constitue en cela un TH. Dans les années 1960 déjà, des cas d'intégration de génomes rétroviraux avaient été mis en évidence dans des génomes de vertébrés (Temin 1964 ; Sambrook et al. 1968). Aujourd'hui, l'analyse des génomes eucaryotes, avec le nombre croissant de génomes séquencés, permet de détecter la présence de séquences virales, voire de génomes viraux entiers intégrés au génome hôte. D'un point de vue évolutif, si ces éléments viraux endogènes (EVE) ne sont pas (trop) délétères pour l'hôte, leur intégration dans la lignée germinale pourrait être transmise à une descendance viable et ainsi perdurer, voire se fixer au sein de l'espèce. Cette considération peut être élargie à tout matériel génétique (viral ou non viral) transféré horizontalement. Si les rétrovirus endogènes ont représenté la majeure partie des séquences provirales détectées dans les génomes, dues à leurs capacités d'intégration, tous les types de virus ont aujourd'hui été détectés dans les génomes eucaryotes. Il y a par exemple l'étude publiée par Flynn et Moreau qui a révélé la présence d'EVE correspondant à des virus à ARN simple brin, ADN simple brin, ADN double brin ou à des rétrovirus dans les génomes de fourmis (Flynn et Moreau 2019). Les auteurs proposent que la présence en majorité d'EVE de virus à ARN simple brin serait lié au fait que ce type de virus est connu pour être le groupe viral principal infectant les fourmis.

On peut également citer différents rétrovirus endogènes détectés dans les génomes de drosophiles comme ZAM chez *D. melanogaster* (Leblanc et al. 2000) ou *tirant* chez *D. simulans* (Akkouche et al. 2012). Les cycles de réplication de ces rétrovirus ont été décortiqués. Ces séquences endogènes ne sont pas toujours fixées dans les populations et pourraient, comme chez les vertébrés, avoir été cooptées par l'hôte, remplissant aujourd'hui des fonctions cellulaires (Fablet et al. 2019).

En effet, des études ont montré, en particulier chez les mammifères, que certains EVE avaient été cooptés dans le génome hôte (Frank et Feschotte 2017). On peut citer l'exemple du spermophile rayé (*Ictidomys tridecemlineatus*) chez qui la présence d'un EVE de bornavirus (virus à ARN simple brin) protège les individus porteurs contre des bornavirus exogènes (virus libres, non intégrés au génome hôte et se propageant horizontalement en infectant les individus)

(Fujino et al. 2014). Un autre exemple démontrant l'impact fonctionnel de virus insérés dans les génomes animaux est celui des syncytines, protéines jouant un rôle essentiel au moment de la formation du placenta lors d'une grossesse chez les mammifères, qui sont en réalité des glycoprotéines d'enveloppe d'origine virale (Dupressoir et al. 2009).

## VI. Transferts horizontaux de gènes entre une guêpe parasitoïde utilisant un virus domestiqué et son hôte

Il est intéressant de s'arrêter un moment sur une catégorie de virus dits « domestiqués » présents au sein du génome de guêpes parasitoïdes, les polydnavirus. Ces virus constituent un exemple unique. En effet, des guêpes ont maintenu la machinerie virale complexe de ces virus au sein de leur génome, cette dernière étant alors impliquée dans des fonctions essentielles pour le succès reproducteur de ces guêpes. Les polydnavirus sont présents dans le génome de milliers de guêpes parasitoïdes, qui pondent leurs œufs à l'intérieur de larves de lépidoptères (Gundersen-Rindal et al. 2013). La guêpe utilise ces virus pour parasiter avec succès la larve hôte en déjouant son système immunitaire (c'est pourquoi on parle dans ce cas de virus « domestiqués » ; Strand et Burke 2013). Les polydnavirus, comprenant deux catégories de virus, les bracovirus et les ichnovirus, sont respectivement associés aux guêpes braconides et ichneumonides. Chez les unes ou les autres guêpes, ces virus sont utilisés pour délivrer des gènes de virulence dans les hôtes parasités (Herniou et al. 2013). Ils ont évolué par convergence, à travers au moins deux événements d'intégration d'ADN viral dans le génome de guêpes (Drezen et al. 2017; Gauthier, Drezen, et Herniou 2018). On sait aujourd'hui que les bracovirus proviennent d'une intégration d'un nudivirus dans le génome d'une ancienne espèce de guêpe, cela ayant eu lieu il y a environ 100 millions d'années (Bézier et al. 2009; Theze et al. 2011). Chez les guêpes associées au bracovirus, presque toutes les fonctions principales du virus ont été conservées (Bézier et al. 2009; Burke et al. 2013; Wetterwald et al. 2010). Le cycle de vie du bracovirus comporte ainsi les différentes étapes retrouvées classiquement dans le cycle de réplication d'un virus. En effet, le virus infecte des cellules cibles et se réplique, formant de nouvelles particules virales. En revanche, ces étapes ne dépendent pas que du virus, mais sont partagées entre la guêpe et l'hôte parasité (Drezen et al. 2017). Ainsi, les bracovirus ont leur génome éparpillé en plusieurs cercles viraux codant des facteurs de virulence le long du génome de la guêpe.

Récemment, les recherches de transfert d'ADN médié par les bracovirus ont mis en évidence la présence de séquences de bracovirus dans divers génomes de lépidoptères comme le

monarque (*Danaus plexippus*), les noctuelles du maïs et de la betterave (*Spodoptera frugiperda et Spodoptera exigua, respectivement*), ou le bombyx du mûrier (*Bombyx mori*). Ces résultats apparaissent à première vue surprenants dans la mesure où les guêpes parasitoïdes tuent l'hôte parasité, empêchant donc tout impact évolutif. Ces fragments d'ADN viraux détectés dans les génomes de lépidoptères présentent jusqu'à 90% similarité (nucléotidique) avec l'ADN de bracovirus, cela sur plusieurs kilobases. De tels niveaux de similarité sur une telle longueur correspondent à un niveau de conservation de séquence inattendu entre un virus et un lépidoptère. La plus grande des insertions détectées (6,5 kb) correspond à plus de la moitié de la séquence d'un cercle de bracovirus. Les autres inserts sont généralement limités à un gène avec quelques séquences régulatrices autour (Gasmi et al. 2015). Dans un cas, une séquence régulatrice (impliquée dans la production d'un cercle de bracovirus) est présente dans la séquence transférée chez un lépidoptère, ceci nous renseignant sur le sens du transfert, à savoir du bracovirus vers le lépidoptère (Gasmi et al. 2015). Dans cette étude, l'intégration d'ADN de bracovirus dans les génomes d'insectes a été vérifiée en séquençant les jonctions entre l'ADN du lépidoptère concerné et l'insert viral, cela sur plusieurs individus provenant de différentes collections. Ces pratiques robustes apportent du poids quant à la présence véritable de fragments d'ADN de bracovirus dans les génomes de lépidoptères. Cela nous permet de penser que ces intégrations sont fixées dans les populations des lépidoptères concernés. De plus, beaucoup de courtes séquences originaires de bracovirus (quelques centaines de paires de bases de long) ont également été détectées dans les génomes du monarque et du bombyx du mûrier, par des méthodes bio-informatiques (Schneider et Thomas 2014). Des analyses fonctionnelles suggèrent que ces gènes de bracovirus dans les génomes de lépidoptères pourraient jouer un rôle dans la défense contre un baculovirus (Gasmi et al. 2015). Ainsi se dessine une course à l'armement dans laquelle des lépidoptères utiliseraient des gènes d'un virus pour lutter contre un autre virus, cela entraînant potentiellement des TH de matériel génétique au passage (Figure 0.3).

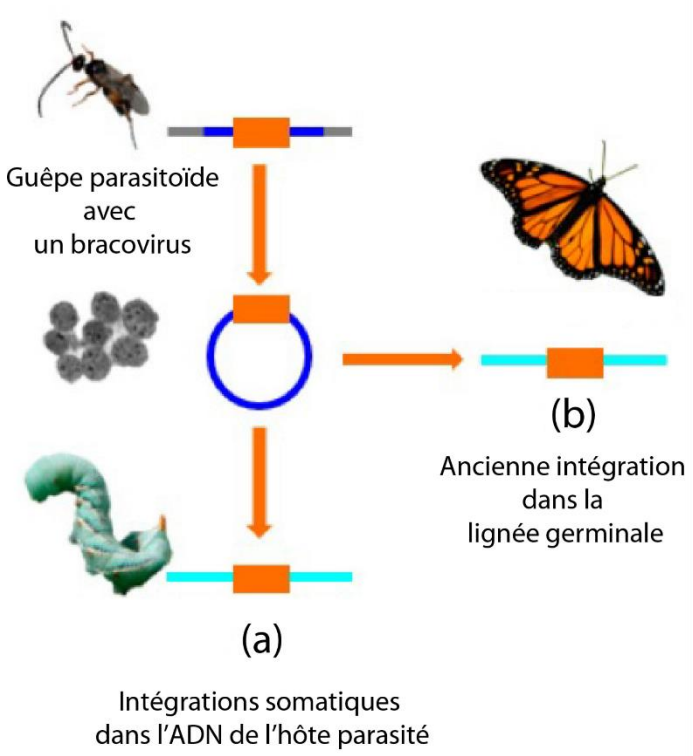



**Figure 0.3 : Transfert d'ADN par l'intermédiaire d'un bracovirus.** (**a**) Les bracovirus endogènes associés aux guêpes parasitoïdes produisent des particules virales et les cercles d'ADN empaquetés dans ces particules s'intègrent dans l'ADN des cellules hôtes parasitées. (b) La présence de séquences de bracovirus fixées dans les génomes de plusieurs lignées de lépidoptères suggère que quelquefois les cercles bracoviraux peuvent aussi s'intégrer dans le génome de cellules germinales. Ces insertions conféreraient un avantage sélectif aux individus produisant certains produits de gènes bracoviraux comme la protection contre d'autres virus. Les rectangles de couleur représentent les gènes transférés, les lignes de couleur représentent l'ADN génomique des espèces représentées, le cercle correspond à un cercle de bracovirus empaqueté dans une particule. Crédits Image : A. Bézier, J. Gaillard & J. Herbinière. Figure modifiée à partir de Drezen et al. 2017.

Dans une étude récente, Chevignon et al. (2018) ont détecté des fragments d'ADN de bracovirus intégrés au sein des hémocytes de l'hôte parasité, le sphinx du tabac (*Manduca sexta*). Une approche par PCR réalisée sur un échantillon des cercles indiquait qu'ils ont persisté dans les hémocytes de *M. sexta* sous forme linéaire et ont été potentiellement intégrés au génome de ces cellules. De plus, les auteurs ont mis en évidence que huit cercles de bracovirus étaient intégrés dans le génome des hémocytes dans des régions spécifiques reconnues par le motif d'insertion de l'hôte (ou « host insertion motif », HIM). Ainsi, un mécanisme d'insertion spécifique médié par ces HIM est à l'œuvre dans ces intégrations. De plus, des régions de génome de l'hôte *M. sexta* apparaissent être enrichies en sites d'insertion des cercles viraux. L'identification d'un mécanisme d'intégration efficace et spécifique partagé par plusieurs espèces de bracovirus pose la question du rôle de ce mécanisme dans le succès de parasitisme des guêpes braconides. Les

résultats obtenus par Chevignon et al. montrent l'intégration d'ADN de bracovirus dans des cellules immunitaires à chaque événement de parasitisme. Étant donné que les bracovirus ne se répliquent pas dans les cellules qu'ils infectent, l'intégration de séquences virales dans l'ADN hôte pourrait permettre la production de protéines de virulence au sein des cellules hôtes en division. De plus, ce processus d'intégration pourrait servir de base pour comprendre comment les polydnavirus servent de vecteurs d'ADN entre les guêpes parasitoïdes et leurs hôtes lépidoptères. Bien que ces TH concernent des cellules somatiques, ils soulèvent la possibilité que des TH se produisent aussi dans la lignée germinale, ce qui est d'autant plus intéressant que toutes les chenilles ne meurent pas à la suite d'un événement de parasitisme, ce qui permettrait un possible héritage vertical du matériel génétique acquis lors de ces TH.

# VII. Les transferts horizontaux chez les métazoaires impliquent surtout les éléments transposables

## a. Qu'est-ce qu'un élément transposable ?

La grande majorité des TH de matériel génétique identifiés jusqu'à présent entre métazoaires impliquent des éléments transposables (ET) (Gilbert et Feschotte 2018). Plusieurs milliers de ces TH d'ET ont été identifiés (référencés dans la base de données HTT-DB, Dotto et al. 2018), alors que seulement un exemple de TH d'un gène non-ET entre animaux (une protéine antigel partagée entre diverses espèces de poissons) a été décrit à ce jour (Graham et al. 2008). Les ET sont des éléments génétiques mobiles égoïstes, capables de se déplacer et se dupliquer au sein des génomes. On les classe généralement en deux catégories : les ET de classe I ou rétrotransposons qui transposent par un mécanisme de copier/coller via un intermédiaire ARN ; et les ET de classe II ou transposons à ADN qui transposent par un mécanisme de couper/coller, via un intermédiaire ADN (Wicker et al. 2007). De par les nombreuses séquences répétées d'ET dans les génomes, leur étude s'est révélée et s'avère toujours compliquée car ils entravent l'assemblage de génomes de bonne qualité. Si leur nombre de copies au sein des génomes de métazoaires est variable, on considère que virtuellement tous les eucaryotes portent des ET dans leur génome. On peut citer en exemple le génome humain dont on estime qu'environ 50% correspondent à des ET (Lander et al. 2001), ou encore le génome du maïs composé à plus de 85% d'ET (Schnable et al. 2009). Les ET sont ainsi une part de l'ADN répété non génique des génomes et ont longtemps été perçus comme inutiles. De plus, les vagues de transposition des ET dans les génomes entraînent souvent des mutations délétères, impactant négativement la

valeur sélective de l'hôte. Aujourd'hui, cette vision tend à être nuancée au fur et à mesure des études montrant que ces séquences peuvent façonner les génomes hôtes et influencer l'évolution des organismes, en étant une source de mutations, de polymorphisme génétique, de réarrangements chromosomiques et en participant à la régulation de réseaux de gènes (Biémont et Vieira 2006; Feschotte et Pritham 2007; Cordaux et Batzer 2009; Chénais et al. 2012; Chuong, Elde, et Feschotte 2017; Bourque et al. 2018).

### b. Les transferts horizontaux entre organismes comme moyen de persistance des éléments transposables

La question de savoir comment les ET réussissent à perdurer dans les génomes sans procurer de bénéfices directs à leurs hôtes a suscité beaucoup de travaux (Le Rouzic et al. 2007; Brookfield et al. 1997). Étant donné leur impact principalement délétère sur la valeur sélective de l'hôte, des mécanismes d'extinction des séquences actives d'ET sont apparus, comme la compaction de la chromatine rendant la région génomique peu accessible à la transcription, ou par la voie des petits ARN interagissant avec les protéines de la famille PIWI, appelés piARN (Slotkin et Martienssen 2007). Une fois l'activité des séquences d'ET limitée, elles vont évoluer de façon neutre et ainsi accumuler des mutations par dérive génétique, les rendant non fonctionnelles (Szitenberg et al. 2016; Arkhipova 2018). En l'absence d'autres mécanismes, ces ET seraient ainsi voués à disparaître. Pour expliquer le fait que l'on trouve toujours aujourd'hui des ET dans les génomes, deux hypothèses, qui ne s'excluent pas, peuvent être émises : soit les ET apparaissent *de novo* au sein des génomes eux-mêmes, soit les ET ont un moyen d'échapper à leur sort funeste au sein d'un génome.

La première hypothèse ne peut être simplement rejetée puisque les ET sont bien apparus un jour et qu'on leur trouve des similarités d'une part avec des virus, et d'autre part avec d'autres séquences dans les génomes comme des ARN de transfert ou des ARN codant des sous-unités ribosomiques, qui sont similaires à certaines séquences d'ET non autonomes (Krupovic et Koonin 2015; Ohshima et Okada 2005).

Mais bon nombre d'ET similaires sont détectés entre espèces différentes, suggérant fortement qu'un des moyens de leur persistance soit lié à leur transfert horizontal, qui serait facilité par l'activité de transposition de copies actives d'un génome hôte à un génome naïf, ces copies ne pouvant alors être réprimées chez ce dernier. Il est techniquement possible d'inférer un TH d'ET en comparant en autres les séquences nucléotidiques de ces éléments présents chez différentes espèces. Si la similarité entre ces séquences est trop élevée et incompatible avec

l'hypothèse de leur présence par transmission verticale étant donné le temps de divergence entre ces espèces, l'hypothèse d'un TH de ces ET est alors privilégiée (Peccoud, Cordaux, et Gilbert 2018).

Dans une étude de 2017, nous avions recherché de manière systématique les TH d'ET chez les insectes, et mis en évidence 2 248 de ces TH ayant eu lieu au cours des dix derniers millions d'années, à partir de l'analyse de 195 génomes (Peccoud et al. 2017). La majorité de ces TH impliquent des ET de classe II. La plus forte proportion de TH de transposons à ADN que de TH de rétrotransposons pourrait être due au fait que l'intermédiaire de transposition sous forme ADN est plus stable que celui sous forme d'ARN. Le complexe transposase/ADN pourrait mieux supporter les conditions de TH, impliquant certainement une étape extracellulaire, que le complexe transcriptase-inverse/ARN. De plus, il est possible que les rétrotransposons nécessitent plus d'interactions avec des facteurs cellulaires hôtes pour une rétrotransposition réussie, contrairement aux transposons à ADN. Le nombre de facteurs cellulaires impliqués dans un événement de transposition pourrait ainsi être également limitant dans le cadre d'un TH, les facteurs cellulaires différant d'une espèce à l'autre.

# VIII. Mécanismes responsables des transferts horizontaux d'éléments transposables chez les métazoaires

Il est intéressant de découvrir toujours plus de TH d'ET chez les métazoaires et de mettre en évidence leur impact dans l'évolution de l'espèce receveuse. Mais ces TH soulèvent la question des mécanismes sous-jacents. Il paraît difficile d'imaginer que ces mécanismes soient les mêmes que ceux ayant lieu chez les procaryotes, spécialement à cause de la séparation entre les lignées germinale et somatique, et le nombre relativement faible de ces cellules germinales. De plus, aucun mécanisme spécifiquement dédié au TH entre eucaryotes n'est connu. Néanmoins ces TH ont lieu, et des mécanismes impliquant des conditions écologiques particulières doivent avoir lieu, permettant ainsi aux TH de se produire.

## a. Les vésicules extracellulaires

L'un des candidats potentiels qui pourraient permettre le passage d'ADN d'un donneur à un receveur sont les vésicules extracellulaires (VE). Les VE sont de petites vésicules membranaires, telles que les exosomes et les vésicules excrétées, qui peuvent être libérées de presque tous les types cellulaires, et dont la formation peut être spontanée ou induite par divers

stimuli. Les exosomes sont de petits véhicules extracellulaires dont la taille varie entre 50 nm et 100 nm de diamètre, avec une densité de 1,10–1,21 g/ml. Ils sont libérés par exocytose des corps multivésiculaires, déclenchée par des molécules clés comme le céramide sphingolipide et la protéine X interagissant avec l'ALG-2 (ALIX). Les vésicules excrétées ont une plus grande taille, entre 100 nm et 1 µm de diamètre (Cai et al. 2016). Elles sont relarguées des cellules maternelles par bourgeonnement d'une membrane plasmique suivie d'une fission de leur tige membranaire (Cocucci, Racchetti, et Meldolesi 2009; Cai et al. 2016; Figure 0.4). Les VE peuvent varier dans leur formation, leur taille, leur abondance et leur composition, mais elles contiennent souvent d'abondantes protéines transmembranaires et cytosoliques, ARNm, miARN et ADN (Trajkovic et al. 2008; Cai et al. 2015; L. Zhang et al. 2015; Zomer et al. 2015). Les VE permettent le TH de molécules transportées d'une cellule à une autre. Ces vésicules peuvent être excrétées de presque tous les types cellulaires à la fois en conditions physiologiques et pathologiques (Théry, Ostrowski, et Segura 2009; Arita et al. 2008; Cocucci, Racchetti, et Meldolesi 2009; Trajkovic et al. 2008). De plus, les composants transférés dans les VE sont fonctionnels et peuvent réguler les fonctions biologiques des cellules réceptrices. Jusqu'à présent, on pensait que la plupart des cellules libéraient une abondance de VE contenant un ensemble sélectionné de protéines et d'ARN (Skog et al. 2008; Valadi et al. 2007). En outre, de nombreuses études se sont concentrées sur les microARN (miARN) dans les VE qui sont connus pour contrôler l'expression des gènes en régulant le renouvellement de l'ARNm dans les cellules réceptrices, et également pour leur implication dans les cancers (Moldovan et al. 2013; Pfeifer, Werner, et Jansen 2015; Ramshani et al. 2019; Yoshikawa et al. 2019; Battaglia et al. 2019; Groot et Lee 2020). Cependant, de récentes études ont montré que ces VE pouvaient contenir de l'ADN qui serait transporté entre cellules du même organisme, provenant généralement de cellules cancéreuses (Cai et al. 2016; Kalluri et LeBleu 2016; Klump et al. 2018; Vagner et al. 2018; Jabalee, Towle, et Garnis 2018; Lázaro-Ibáñez et al. 2014; 2019). Cet ADN pourrait influencer la fonction des cellules réceptrices en augmentant les niveaux de protéines et d'ARNm (Cai et al. 2013; 2014). Un certain nombre de gènes transférés d'une cellule à l'autre par la voie des vésicules extracellulaires et impliqués dans des maladies humaines ont également été mis en évidence (Cai et al. 2016, Tableau 0.1).

**Tableau 0.1: Rôle des ADN impliqués dans diverses pathologies transférés par des vésicules lors de communication extracellulaire.** Tableau modifié à partir de Cai et al. (2016).

| Gènes transférés | Cellules maternelles | Cellules réceptrices | Maladies associées |
|---|---|---|---|

| | | | |
|---|---|---|---|
| gène $AT_1$ | cellules musculaires lisses vasculaires | cellules HEK293/ cellules musculaires lisses | hypertension/athérosclérose |
| ADN hybride BCR/ABL | cellules de la leucémie myéloïde chronique | cellules HEK293/ neutrophiles | leucémie |
| gène c-Myc | cellules de médulloblastome | - | tumeur |
| gènes MLH1, PTEN, TP53 | cellules cancéreuses de prostate | - | tumeur |
| gènes KRAS, p53 | cellules cancéreuses pancréatiques | - | tumeur |
| gène résistant aux médicaments | érythrocytes infectées par *Plasmodium falciparum* | érythrocytes infectées par *P. falciparum* | infection parasitaire |
| ADN extracellulaire | cellules apoptotiques | cellules de l'immunité | Lupus |
| ADN mitochondrial | astrocytes | astrocytes | maladie d'Alzheimer |
| ADN mitochondrial | myoblastes | myoblastes | maladies des muscles squelettiques |
| gène SRY | leucocytes | leucocytes/ cellules endothéliales | athérosclérose |
| gène chromosomal 333 | cardiomyocytes | fibroblastes | cardiomyopathies |
| ADN chromosomal | cellules épithéliales prostatiques | spermatozoïdes | - |
| ADN H-ras | cellules épithéliales transformées par H-ras | cellules épithéliales | - |

Au fur et à mesure que les VE sont étudiées, on découvre qu'elles peuvent contenir des protéines, différents types d'ARN et même de l'ADN génomique comme des oncogènes. Mais ces VE pourraient-elles également contenir des ET (sous forme d'ARN et/ou d'ADN) et ainsi être un vecteur possible d'ET entre cellules ? Dans une étude de Balaj et al. (2011), les auteurs ont mis en évidence la présence d'ADN et d'ARN dans des microvésicules issues de cellules tumorales en infectant des cellules saines. Cet ADN correspondait à des oncogènes comme c-Myc. De grandes quantités de transcrits ARN ont aussi été détectées, correspondant à des rétrotransposons humains, i.e., des éléments LINE-1 (éléments autonomes) et Alu (éléments non autonomes). De plus, des évènements d'insertion de LINE-1 dans des cellules saines apportées par des VE ont été mis en évidence expérimentalement (Kawamura et al. 2019). Un TH d'ADN incorporé dans des VE a aussi été mis en évidence en utilisant des cellules de la plante *Arabidopsis thaliana* (Fischer et al. 2016). Ces études mettent en avant le rôle de vecteurs des VE dans la communication intercellulaire au sein d'un organisme. Cependant, pour qu'un TH ait un impact évolutif, il est nécessaire que ce dernier se produise entre organismes. Or, les VE n'ont, jusqu'à présent, pas démontré leur capacité de vecteur entre organismes différents. Néanmoins, les VE pourraient être impliqués dans des TH entre animaux dans le cadre des relations hôtes-parasites. En effet, il a été montré que les VE sont impliquées dans l'interaction hôte-parasite et dans la communication entre parasites (i.e. parasite des genres *Trypanosoma* ou *Leishmania*), induisant un dysfonctionnement des réponses immunitaires ou manipulant la physiologie et le métabolisme de l'hôte (Wu et al. 2019; Dong, Filho, et Olivier 2019; Coakley, Maizels, et Buck 2015; Sampaio, Cheng, et Eriksson 2017). C'est pourquoi certains auteurs ont

évoqué la possibilité que les VE puissent agir comme vecteurs de matériel génétique entre espèces, par exemple dans le cadre des relations hôtes-parasites et spécifiquement dans le cas de la malaria (infection humaine par le parasite *Plasmodium falciparum*, Correa et al. 2020) qui a été étudié en détail. On peut également mentionner l'exemple d'un ver parasite de l'intestin chez la souris *Trichuris muris* dont le contenu des VE a été analysé. De nombreuses protéines interagissant avec des cellules de souris ont été détectées, ainsi que des ARNm qui correspondraient à des ET (Eichenberger et al. 2018). Cela permet de penser que ces ET du ver parasite pourraient interagir avec des cellules, voire le génome cellulaire de la souris. Bien que les cellules de souris ne soient pas des cellules de la lignée germinale, ces résultats sont prometteurs pour souligner le possible rôle des VE comme vectrices d'ET entre animaux.

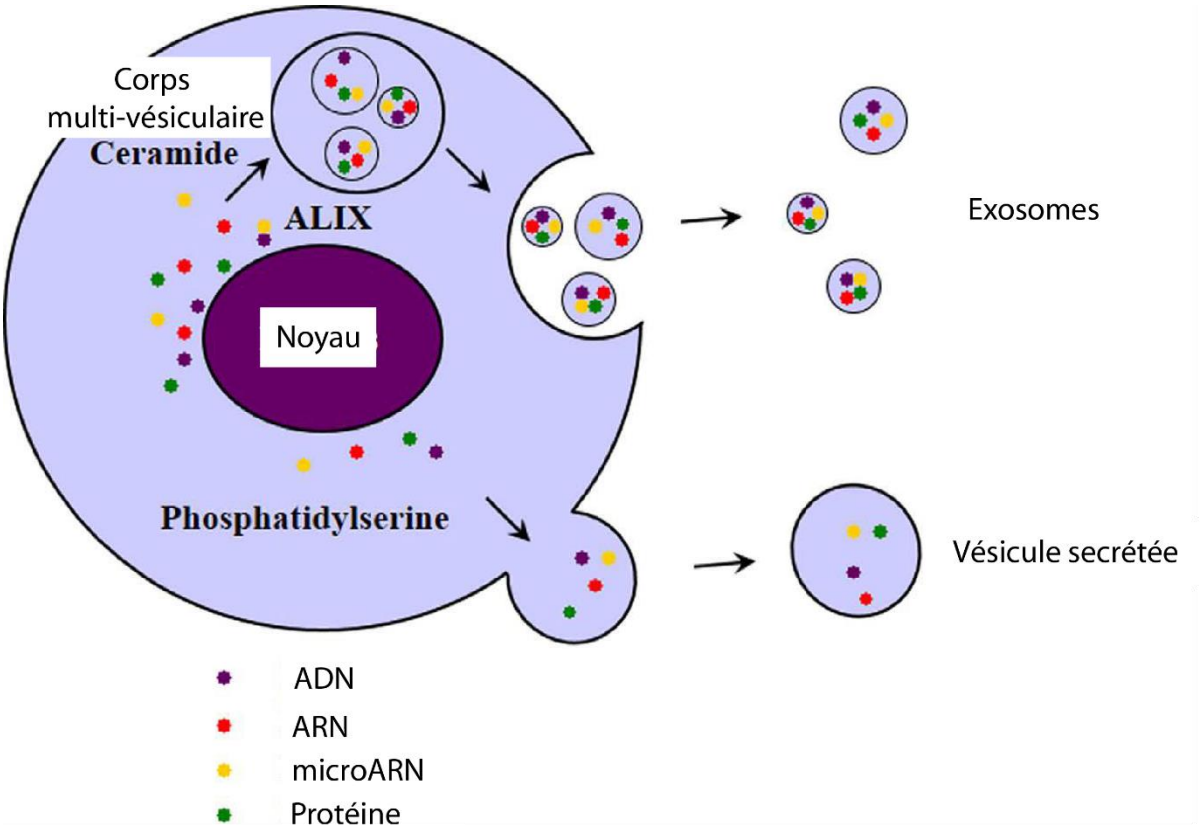



**Figure 0.4: Schéma de la libération et des composés potentiels des vésicules extracellulaires.** Les VE sont libérées par exocytose des cellules (exosomes) et par bourgeonnement à partir de la membrane plasmique (vésicules excrétées). ALIX : ALG-2 interagissant avec la protéine X. Protéine, ARNm, miARN, et ADN sont sélectivement empaquetés dans les VE, puis sécrétés de la cellule. Figure modifiée issue de Cai et al. 2016.

### b. Proximité phylogénétique et géographique

Il est possible que d'autres facteurs puissent expliquer les TH et que les relations hôtes-parasites n'expliquent pas à elles seules le spectre des TH d'ET chez les métazoaires. Dans notre étude

de 2017 (Peccoud et al. 2017), nous avons cherché à détecter de façon systématique les TH d'ET dans tous les génomes d'insectes disponibles au moment de l'étude (195 génomes). L'analyse de ces TH a mis en évidence davantage de TH de transposons à ADN que de rétrotransposons, et a surtout mis en exergue deux facteurs majeurs semblant influencer les TH d'ET (au moins chez les insectes), à savoir la proximité phylogénétique des espèces donneuse et receveuse, et leur proximité géographique. Le fait que des espèces ont d'autant plus de chances d'échanger de l'ADN via TH qu'elles sont proches géographiquement ou physiquement peut paraître intuitif. Néanmoins cela n'avait jamais été formellement montré chez les métazoaires. De plus, la proximité phylogénétique comme facteur facilitant les TH d'ET entre insectes soutient l'hypothèse selon laquelle les ET interagissent avec des facteurs cellulaires (i.e. des protéines). Plus les espèces impliquées dans un transfert sont proches, moins leurs facteurs cellulaires sont divergents et l'ET arrivant dans cette nouvelle espèce aura plus de chances de pouvoir interagir avec ces facteurs cellulaires.

### c. Les relations hôtes-parasites

Bien qu'il n'y ait à ce jour aucun cas formellement avéré de TH d'ET entre un hôte et son parasite par l'intermédiaire de VE (malgré des avancées conséquentes ces dernières années), l'importance des relations hôtes-parasites dans les TH a déjà été mise en lumière. Une étude de Kuraku et al. (2012) a mis en évidence des TH d'ET entre des poissons-téléostéens et des lamproies, parasites de ces poissons. La relation hôte-parasite est avancée pour expliquer ces TH. Un autre exemple frappant a été révélé dans une étude de 2010, dans laquelle les auteurs ont mis en évidence des TH d'ET entre des vertébrés et un insecte, ayant possiblement pu servir de vecteur aux ET pour infecter ensuite diverses espèces (Gilbert et al. 2010). En effet, *Rhodnius prolixus*, une punaise triatomine se nourrissant du sang de différents tétrapodes et vecteur de la maladie de Chagas chez l'homme, porte dans son génome quatre familles de transposons distinctes qui ont également envahi les génomes d'un ensemble de tétrapodes. Les ET d'insectes sont identiques à environ 98% et se regroupent phylogénétiquement avec ceux de l'opossum et du singe-écureuil, deux de ses hôtes mammifères préférés en Amérique du Sud. Les auteurs ont également identifié l'une de ces familles de transposons chez la limnée *Lymnaea stagnatilis*, un vecteur cosmopolite de trématodes pouvant infecter plusieurs vertébrés, dont la séquence ancestrale est presque identique et se regroupe avec celles trouvées chez les mammifères de l'Ancien Monde. Ensemble, ces données soutiennent, sans toutefois le démontrer, le rôle supposé des interactions hôtes-parasites dans la facilitation de TH d'ET chez

les animaux. De plus, la grande quantité d'ADN générée par l'amplification des ET transférés horizontalement soutient l'idée que l'échange de matériel génétique entre les hôtes et les parasites influence leur évolution génomique. Enfin, une étude récente de Reiss et al. (2019) a révélé, parmi divers ordres d'insectes, un excès de TH d'ET chez les lépidoptères. Les auteurs de l'étude mettent en avant, outre les mécanismes déjà évoqués plus haut, l'importance de la biologie de l'hôte dans ces TH. Justement, ils mettent en perspective l'importance des TH d'ET chez les papillons avec la circulation de baculovirus, virus ayant un large tropisme chez les lépidoptères. Il a d'ailleurs été montré à travers diverses études, des années 1980 à aujourd'hui, que ce virus est capable de porter des ET de lépidoptères dans son génome et pourrait ainsi être un vecteur potentiel d'ET entre animaux (Miller et Miller 1982; Fraser, Smith, et Summers 1983; Gilbert et al. 2014; 2016; Loiseau et al. 2020).

### d. Les virus : vecteurs modèles d'éléments transposables entre animaux ?

Certains virus étant capables d'infecter plusieurs espèces ont naturellement été proposés comme potentiels vecteurs d'ADN entre ces espèces (Loreto, Carareto, et Capy 2008). Parmi la grande diversité de virus existants, les virus à grands génomes à ADN double-brin sont particulièrement intéressants puisque le support de leur information génétique est identique à celui de leurs hôtes et, par extension, à celui des ET. Encore faut-il que de l'ADN étranger s'insère dans les génomes viraux, c'est-à-dire que ces virus sont aptes à porter de l'ADN étranger dans leur génome, que cet ADN n'est pas perdu ou modifié dans les génomes viraux de manière que les virus puissent ensuite le transposer dans le génome d'une espèce receveuse lors d'une future infection. Cette idée a été encouragée par la découverte d'ET intégrés au sein de génomes de virus à ADN double-brin comme les poxvirus ou les iridovirus (Piskurek et Okada 2007; Piégu et al. 2013). Il convient de noter que des ARN d'ET ont également été identifiés dans les capsides de certains virus à ARN, ce qui étend potentiellement la capacité cargo à tous les types de virus (Routh, Domitrovic, et Johnson 2012).

#### i. Virus, vecteur d'ADN *in vitro*…

Pour qu'un TH d'ET médié par un virus puisse réussir, il faut que l'ET saute du génome d'un premier hôte à celui du virus puis du virus au génome d'un second hôte. Justement, la capacité des ET à sauter dans certains génomes viraux est connue depuis le début des années 80. Le premier ET découvert dans un virus est TED, un rétrotransposon intégré dans le génome du baculovirus *Autographa californica* multiple polyhedrosis virus (AcMNPV) (Miller and Miller

1982). Par la suite, d'autres ET comme IFP2 (aujourd'hui appelé piggyBac), TFP3 (aujourd'hui appelé tagalong, appartenant aussi à la superfamille piggyBac) et Hitchicker (de la superfamille PIF/Harbinger) ont également été trouvés intégrés au génome d'AcMNPV après infection d'une lignée cellulaire de la fausse arpenteuse du chou (*Trichoplusia ni*, Noctuidae, Lepidoptera), les cellules TN-368 (Bauser, Elick, et Fraser 1996; Fraser, Smith, et Summers 1983; Wang et Fraser 1993). Ces travaux montrent que l'intégration d'un ET de l'hôte dans un génome viral est possible en conditions *in vitro*.

### ii. …Mais aussi *in vivo*

Quelques années plus tard, le même constat a pu être fait après l'infection de chenilles de papillons par un autre baculovirus, *Cydia pomonella* Granulosis Virus (CpGV) (Jehle et al. 1995 ; 1997). Ce sont cette fois-ci des ET de la superfamille Tc1/Mariner qui ont été identifiés au sein de génomes de CpGV (Jehle et al. 1995 ; 1997).

Plus récemment, grâce à une approche de génomique des populations, Gilbert et al. (2014) ont pu arriver au même résultat en utilisant les nouvelles technologies de séquençage. Ils ont séquencé une population d'AcMNPV, après infection de chenilles de *T. ni*. La profondeur du séquençage leur a permis d'identifier deux ET de la fausse arpenteuse du chou appartenant à deux familles différentes (*piggyBac* et *mariner*) insérés à différents loci dans les génomes séquencés d'AcMNPV. Ils ont aussi pu montrer que ces deux transposons avaient été transférés horizontalement entre la fausse arpenteuse du chou et d'autres espèces de papillons très éloignées et connues pour être également susceptibles au virus AcMNPV, le sphinx du tabac (*Manduca sexta*) et la noctuelle du chou (*Helicoverpa armigera*). La très forte identité entre les ET des différentes espèces suggère que les transferts sont extrêmement récents (Figure 0.5). Ces données ont aussi permis de calculer pour la première fois une fréquence d'insertion d'ET de papillons dans les populations d'AcMNPV, équivalente à au moins un transposon dans une population de 8 500 génomes de baculovirus.

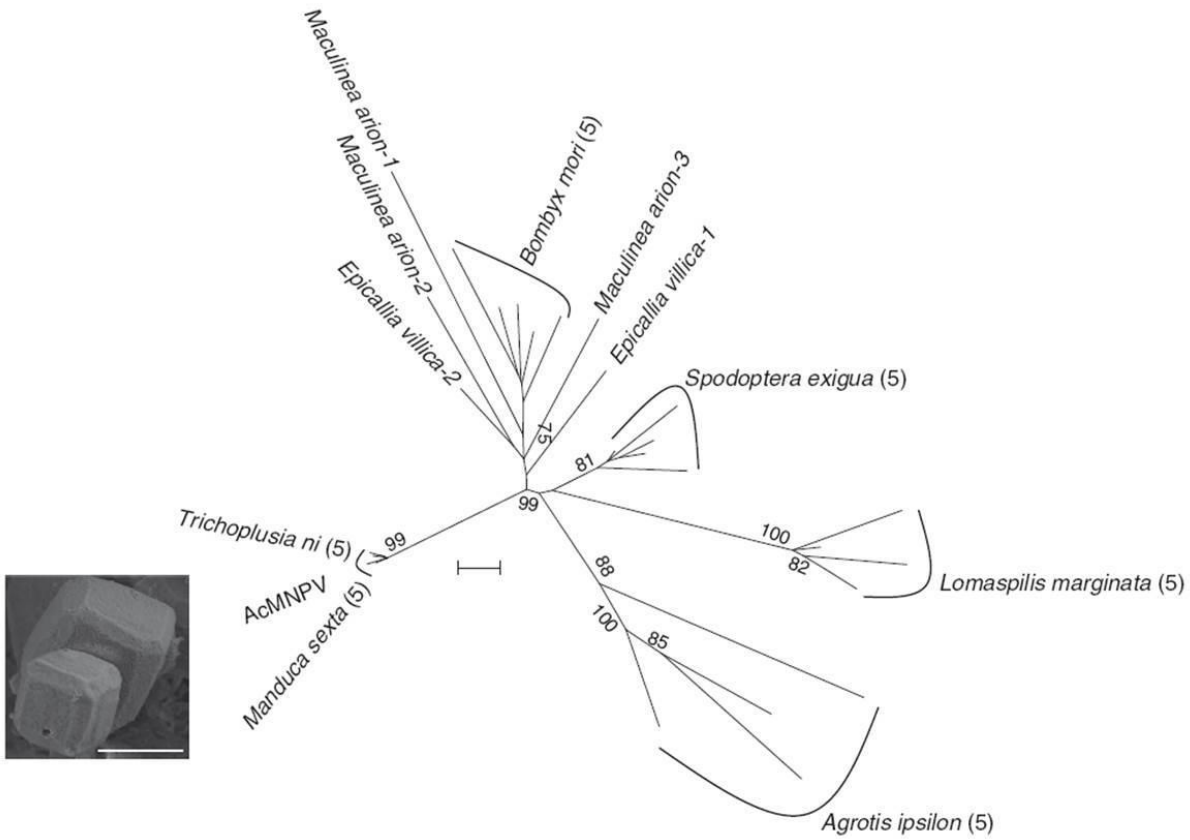



**Figure 0.5: Phylogénie de l'élément MAR1 (Tc1/Mariner), qui a été trouvé intégré dans le génome d'AcMNPV**. L'arbre est un arbre de copies de MAR1. La barre d'échelle pour la longueur des branches est de 0,01 substitution par site. Les valeurs de bootstrap > 70% sont indiquées. Le nombre de copies utilisées pour les analyses phylogénétiques est indiqué entre parenthèses pour chaque espèce. L'ET d'AcMNPV correspond au consensus de toutes les copies trouvées intégrées dans le génome du virus. La photo du virus a été prise en microscopie électronique à balayage. Barre d'échelle blanche = 1 μm. Figure modifiée à partir de Gilbert et al. (2014).

Cette approche qui consistait à chercher des ET de l'hôte à basse fréquence dans des populations de génomes viraux s'est avérée pertinente et a permis de confirmer et de détecter la présence d'ET hôtes insérés dans les génomes viraux à basse fréquence. De plus, cette étude était la première à montrer que des transposons s'intégrant *in vivo* dans un génome viral avaient récemment été transférés horizontalement entre espèces d'insectes susceptibles au virus. Cependant, ces recherches étaient limitées par le faible nombre de séquences hôtes disponibles, notamment l'absence d'un génome disponible de *T. ni*. Les auteurs ont conclu que la diversité des ET sautant dans les génomes viraux ainsi que la fréquence d'un transposon dans 8 500 génomes viraux étaient probablement très sous-estimées.

### e. Flux continu d'éléments transposables dans des populations de baculovirus

Le transcriptome de *T. ni* ayant été publié entre temps (Chen et al. 2014), les auteurs ont réitéré les analyses précédentes en utilisant ce transcriptome comme séquence cible pour la recherche d'ET et autres séquences intégrées dans les populations du même baculovirus (Gilbert et al. 2016). Dans cette seconde étude, ils ont donc réanalysé la population déjà étudiée dans Gilbert et al. (2014) et ont inclus 20 autres populations d'AcMNPV dont 10 issues de séries de 10 cycles successifs d'infections de chenilles de *T. ni* et 10 issues de séries de 10 cycles successifs d'infections de chenilles de la légionnaire de la betterave (*Spodoptera exigua*, Noctuidae, Lepidoptera). Les populations issues de *S. exigua* ont été analysées en utilisant le transcriptome de cette espèce, déjà disponible (Pascual et al. 2012) comme séquence cible. Au total, 86 séquences différentes ont été détectées, responsables de 27 504 insertions dans ce qui équivaut à environ 500 000 génomes viraux. La majorité de ces 86 séquences sont des ET (n = 69) appartenant à 10 superfamilles de transposons à ADN et à 3 superfamilles de rétrotransposons, qui se sont insérés plusieurs fois à plusieurs positions le long du génome viral. Ce jeu de données plus complet a permis de préciser l'estimation de la fréquence de génomes d'AcMNPV portant au moins un ET hôte à au moins 4,8% en moyenne (de 1,1 à 14,3%, selon les populations). Cette fréquence est très intéressante au niveau biologique, car pendant l'infection d'une chenille par ingestion de particules virales, plusieurs dizaines voire plusieurs centaines de milliers de génomes d'AcMNPV sont ingérés par la chenille, donc plusieurs dizaines d'ET d'un hôte précédent aussi sont ingérés. Ceci implique que chaque infection qui ne serait pas létale pour une chenille représente une opportunité de TH entre chenilles pour les ET insérés dans les génomes viraux. Les auteurs ont d'ailleurs pu montrer que 21 ET intégrés dans les génomes d'AcMNPV avaient été transférés horizontalement chez les insectes, dont 5 espèces de lépidoptères connues pour être susceptibles aux baculovirus. Enfin, si toutes les populations de virus analysées portent des séquences hôtes, aucune des insertions d'ET trouvées dans la population initiale d'AcMNPV n'a été retrouvée dans les populations 10 cycles d'infection plus tard. Ce résultat suggère un flux continu d'ET dans les populations virales, couplé à une élimination rapide de ces séquences, potentiellement due à leur caractère généralement délétère.

Ainsi, il semble que les virus soient des agents pouvant effectivement porter au sein de leur génome des ET de l'hôte pendant l'infection. En infectant un second hôte, ces ET pourraient potentiellement à nouveau transposer dans le génome de ce nouvel hôte, et ainsi être vecteur d'un TH d'ET entre animaux. Pour peu que ces insertions aient lieu dans une cellule germinale, ce TH d'ET pourrait potentiellement avoir une importance évolutive impactant de fait le

génome de l'espèce receveuse ainsi que la dynamique évolutive de l'ET ayant subi ce TH (Figure 0.6).

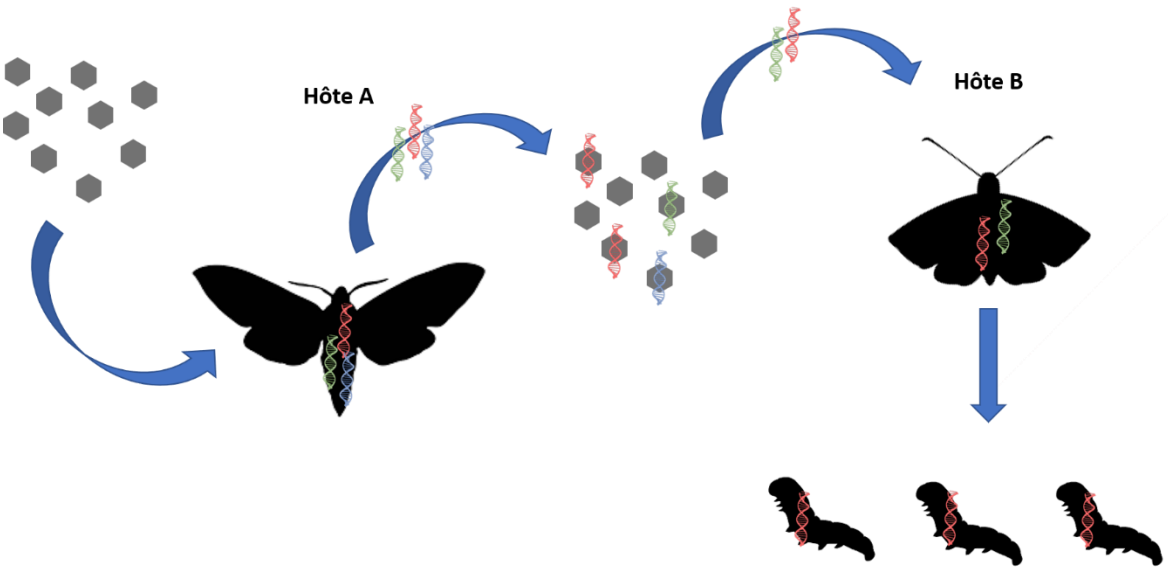



**Figure 0.6: Schéma d'un TH d'ET entre deux espèces différentes de papillons par l'intermédiaire d'une population virale.** La première phase correspond à l'infection d'un papillon par une population virale naïve. Au cours de l'infection, certains ET de l'hôte A transposent du génome de l'hôte aux génomes viraux. Ainsi, la population virale post-infection possède certains génomes viraux avec des ET de l'hôte A. Ce virus peut ensuite infecter un autre hôte et ainsi les ET de l'hôte A pourraient transposer des génomes viraux au génome des cellules infectées de l'hôte B. Si jamais l'hôte survit à l'infection et si les cellules infectées appartiennent à la lignée germinale et sont impliquées dans la reproduction de cet hôte, alors des descendants de l'hôte B pourront porter au sein de leur génome des ET de l'hôte A. Ainsi, la population virale aura été vectrice d'ET entre les deux espèces animales.

### f. Quelques informations concernant le virus AcMNPV

Ce virus a été évoqué précédemment puisque différentes études ont mis en évidence des ET insérés dans des génomes d'AcMNPV. De plus, il a également été utilisé dans plusieurs études au cours de cette thèse. Il apparaît donc nécessaire d'apporter quelques précisions quant à sa biologie.

Ce virus appartient à la famille des *baculoviridae*, qui sont de grands virus à ADN double brin circulaire infectant des invertébrés, en particulier les larves d'insectes. Les virus de cette famille partagent des caractéristiques communes comme une nucléocapside enveloppée en dorme de tige de 30-60nm de diamètre et 25-300 nm de long. La taille du génome varie de 80 à 180 kb (celui d'AcMNPV fait 134 kb environ) et possèdent deux phénotypes différents de virions. La taxonomie des *baculoviridae* a été décrite pour la première fois en 1976 et comprend quatre genres, déterminés selon des caractéristiques phylogénétiques, biologiques et morphologiques.

Le premier genre est appelé *Alphabaculovirus* et inclut tous les nucléopolyhédrovirus spécifiques aux lépidoptères. L'espèce type de ce genre est AcMNPV. Le second genre, les *Betabaculovirus*, incluent les granulovirus dont l'espèce type est *Cydia pomonella* granulovirus (CpGV). Les *Gammabaculovirus* regroupent les baculovirus infectant les hyménoptères alors que les *Deltabaculovirus* correspondent aux baculovirus infectant les diptères. La plupart des *baculoviridae* infectent une seule ou quelques espèces apparentées, mais AcMNPV est connu pour avoir une gamme d'hôtes lépidoptères relativement importante (Rohrmann 2019).

Au cours d'une infection à AcMNPV, deux formes morphologiquement différentes de virions sont produites, chacun ayant des propriétés et une action spécifique dans le cycle de ces virus. Tout d'abord, les larves d'insectes sont infectées lors de l'ingestion de corps d'occlusion, qui sont des corps composés d'une matrice protéique (la polyhédrine) contenant des virus dérivés de l'occlusion (ou ODV pour *occlusion-derived virus*), virions enveloppés présents dans cette matrice protéique. Au cours de l'ingestion, les corps d'occlusion vont parvenir dans l'intestin des larves. Le pH basique s'y trouvant va entraîner la dissolution de ces corps d'occlusion, libérant ainsi les ODV qui vont infecter les cellules épithéliales intestinales. Les cellules infectées vont ensuite produire le second type de virions, des virus bourgeonnant (ou BV pour *budded virus*). Ces BV vont répandre l'infection virale de cellule en cellule au sein de la larve infectée. Chez les *Betabaculovirus* (par exemple CpGV), la matrice protéique est composée non pas de polyhédrine mais de granuline, et chaque ODV ne contient qu'un seul génome viral (Rohrmann 2019).

## IX. Objectifs de la thèse

Au cours de cette thèse, l'objectif était donc de disséquer le processus permettant aux virus d'agir comme vecteurs d'ADN entre différents hôtes. Pour cela, nous nous sommes posé différentes questions afférant à chaque étape d'un TH d'ADN médié par une population virale (figure 0.6).

La première d'entre elles concernait la première étape du TH, à savoir l'infection d'un hôte par un virus. Comment et pourquoi des ET pourraient-ils transposer dans des génomes viraux ? Plus précisément, nous nous sommes demandé si le stress provoqué par une infection virale pouvait induire l'activation des ET de l'hôte infecté, et ainsi leur possible transposition dans les génomes viraux. Aussi, nous avons cherché à savoir si les ET intégrés dans des génomes viraux pouvaient être transcrits, condition indispensable pour que ces ET puissent ensuite transposer des génomes viraux au génome d'un autre hôte. Une étude menée sur ce sujet par analyse de données transcriptomiques tente d'apporter des réponses à ces questions et constitue le premier chapitre de cette thèse.

Un autre point important d'un TH d'ET par l'intermédiaire d'un virus et pour lequel les données dans la littérature sont assez peu fournies concerne la diversité et la quantité des ET que l'on peut trouver intégrés dans des génomes viraux. En effet, mis à part les quelques systèmes papillon-baculovirus décrits précédemment, aucun autre système hôte-virus n'a à ce jour été étudié par séquençage haut débit du point de vue de l'insertion d'ET dans les génomes viraux. Nous avons ainsi cherché à étendre le spectre des connaissances en étudiant une diversité de systèmes hôtes-virus pour mieux cerner cette dynamique. Ci-après sont énoncées quelques-unes des questions que nous nous sommes posées : observe-t-on un biais en termes de type d'ET inséré dans les populations virales ? La fréquence d'insertion est-elle la même d'un système à l'autre ? Si ce n'est pas le cas, à quoi cela est-il dû ? Les ET peuvent-ils transposer de génomes viraux à génomes viraux ? Les ET portés par des génomes viraux peuvent-ils persister pendant plus d'un cycle d'infection au sein d'une population virale ? Ces différentes questions font l'objet du deuxième chapitre de la thèse.

Bien que l'insertion d'ET dans des génomes viraux ait précédemment été montrée comme nous l'avons vu en introduction, la détection de ces insertions a été le plus souvent basée sur l'analyse de lectures courtes de séquençage (Illumina). Ce faisant, l'état et la structure de la plupart des ET intégrés dans les génomes viraux restent à être précisés. Ce point est important, car, pour

qu'un ET subisse un TH d'un animal à un autre, il est important que sa séquence soit insérée complètement dans le génome viral pour pouvoir ensuite être capable de coder des protéines fonctionnelles, et transposer à nouveau lors de l'infection d'un nouvel hôte par le virus. Nous avons de ce fait cherché à combler ce vide par l'analyse de longues lectures, générées grâce à la technologie PacBio, capables de couvrir l'insertion d'un ET sur toute la longueur de sa séquence, et pas seulement ses extrémités. La richesse des données générées dans cette étude nous a amenés à nous intéresser non pas seulement aux insertions d'ET dans les génomes viraux, mais à l'ensemble des variations structurales présentes au sein des génomes d'une population virale, apportant à nouveau des résultats jamais mis au jour jusqu'à présent chez des virus. Ces données représentaient une opportunité unique de mesurer la variabilité de ces variants génomiques structuraux présente au sein de populations virales. Ceci est traité dans le troisième chapitre.

Enfin, bien que dépassant à proprement parler le propos du sujet de cette thèse, nous nous sommes intéressés au TH d'un rétrovirus murin dans une lignée cellulaire humaine. Cette étude s'est présentée par l'opportunité d'analyser plusieurs jeux de données (génomiques et transcriptomiques) afin de mettre en avant de nouveaux pathogènes affectant des cultures cellulaires humaines, dans le cadre d'une collaboration avec le laboratoire de virologie du Centre Hospitalier Universitaire de la ville de Poitiers. Au cours de cette étude, aucun nouveau pathogène n'a pu être détecté, mais un TH de rétrovirus murin dans la lignée cellulaire Hep2 clone 2B a été révélé et caractérisé en détail. Le sujet de la thèse s'inscrivant plus largement dans la thématique des TH chez les animaux, cette étude nous a semblé pertinente et constituera ainsi le quatrième et dernier chapitre de cette thèse.

# Chapitre 1

# *Chapitre 1 : Expression des éléments transposables chez la fausse-arpenteuse du chou* Trichoplusia ni *lors d'une infection à AcMNPV*

L'étude de l'influence du stress sur l'activité des éléments transposables (ET), éléments génétiques mobiles égoïstes présents dans les génomes, nous a semblé pertinente dans le cadre de la thématique des transferts horizontaux (TH) d'ADN entre métazoaires, effectués par l'intermédiaire d'un virus. La première étape de ce TH consiste en l'infection d'un hôte par un virus. Si le stress provoqué par l'infection virale entraîne une augmentation de l'activité des ET, alors ces derniers auraient une probabilité plus importante de s'intégrer au sein des génomes viraux, renforçant ainsi l'hypothèse des virus comme vecteurs d'ADN entre animaux.

Pour tester cette hypothèse, nous avons utilisé les données transcriptomiques d'un virus (le baculovirus *Autographa californica* multiple nucleopolyhedro virus ou AcMNPV) et de son hôte la fausse-arpenteuse du chou (*Trichoplusia ni*) produites par Chen et al. (2013) et Shrestha et al. (2018). La première étude visait à analyser l'expression des gènes d'AcMNPV lors de l'infection d'une lignée cellulaire de *T. ni* (Tnms42). L'autre visait à étudier l'expression des gènes de ce même virus dans l'épithélium intestinal de larves de *T. ni* infectées par AcMNPV. L'étude de l'expression des ET dans ces deux ensembles de jeux de données nous a permis de montrer que la plupart des familles d'ET n'étaient pas exprimées tout au long de l'infection virale. Seules quelques familles d'ET apparaissent surexprimées (13 familles d'ET dans l'intestin moyen des larves et 30 dans les cellules Tnms42), principalement l'ET non autonome TFP3, appartenant à la superfamille piggybac. Il a été trouvé surexprimé en la lignée cellulaire. De plus, nous avons trouvé 11 TE insérés dans les génomes AcMNPV qui ont été cotranscrits avec des gènes viraux après leur insertion, TFP3 étant le plus cotranscrit. Cette étude met en évidence une surexpression spécifique à certains ET au cours d'une infection virale, mais pas de dérépression globale des ET, et elle montre que les ET hôtes insérés dans les génomes viraux peuvent être transcrits, ce qui soutient l'hypothèse selon laquelle les virus pourraient jouer le rôle de vecteurs de transfert horizontal d'éléments transposables entre insectes. De façon plus

générale, elle vient compléter une littérature encore peu fournie concernant l'influence des stress biotiques sur l'activité des ET chez les animaux.

Cette étude a donné lieu à un article dont je suis le premier auteur et qui est destiné à être soumis pour publication dans *PeerJ*.

**Assessing the impact of a viral infection on the expression of transposable elements in the cabbage looper moth (*Trichoplusia ni*)**

Vincent Loiseau[1], Sandra Guillier[1], Richard Cordaux[2], Clément Gilbert[1]

[1] Laboratoire Evolution, Génomes, Comportement, Écologie, Unité Mixte de Recherche 9191 Centre National de la Recherche Scientifique et Unité Mixte de Recherche 247 Institut de Recherche pour le Développement, Université Paris-Sud, 91198, Gif-sur-Yvette, France.

[2] Université de Poitiers, Laboratoire Ecologie et Biologie des Interactions, Equipe Ecologie Evolution Symbiose, 5 Rue Albert Turpain, TSA 51106, 86073, Poitiers Cedex 9, France.

**Summary**

Transposable elements (TEs) are selfish genetic elements able to move in host genomes. They account for a large part of eukaryotic genomes. Due to their deleterious effect during transposition, most of them are repressed. Stress can lead to TE de-repression due to epigenetic modifications or transcription factor activation. However, the relation between TE expression and stress is not so direct, some stresses leading to up- or downregulation. Here we analyzed the cabbage looper larvae (*Trichoplusia ni*) migduts and Tnms42 cells (a *T. ni* cell line) TE expression facing a viral infection by Autographa californica multiple nucleopolyhedrovirus (AcMNPV). We found most TE families were repressed all along the viral infection. Only specific TE families were upregulated (13 TE families in midgut larvae and 30 in Tnms42 cells), mainly the TFP3 non-autonomous TE, belonging to the piggybac superfamily. It was found upregulated in cell line. Moreover, we found 11 TEs inserted into AcMNPV genomes that were transcribed after their insertion, TFP3 being the most transcribed, at the level of some AcMNPV genes. Finally, we investigated if there are some links between TE expression during a viral infection in living animals and the TE landscape of the host genome. We found no correlation between TE copy number, TE copy divergence to consensus or TE distance to gene and TE expression, likely due to the very low expression level of most TE families. This study highlights specific TE upregulation during a viral infection but no global unleashing of TEs, and it shows that host TEs inserted in viral genomes can be transcribed, which further supports viruses as potential vectors of horizontal transfer of transposable element between insects.

## Introduction

Transposable elements (TEs) are selfish genetic elements able to move in the genome of their hosts. They were first discovered in maize (McClintock 1950). They account for a large fraction of eukaryotic genomes (Schnable et al. 2009; Sotero-Caio et al. 2017). Based on their ability to transpose, TEs are classified into two categories: TEs that move through an RNA intermediate are class I TEs, and those moving through a DNA intermediate are class II TEs (Wicker et al. 2007). The raw genetic material deposited by each new transposition event has sometimes been recycled during evolution, fueling genomic novelty and adaptation (I. R. Arkhipova 2018; Bourque et al. 2018). While domestication of many TE-coding sequences has been reported (Volff 2006), most co-option events involve TE regulatory sequences, which have sometimes led to profound changes into expression landscapes (Chuong, Elde, and Feschotte 2017). However, like many other mutation types, most transposition events are neutral or harmful and are thought to negatively impact the host fitness (Barrón et al. 2014; Mita and Boeke 2016; Brookfield and Badge 1997). In response to the deleterious effects of TEs, several TE-repressing mechanisms have evolved in host genomes, such as DNA methylation, histone modifications or post-transcriptional repression through the PIWI-interacting RNA pathway (Slotkin and Martienssen 2007; Deniz, Frost, and Branco 2019). Thus, host-TE interactions are best described as an evolutionary arms race, which often lead to complete extinction of TE families and degradation of TE copies that are purged from the genome with time, mainly due to neutral evolution of most TE sequences (Le Rouzic, Boutin, and Capy 2007; Blumenstiel 2019).

Typically, few TE families are active in a genome, most of them being repressed and thus not expressed (Yoder et al. 1997; Zilberman et al. 2007). However, a perturbation of genomic stability like environmental changes or infections leading to a stress can modify the expression state of TEs (Miousse et al. 2015). Several examples of TE de-repression due to environmental stress have been reported in plants, and this phenomenon appears to also occur in other eukaryotes such as yeasts, human and other mammals, insects and nematodes (Menees and Sandmeyer 1996, Van Meter et al. 2014; Voronova et al. 2014; Romero-Soriano and Garcia Guerreiro 2016; Ryan, Brownlie, and Whyard 2017; Zovoilis et al. 2016; Huang et al. 2017; Hummel et al. 2017, Dubin, Mittelsten Scheid, and Becker 2018). Such de-repression is thought to often be caused by epigenetic modifications or activation of transcription factors (Capy et al. 2000; Horváth, Merenciano, and González 2017). Interestingly, some TEs even bear a stress response element, i.e., a regulatory sequence activated in response to a stress, enabling TEs to

be upregulated in stressful conditions (Bucher, Reinders, and Mirouze 2012; Casacuberta and González 2013). However, the impact of stress on TE expression appears hardly predictable. For instance, studies of stress-induced TE-expression in *Drosophila* have shown that depending on cases, TEs can be upregulated, downregulated or transiently upregulated before being downregulated in response to a stress (Horváth, Merenciano, and González 2017). The complexity of the interplay between stress and TE expression is likely due to several factors. First, the impact of stress on transcription varies along the genome, being seemingly higher in facultative heterochromatin, which is generally gene-rich and poorer in TEs than constitutive heterochromatin, which is generally associated with gene-poor, TE-rich regions (Trojer and Reinberg 2007; Saksouk, Simboeck, and Déjardin 2015). Consistently, the distribution of a TE family along the genome is often highly correlated to chromatin state (Lanciano and Mirouze 2018). Moreover, stress-induced TE activation can generate new copies in the genome via transposition. These new copies can bear *cis*-regulatory elements that can contribute to rewire the stress response network, in turn modulating the interaction between stress and TE expression during a stress (Cowley and Oakey 2013; Galindo-González et al. 2017). Finally, the epigenetic landscape influencing TE repression is variable between closely related species and even between populations of a same species (Barah et al. 2013; Niederhuth et al. 2016; Fouché et al. 2020).

In the study of eukaryotic TE response to stress, most of the effort focused on stress in plants. To our knowledge, few studies have investigated the impact of a biotic stress like a viral infection on the TE expression in animals. A recent study reanalyzed transcriptomic data of several human and mouse cell lines infected by various viruses and found a genome-wide TE upregulation in host cells (Macchietto, Langlois, and Shen 2020). This pattern was observed particularly near antiviral gene responses and was common to analyzed datasets, whatever the kind of virus, the host or the cell type studied. The authors concluded that TE upregulation during a viral infection could be a common phenomenon in human and mouse. A second study analyzed the impact of Sindbis virus (SINV), a single-stranded RNA virus, on *Drosophila simulans* and *D. melanogaster* flies (Roy et al. 2020). They found viral infection can modulate the piRNA and siRNA repertoires, pathways known to be involved in the TE expression control. For instance, a global decrease of TE transcript amounts was observed in *D. simulans* and *D. melanogaster* flies during the exponential phase of SINV replication. On the contrary, no difference in TE transcript amounts was observed in *D. simulans* ovaries. TE activity was

sensitive to SINV infection that may affect TE mobilization rates. Overall, these studies suggest that viral infection impacts TE activity in animals.

Interestingly, several other studies reporting host TEs integrated in baculovirus genomes provide direct evidence that some TEs are active during infection (Fraser et al. 1985; Jehle et al. 1998; Gilbert et al. 2014; 2016; Loiseau et al. 2020). For example, Gilbert et al. (2016) found thousands of TE copies belonging to 13 TE superfamilies integrated in the genome of the AcMNPV baculovirus after infection of noctuid moth larvae. They estimated that in these viral populations, 4.8% of AcMNPV genomes, on average, carried at least one host TE. Furthermore, long read sequencing revealed that many of the TE copies were integrated in AcMNPV genomes as full-length copies, bearing all the components necessary to transpose (Loiseau et al. 2020). These studies clearly revealed that many Class I and Class 2 TEs are expressed and capable of actively transposing during infection by the AcMNPV baculovirus. They also raised several questions regarding the possible interaction between AcMNPV and host TEs. First, host TE expression has never been measured during infection by large dsDNA viruses. Thus, it is unknown whether the TEs found in viral genomes are expressed in the host genome in normal, non-infected conditions, or whether they are normally repressed but become activated or overexpressed in infected hosts. Whether TEs found in viral genomes during an infection are also those that are the most highly expressed in the host genome is also unknown. Furthermore, the influence of factors such as TE age, TE copy number and location in the host genome on the level of host TE expression remains unclear. Finally, whether TE copies integrated into viral genome are expressed during infection has never been measured.

We addressed these questions by reanalyzing RNA-seq datasets produced by Chen et al. (2013) and Shrestha et al. (2018), in which the expression of host genes was measured in Tnms42 cells (a *Trichoplusia ni* cell line) and *Trichoplusia ni* fifth instar larvae infected by AcMNPV, respectively. Viral infections were monitored from 0 hours post infection (hpi) to 48 hpi and 72 hpi, respectively. In both datasets, we found few differentially expressed (DE) TEs. None of the *T. ni* TEs previously found inserted in AcMNPV genomes were particularly overexpressed during the infection in these experiments. Finally, we found no correlation between TE copy number, TE copy divergence to consensus or TE distance to gene with TE expression, mostly due to the very low expression level of most TE families. Interestingly, however, we were able to measure the expression of some TEs inserted in AcMNPV genomes in the Tnms42 cells RNA-seq dataset. The most highly expressed TE (TFP3, piggybac) was expressed at levels

similar to those of many AcMNPV genes. This TE was also one of the most upregulated TEs during infection, due to inserted copies into AcMNPV genomes. This study highlights specific TE upregulation during a viral infection but no global unleashing of TEs, and it shows that TEs inserted in viral genomes can be transcribed, which further supports viruses as vectors of horizontal transfer of transposable element between insects.

## Materials & Methods

### RNA-seq data of Tnms42 cells infected by AcMNPV

RNA-seq data were retrieved from Chen et al. (2013) (Sequence Read Archive [SRA] accession number SRA057390). Briefly, *T. ni* cells from the Tnms42 cell line, which derives from HighFive cells, were infected with the wild-type AcMNPV strain E2 (Chen et al. 2013). For infections, $3x10^6$ Tnms42 were infected at a multiplicity of infection (MOI) of 10 in a T25 flask. After a 1-hour incubation, the inoculum was removed and the cells were rinsed with Grace's medium and cultured in TNM-FH medium supplemented with 10% FBS at 28°C. The time at which the inoculum was removed was designated 0 hpi. Total RNA was isolated from AcMNPV-infected cells, as well as from a set of parallel control cells (uninfected or mock infected), at 0, 6, 12, 18, 24, 36 and 48 hpi using a Qiagen RNeasy minikit. Polyadenylated RNA isolated from 20g total RNA using Dynabeads oligo$(dT)_{25}$ (Invitrogen) was used for sequencing. The sequencing library was constructed with the TruSeq protocol and sequencing was performed on an Illumina HiSeq2000 platform. Single-end reads of 101-bp long were produced. Further information can be found in Chen et al. (2013). Please note that reads corresponding to negative control at 24 hpi cannot be retrieved from the SRA.

### RNA-seq data of midgut *T. ni* larvae infected by AcMNPV

RNA-seq data were also retrieved from Shrestha et al. (2018) (SRA accession number PRJNA484772). In this study, *T. ni* fourth-instar larvae (Cornell strain) that were ready to molt were held for 0 to 5 h without diet, and newly molted 5th-instar larvae (0 to 5 h old) were used for oral infections. Larvae were orally inoculated with 5 µl of a 10% sucrose solution containing a total of $7x10^4$ occlusion bodies of wild type AcMNPV strain E2 (as in Chen et al. 2013). Mock-infected control larvae were fed a similar sucrose solution containing no virus. Midgut tissue was dissected at eight time points post infection: 0, 6, 12, 18, 24, 36, 48, and 72 hpi. For each time point sampled post infection, a parallel mock-infected control midgut sample was

dissected, to mitigate possible artifacts resulting from developmental changes that may occur over the course of the experiment. For each time point and treatment (infected or control), three replicate samples were prepared, with midgut samples from six larvae pooled for each replicate. Total RNA extraction was performed on pooled midgut samples with the TRIzol reagent (Ambion). Poly(A) mRNAs isolated from 3g of total RNA using oligo$(dT)_{25}$ Dynabeads (Invitrogen) were used for sequencing. The library was constructed with the TruSeq protocol. Sequencing was performed on an Illumina HiSeq4000 platform. Single-end reads of 51-bp were generated. Further information is provided in Shrestha *et al.* (2018).

**_T.ni_ genomes used in TE differential expression analyses**

Two *T. ni* genome assemblies were retrieved from GenBank: (i) one derived from a single male *T. ni* larva (accession number PPHH01000000; Chen et al., 2019), and (ii) one derived from the *T. ni* Hi5 germ cell line (accession number NKQN00000000; Fu et al. 2018). The first genome was used to map the *in vivo* RNA-seq reads. As the cell line infected with AcMNPV (Tnms42) is derived from the Hi5 cell line (Chen et al., 2013), we mapped the cell line RNA-seq reads on the second genome.

**TE identification and database**

The TE library that we used to annotate TEs in *T. ni* genomes was compiled as follows. First, RepeatModeler version 1.0.11 (http://www.repeatmasker.org) was run with default options on the *in vivo T. ni* genome, which allowed us to identify 567 TE consensus sequences. In addition, 458 TE consensus sequences of the *T. ni* Hi5 genome were retrieved on https://cabbagelooper.org/. We also added to our TE library 94 *T. ni* TEs previously found inserted in viral genomes (Wang and Fraser 1993; Fraser et al. 1995; Gilbert et al. 2016). Finally, we annotated TEs that we could identify in the RNA-seq data. The 48 datasets produced by Shrestha et al. (2018) were assembled with Trinity version 2.1.1 (Grabherr et al. 2011). The resulting 45,094 contigs were then mapped onto the AcMNPV strain E2 genome (GenBank accession number KM667940.1), which led us to remove 45 viral contigs. RepeatModeler version 1.0.11 was then run on the remaining contigs, which yielded 183 TE families. We also aligned the 45,049 non-viral contigs on a library of TE proteins ("RepeatPeps") provided in the RepeatModeler package using diamond (Buchfink et al. 2015, options: ‘diamond blastx -more-sensitive’). We retained 151 contigs which aligned over at least half of a TE protein. The same approach was applied to the RNA-seq datasets from Chen et al. (2013). After Trinity assembly, we found 103,650 non-viral out of 103,790. Among them, 472 TE families were identified by

RepeatModeler and 612 by alignment on the RepeatPeps library. A total of 2,535 TE sequences were retrieved. Clustering of these sequences using Vsearch (options used: '--target_cov 80.0 --query_cov 80.0 –id 0.95') (Rognes et al. 2016) revealed that they were all unique. Finally, to remove TE sequences for which a robust annotation could not be achieved, we aligned the 2,535 TE sequences on the RepeatPeps library and kept only TEs being >300 bp in length and aligning on at least half of a TE protein. All sequences identified as 'SINE','tRNA','rRNA' or 'Unknown' were discarded. Our final TE library containing 849 TE families was used to annotate TE copies in the two *T. ni* genomes using RepeatMasker version 4.0.7 (http://www.repeatmasker.org). Only TE copies >300 bp in length and aligning on at least >80% of the length of a TE consensus were retained for downstream analyses. This filter led us to consider 410 and 461 TE families for the larvae midgut and cultured cells data, respectively.

**TE mapping with the TEtools pipeline**

The RNA-seq data were trimmed using Trimmomatic version 0.38 (Bolger, Lohse, and Usadel 2014) to remove adaptors and low-quality bases. After trimming, reads <40 bp in length were discarded (command line used: java -jar trimmomatic-0.38.jar SE -threads 30 -phred33 reads_R1.fastq reads_R1_TRIMMED.fastq ILLUMINACLIP:TruSeq2-3-SE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:40).

Given the relatively short length of RNA-seq reads (51 or 101 bp), it seemed difficult to study the TE expression at the copy level. We thus chose to study TE expression at the family level, by mapping reads on TE consensus sequences. Due to the 'very-sensitive' option used for the mapping with TEtools, some mismatches between reads and consensus sequences did not prevent the alignment. Given the option used for mapping, reads could align on the consensus sequences, especially for TE copies being highly similar to their consensus sequence. These copies are the most relevant ones, as they are more likely to be active, contrary to more divergent copies, likely non-functional. The minimum criteria for a read to align on a TE sequence with the 'very-sensitive' option is at least an alignment of 20 bp substring without any mismatch, with a six bp interval. It corresponds to six and 14 20 bp-substrings for a read of 51 or 101 bp, respectively. These substrings have to align at least once on a TE sequence of at least 300 bp, which seems sensitive, even for divergent TE copies to consensus sequence.

The RNA-seq data were mapped to the TE library with the TEtools pipeline version 1.0.0 (Lerat et al. 2016). Reads were aligned to our TE library using Bowtie2 v2.2.4 (Langmead and Salzberg 2012), with the most sensitive option and keeping a single alignment for reads mapping to multiple positions (–very-sensitive for Bowtie2). Read counts were then computed

per TE family (410 and 461 TE families in the larvae midgut and cultured cells data, respectively).

**Differential expression analysis with DESeq2**

We performed the DE analyses for all TE families using the R Bioconductor package DESeq2 (Love, Huber, and Anders 2014) on the raw read counts, using the Benjamini–Hochberg multi ple test correction (FDR level of 0.05, Benjamini and Hochberg, 1995). A TE family was cons idered differentially expressed between samples when the adjusted p-value was <0.05. TE fam ilies with at least 2-fold expression differences between conditions were considered. All analy ses were performed using R version 3.6.2 (R Core Team 2019, https://www.R-project.org/).

**RPKM computation**

Based on read count outputs of TEtools, TE family expression derived from different samples was estimated and normalized to RPKM (reads per kilobase of exon model per million mapped reads, Mortazavi et al. 2008) to generate an expression unit enabling the comparison with gene expression computed in by Chen et al. (2013) and Shresta et al. (2018).

**Detection of TE/virus junctions in transcriptomic data**

In addition to the DE analysis, we measured the expression of TEs integrated into viral genomes. For this, we identified RNA-seq reads carrying a junction between a moth TE sequence and the AcMNPV genome. Such chimeric reads correspond to portions of transcripts that start into a viral gene and continue in a TE sequence integrated in the viral genome. This approach allowed us to make sure host transcripts had been excluded. To identify chimeric reads, all reads were aligned to the AcMNPV WP10 genome (GenBank accession number KM609482) and to a TE library including TEs available in RepBase as of March 2018, TEs identified by Gilbert et al. (2016) and Walsh et al. (2013). Analyses used to identify chimeric reads was performed on R (R Core team, 2019). This pipeline was developed by Gilbert et al. (2016). Briefly, reads are aligned separately on host sequences and the viral genome using blastn (-task megablast). Chimeric reads for which a portion aligns on a host sequence *only* and the other portion aligns on the viral genome *only* are then identified based on alignment coordinates. This approach was also used in Loiseau et al. (2020) to identify TEs from *Spodoptera exigua* integrated in genomes of another AcMNPV population purified from *S. exigua* larvae, as well as by Peccoud et al. (2018) to characterize artificial chimeras generated during the construction of sequencing libraries.

**Identification of target site duplications**

To confirm host TE insertions in viral genomes, we searched for target site duplications (TSD), that are a signature of canonical transposition. We separated chimeric reads in 5' of a TE sequence from those in 3'. To be sure reads in 5' and 3' corresponded to the same insertion, we used different criteria. The viral insertion coordinate had to be equal to more or less 5 bp between the 5' and 3' chimeric reads. The same TE had to be detected at this insertion point. The 5' and 3' chimeric reads had to have a concordant orientation (if the left part of some reads corresponded to the viral sequence, the right part of the other reads had to correspond to the viral sequence). Ten nucleotide sequences upstream or downstream the TE insertion point were considered for each read and TSD were identified by building sequence logos (Wagih 2017).

**Results and discussion**

**Genome-wide TE differential expression during AcMNPV infection of Tnms42 cells**

Reads produced by Chen et al. (2013) were mapped on our TE library containing 461 TE consensus sequences, including 245 DNAs, 120 LINEs, 90 LTRs and 6 Helitrons. Among these, 283 were considered as never expressed (RPKM <1 at 0 and 48 hpi, without consideration of expression at intermediate time points) and 46 as always expressed (RPKM >1 at 0 and 48 hpi, without consideration of expression at intermediate time points). A total of 66 TE families were found to exhibit DE during the course of the AcMNPV infection in *T. ni* cells (Figure 1.1). Among them, 36 were upregulated, including 33 that were activated (RPKM<1 at 0 hpi), and 30 were downregulated, including 22 that were silenced (RPKM<1 at 48 hpi). We found that half of the Helitron families present in the *T. ni* genome (3 out of 6) are DE, compared to only 11%, 19%, and 14% of DNA, LINE and LTR TEs (Table 1.1). Although no TE class seemed to be strikingly either up- or downregulated, at the superfamily level, 4/4 DE Maverick families are upregulated, 6/7 TcMar-m4 families and 4/5 piggyBac families are downregulated. We also detected 10 DE L2 families, with five down- and five upregulated (Table 1.1). These results suggested that some TE superfamilies may be more prone to be overexpressed and others to be underexpressed during a viral infection.

The most upregulated TEs were an RTE and an L2 non-LTR retrotransposons, as well as a TFP3 DNA TE, with a log2FoldChange >8 at 48 hpi (i.e. $2^8$= 256-fold more expressed

compared to the uninfected condition). However, among those three, only TFP3 reached a high expression level (684 RPKM at 48 hpi, Figure 1.1). This TE is a 831 bp-long non-autonomous TE belonging to the piggyBac superfamily. It was first discovered inserted in AcMNPV genomes purified from *T. ni* (TN-368) cells (Fraser et al 1983; Wang and Fraser 1993). TFP3 was strongly overexpressed in infected Tnms42 cells, whatever the time point. It is relevant that TFP3 is the only upregulated TE reaching an expression level > 10 RPKM at 48 hpi (Figure 1.1). Thus, TEs upregulated in the infected condition are generally weakly expressed and their relative overexpression (corresponding to the log2FoldChange value) is due to a small absolute increase in expression (revealed by the RPKM level).

The most downregulated TEs were a piggyBac DNA TE and a Gypsy LTR-retrotransposon, with a log2FoldChange <-4 at 48 hpi (i.e. $2^4$=16 times less expressed compared to normal condition, Figure 1.1). None of the underexpressed TEs have RPKM levels >10 at 48 hpi. One interesting TE is the downregulated TE piggyBac_1 family that shows the highest expression level at 0 hpi (58 RPKM) among underexpressed TEs and drops to 5 RPKM at 48 hpi, which is the strongest decrease. This TE is the only DE TE with TFP3 (that is upregulated contrary to this piggyBac TE) to have been found integrated into AcMNPV genomes (Gilbert et al. 2016; Fraser et al. 1983).

Thirty *T. ni* TEs previously found integrated in AcMNPV genomes were found in the *T. ni* Hi5 genome, including the two DE piggyBac described above. Among the 28 other TEs, 23 TEs were not (or nearly so) expressed (expression level < 1 RPKM at 0 and 48 hpi). The five other *T. ni* TEs previously found inserted in AcMNPV genomes showed a decrease of their expression level between 0 and 48 hpi, although such decreases were not significant compared to the control data.

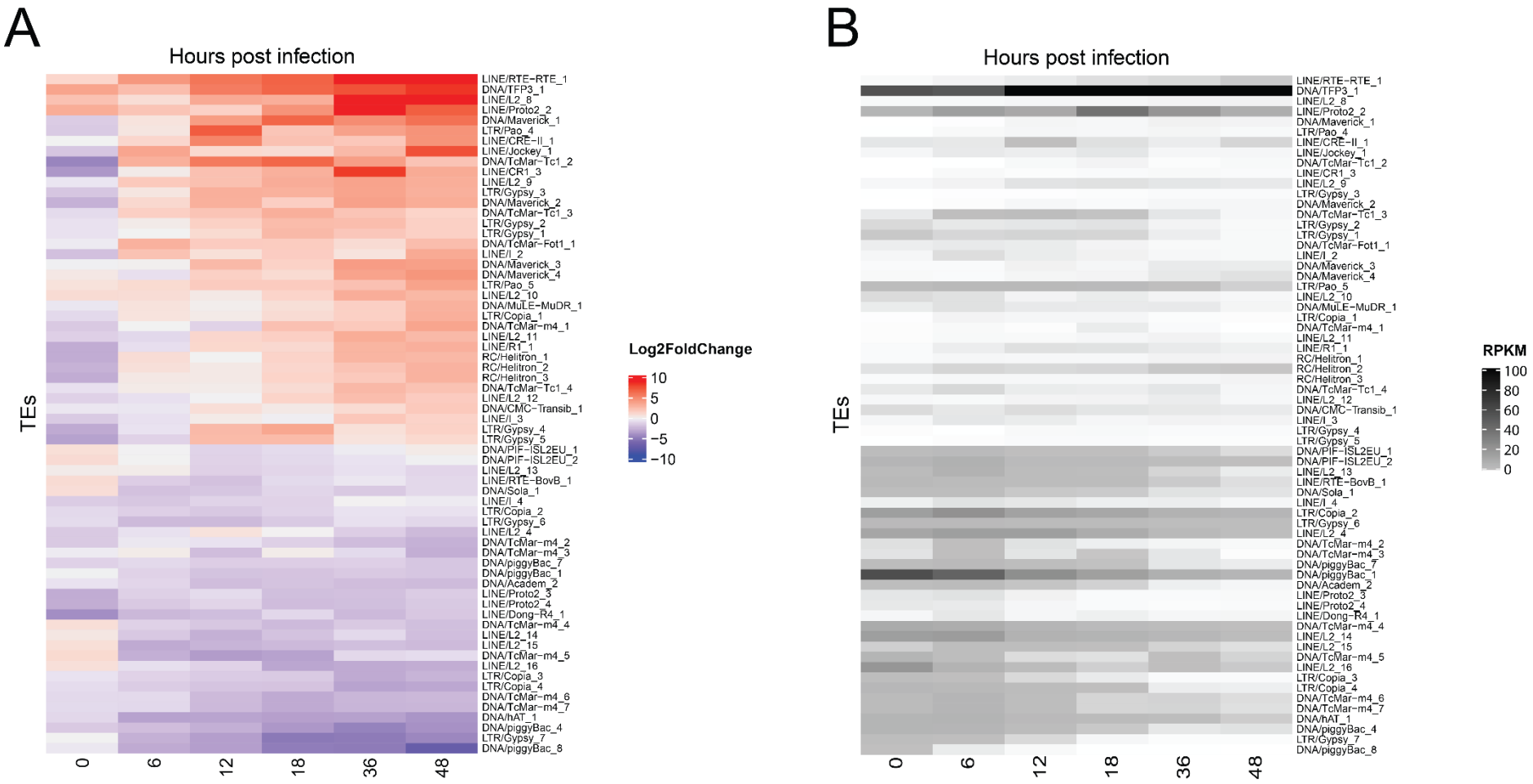



**Figure 1.1: 66 differentially expressed TE families in the cell line data.** Heatmaps represent the log2 fold change (A) of the TE expression and the absolute expression expressed in RPKM (B) during the infection. The differential expression was computed by comparison between virus-infected and mock-infected data. A correspondence of TE names between those on the heatmaps and those in the fasta file is available in Table S1.1.

**Table 1.1: Information about the 66 DE TEs in cell line data.**

| TE expression state during infection by AcMNPV | Number of TE families by TE class | Number of TE families by TE superfamily |
|---|---|---|
| Upregulated | 12 DNA | 1 piggyBac |
| | | 1 MuLE-MuDR |
| | | 1 CMC-Transib |
| | | 1 TcMar-m4 |
| | | 1 TcMar-fot1 |
| | | 3 TcMar-Tc1 |
| | | 4 Maverick |
| | 3 RC | 3 Helitron |
| | 13 LINE | 1 R1 |
| | | 1 RTE-RTE |
| | | 1 Proto2 |
| | | 1 CRE II |
| | | 1 CR1 |
| | | 1 Jockey |
| | | 2 I |
| | | 5 L2 |
| | 8 LTR | 1 Copia |
| | | 2 Pao |
| | | 5 Gypsy |
| Downregulated | 15 DNA | 1 Sola |
| | | 1 Academ |
| | | 1 hAT |
| | | 2 PIF-ISL2EU |
| | | 4 piggyBac |
| | | 6 TcMar-m4 |
| | 10 LINE | 1 Dong-r4 |
| | | 1 I |
| | | 1 RTE-BovB |
| | | 2 Proto2 |
| | | 5 L2 |
| | 5 LTR | 2 Gypsy |
| | | 3 Copia |

Among the 461 TE families annotated in the *T. ni* Hi5 genome, 326 (71%) were not expressed (RPKM<1) and 106 (23%) were weakly expressed (1<RPKM<10) at 0 hpi. At 48 hpi, 404 (88%) TE families were not expressed and 50 (11%) were weakly expressed. Thus, the infection of *T. ni* cells by AcMNPV tended to be associated with an overall decrease in TE expression. This result does not support a scenario whereby the integration of host TEs in AcMNPV genomes during the course of an infection (Fraser et al. 1983; Jehle et al. 1998; Gilbert et al. 2014; 2016; Loiseau et al. 2020) would be due to a general unleashing of TE expression.

**Genome-wide TE differential expression during AcMNPV infection of *T. ni* larvae midguts**

The reads produced by Shrestha et al. (2018) were mapped on our TE library containing 410 TE consensus sequences, including 215 DNAs, 104 LINEs, 85 LTRs and 4 Helitrons. Among these, 250 were considered as never expressed (RPKM <1 at 0 and 72 hpi, without consideration of expression at intermediate time points) and 89 as always expressed (RPKM >1 at 0 and 72 hpi, without consideration of expression at intermediate time points). A total of 27 TE families were found to be DE during the course of the AcMNPV infection in *T. ni* larvae (Figure 1.2). Among them, 13 were upregulated, including six that were activated (RPKM<1 at 0 hpi), and 14 were downregulated, including seven that were silenced (RPKM<1 at 72 hpi). DE TE families comprised 10 DNAs, 11 LINEs and 6 LTR retrotransposons. The 13 TE families being upregulated were four DNAs (one Academ, one PIF-Harbinger and two piggyBac), four LINEs (one Proto2, one CR1, one L2 and one I) and five LTR retrotransposons (one Copia, two Pao and two Gypsy). The 14 downregulated TEs were six DNAs (one PIF-Harbinger, one TcMar-Tc1 and four piggyBac), seven LINEs (one CR1 and six L2) and one LTR retrotransposon (one Pao). Although 14% of TE families identified into *T. ni* Hi5 genome were found to be DE in cell line data (66/461 TE families), only 6.6% (27/410 TE families) detected in the *T.ni in vivo* genome were DE in the midgut larvae. This result is consistent with what is known about cell lines, as they have fewer constraints and generally undergo many chromatin remodeling and chromosomal rearrangements, as is known for the *T.ni* Hi5 cell line (Fu et al. 2018). Such modifications could lead to higher TE activity in cell lines, in particular during stress conditions, as protection pathways like the immune system are less efficient, in part due to the presence of a unique cell type (ovarian germ cells, in the case of *T. ni* Hi5 cells; Granados et 1986; Granados et al. 1994).

It is noteworthy that about half of DE TEs were upregulated and half were downregulated. However, LTRs were the major TE class being upregulated, with a single family (out of 6 DE LTRs) being downregulated. At the TE superfamily level, one L2 was upregulated while six L2 were downregulated. These results show different patterns of expression during an infection among different TE groups. LTRs are seemingly more prone to overexpression. On the contrary, TEs of the L2 superfamily are seemingly more prone to underexpression. A notable difference between the cell line and the *in vivo* data is a majority of L2 families (6/7) being downregulated *in vivo*, whereas half of L2 families (5/10) are up- or downregulated in the cell line. However, the number of DE TE families is too small to perform meaningful statistical tests, precluding any robust conclusion to be drawn.

The most upregulated TE family was the Proto2_1 family (Figure 1.2) with a log2FoldChange of 7.8 at 72 hpi (or $2^{7.8}$=222-fold increase in expression). Significant log2FoldChanges were identified for this TE at 24, 36, 48 and 72 hpi. Although it remained expressed at <0.2 RPKM in the uninfected condition, it goes from 0.1 RPKM at 0 hpi to 48 RPKM at 72 hpi in the infected condition (Figure 1.2). The TcMar-Tc1_1 family had a significant expression at 72 hpi with a log2FoldChange of -4.2 ($-2^{4.2}$= 18 times less expressed than in control condition). However, the absolute expression level was very low in both cases (0.7 RPKM in control condition versus 0.04 RPKM in the infected condition at 72 hpi). We noticed that some TE families reached a peak in their expression variation (either up- or downregulated) during the infection at 12/18 hpi, such as Gypsy_1, Gypsy_2 or PIF-Harbinger_2 families (Figure 1.2). This is reminiscent of what was also observed in the cell line analysis (e.g. Gypsy_4, Gypsy_5, TcMar-Tc1_2 or TcMar-m4_5 at 12/18 hpi in Figure 1.1). It is tempting to interpret these results in regard of host or viral gene expression variation during infection (Chen et al. 2013; 2014; Shrestha et al. 2018; 2019). However, one drawback of our study is our inability to infer TE expression at the copy level. As TE families are made of multiple copies dispersed throughout the genome which may each contribute very differently to the overall expression of their family, it was not relevant to test for correlation between variation in gene and TE expression.

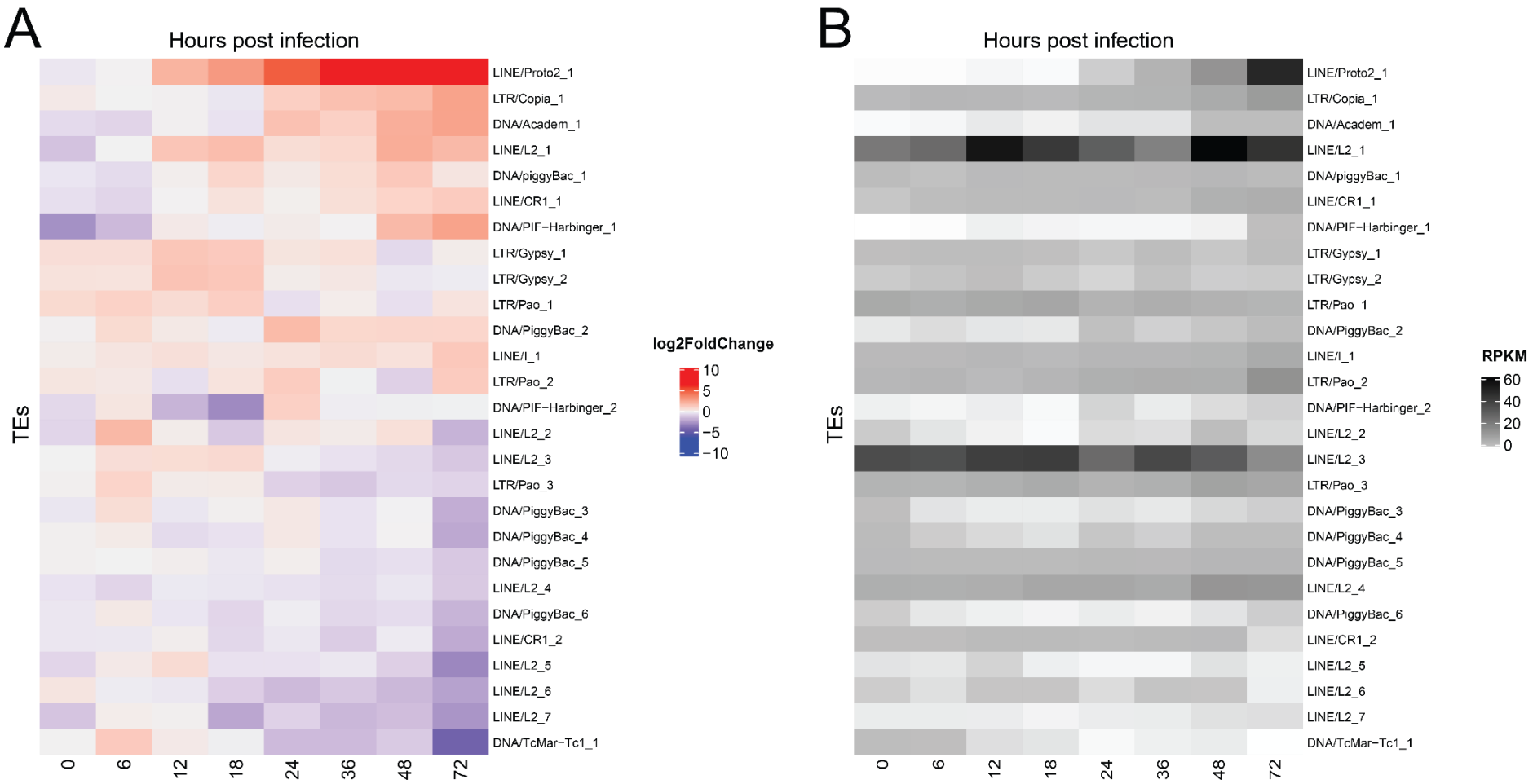



**Figure 1.2: 27 differentially expressed TE families in the in vivo data.** Heatmaps represent the log2 fold change (**A**) of the TE expression and the absolute expression expressed in RPKM (**B**) during the infection. The differential expression was computed by comparison between virus-infected and mock-infected data. A correspondence of TE names between those on the heatmaps and those in the fasta file is available in Table S1.1.

Six TE families (LINE/L2_4, LTR/Copia_1, LTR/Gypsy_1 and LTR/Gypsy_2, DNA/piggyBac_1 and DNA/piggyBac_4) were DE both in the cell line and in the *in vivo* experiment. The piggyBac_4 TE was downregulated in both cases. The Gypsy_1 and Gypsy_2 TEs were both upregulated with a peak at 18 hpi in both cases. The Copia_1 TE was upregulated in both cases with a peak at 48/72 hpi. The L2_4 TE was downregulated in both cases. Finally, the piggyBac_1 TE was upregulated in infected larvae but downregulated in the infected cell line. Contrary to the other shared DE TE families, piggyBac_1 seems to undergo different forces acting on the genomes of the Hi5 cell line and in midgut larvae, possibly due to differences in the regulatory landscapes of the two cell types. Another explanation can be the expression of different copies located in different genomic regions. Whatever the factors causing the opposite variation of expression in infected versus non-infected contexts between larvae and cell line, this TE appears much less expressed, overall, in larvae (1.4 and 2.8 RPKM at 0 and 48 hpi, respectively) than in the cell line (58 and 5.34 RPKM at 0 and 48 hpi, respectively).

Twenty-four *T. ni* TEs found inserted in AcMNPV genomes in previous studies were detected in the *T. ni* larva genome, including the DE piggyBac_1 described above. Among the 23 other TEs, 17 had an expression level <1 RPKM at 0 and 72 hpi. All six other TEs showed an increase of their expression level between 0 and 72 hpi, although not significantly compared to the control data. Among them was the TFP3 TE, expressed at 24 RPKM at 0 hpi and 27 RPKM at 72 hpi, with a peak of 42 RPKM at 12 hpi. It is interesting that most TEs found inserted in viral genomes were not DE, except TFP3 (in cell line) and piggyBac_1, which had an opposite expression variation depending on whether it was expressed in cell line or *in vivo*.

Among the 410 TE families detected in the *T. ni* larva genome, 303 (74% of TE families) were not expressed (RPKM<1) and 86 (21%) were weakly expressed (1<RPKM<10) at 0 hpi. This is consistent with the fact that few TE families are expected to be expressed in normal conditions because of the various mechanisms that have evolved to silence them (Slotkin and Martienssen 2007). The number of silenced TEs decreased at 72 hpi, with 264 (64%) TE families being not expressed and 113 (28%) being weakly expressed. This suggests a global trend towards activation or overexpression during viral infection. However, the magnitude of the variation in expression is weak. This weak trend somewhat contrasts with what we observed in the cell line, whereby infection by AcMNPV tended to be associated with a general decrease in TE expression. This suggests the nature and/or the strength of the interactions between host

cells, TEs and the virus differ between the cell line and a living organ, at least in *T. ni*. In addition to the piRNA pathway that actively represses TEs in lepidopterans, epigenetic marks, such as 5-methylcytosine (5mC), are involved in TE regulation (Deniz et al. 2019). Thus, differences in the strength of the piRNA response and/or in epigenetic landscape may explain the variation of TE expression observed between the larvae and cell line.

Contrary to Macchietto et al. (2020) who found a global upregulation of TE expression after viral infections of various cell lines, our results show there are as many upregulated TE families (13 and 36 *in vivo* and in cell line, respectively) as there are downregulated ones (14 and 30 *in vivo* and in cell line, respectively). The majority of DE TEs are weakly expressed. The *T.ni* TEs previously found to be inserted in AcMNPV genomes are not more represented in upregulated TEs. Only few TEs (as TFP3 in the cell line) show a strong increase in their expression during infection. Overall, the magnitude of the variation in TE expression during infection by AcMNPV is relatively weak in both larvae and the cell line. In both cases, our analyses reveal that variation in TE expression in the infected condition is specific to some TEs, with no sign of global up- or downregulation. These results are in agreement with what is known about the impact of stress on TE expression in eukaryotes, where no consensus clearly emerges (Horváth et al., 2017).

**TE landscape in the *T. ni in vivo* genome.**

We then assessed whether the expression level of TE families in *T. ni* larvae could be associated with factors such as TE copy number, TE age or TE proximity with genes. To do so, we first characterized the TE landscape of the *T. ni* larva genome, which had not been done in the original publication (Chen et al., 2019). We annotated 15,879 copies >300 bp in length in the *T.ni* larva genome, corresponding to 34 RepeatModeler TE superfamilies and 410 TE families. DNA TEs were the most abundant with 9,157 copies, followed by LINEs (3,693 copies), LTRs (2,982 copies) and rolling circle TEs (47 copies). The most abundant superfamilies were the class II DNA/piggyBac (2,175 copies), the class I LINE/L2 (1,660 copies) and the class I LTR/Gypsy (1,469 copies), which collectively accounted for about one third of all TE copies (Figure 1.3A). On the contrary, some superfamilies had few copies, like Mariner (234 copies), Proto2 (173 copies) or Transib (170 copies).

The overall copy divergence to consensus ranged from 0 to 36.6% (median 4.5% and mode 2.8%) (Figure 1.3B). Almost 200 copies were identical to their consensus (0% divergence).

Among the 410 TE families, 184 showed a peak of copy divergence to consensus <5%, with seven TE families peaking at 0%. At the superfamily level, 10 out of 34 superfamilies showed a peak of copy to consensus divergence <5%, 7 of which peaking at 0% (TcMar-m4, Harbinger, L2, IS, Mariner, RTE-RTE and Loa). Together, these results indicate that a relatively large fraction of TE copies transposed very recently in the *T. ni* larva genome, which in turn suggests that several TE families are still likely active.

**Analysis of TE expression in *T. ni* somatic tissues**

Based on the *T. ni* larva TE landscape, we found no correlation between the level of expression of TE families at 0 or 72 hpi and their copy number (>300 bp) or their age (approximated by the distribution mode of copy-to-consensus divergence) (Figure 1.4). Similarly, TE family expression level at 0 or 72 hpi was not correlated with average distance of TE copies to nearest genes. Considering only the expression level of DE TE families did not reveal any further correlation with these variables. We did not perform the analysis only considering TE families with high expression values (>10 RPKM) because of the low number of such families. One TE family (LINE/L2) had a very high expression compared to others (1,499 RPKM at 0 hpi and 323 RPKM at 72 hpi, Figure 1.4) but it was not DE. It has 28 copies in the genome that are 2.6% divergent to the consensus sequence on average and are located at 2,960 bp of nearest gene on average. Lack of correlation between the level of TE expression and the three TE family features we tested is in part likely due to the overall low level of TE expression observed in *T. ni* midgut larvae. To further illustrate this, we compared TE expression to expression of *T. ni* genes as reported in Chen et al. (2014) (Figure 1.5). This analysis showed that TEs are expressed at low levels (median <0.5 RPKM) in all control and infected conditions and that their overall level of expression is not affected by the viral infection. By contrast, genes were expressed at higher levels (median >1.5 RPKM) in all conditions, with marked variation between infected and non-infected conditions, including a strong increase to 4 RPKM at 72 hpi in the infected condition. Together, these results further show that contrary to the expression of genes which is markedly impacted by the AcMNPV infection (as shown by Chen et al. 2014), that of TEs is not affected in the midgut of *T. ni* larvae.

Based on the results of our analysis of TE expression in infected and non-infected *T. ni* larvae, we conclude that the ability of some *T. ni* TEs to transpose in viral genomes upon infection is not linked to stress-mediated overexpression of these TEs in midgut cells. In turn, we propose that the low level of expression of these TEs in midgut cells may be sufficient for them to

transpose in AcMNPV genomes. Alternatively, transposition into viral genomes may occur in tissues other than those constituting the midgut, in which TE expression might be higher than in the midgut. In this regard, it is noteworthy that the tissue tropism of AcMNPV includes most cell types of lepidopteran larvae (Barrett et al. 1998; Engelhard et al. 1994; Rahman and Gopinathan 2004). It would thus be interesting to repeat this analysis on several other tissues and/or on whole larvae.

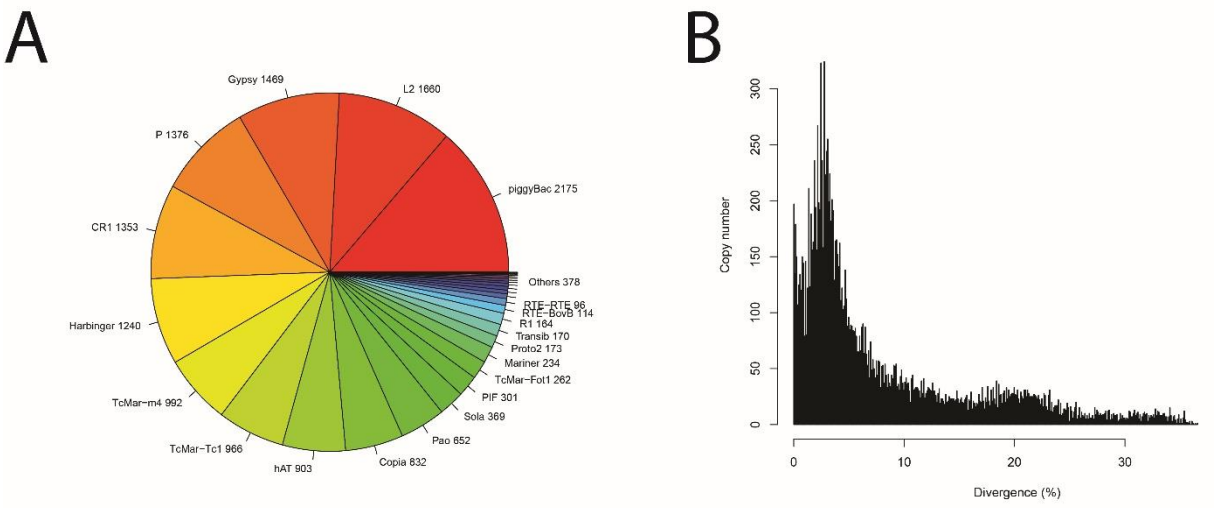



**Figure 1.3: Description of TE landscape in *T. ni in vivo* genome. A**: All the different TE superfamilies detected into the genome with their respective copy number. **B**: TE copy divergence histogram for all TE families.

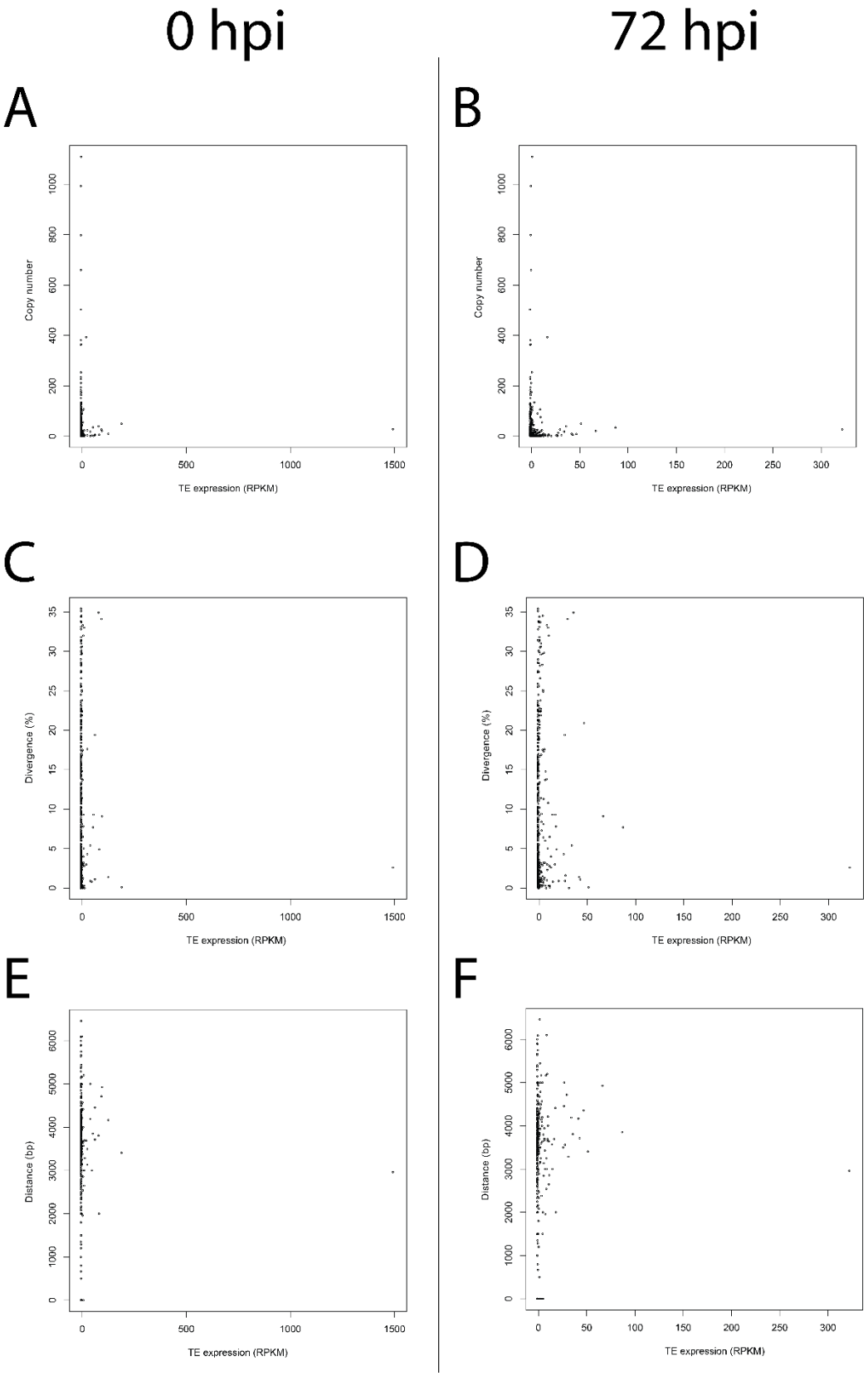



**Figure 1.4: No correlation between TE expression (RPKM) and number of copy (A and B), divergence to consensus (C and D) or distance to genes (E and F).** The left panel corresponds to TE expression at 0 hpi (A, C and E) and the right panel, to the TE expression at 72 hpi (B, D and F).

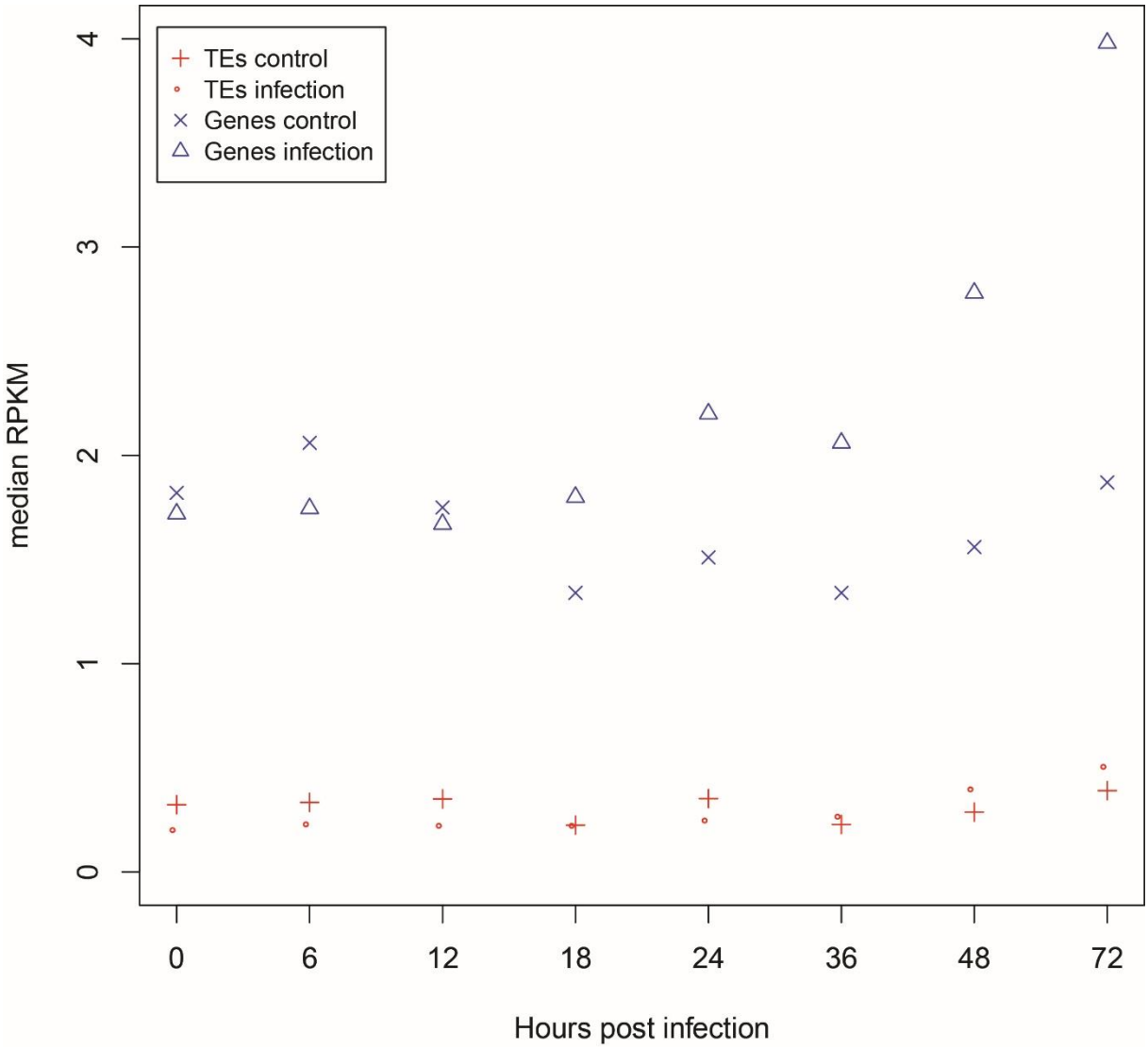



**Figure 1.5: The global TE expression in normal and infected conditions is compared to that of the global gene expression in both conditions (RPKM).**

**Expression of AcMNPV-borne TE copies**

Our search for TE-virus chimeric reads revealed no such read in the RNA-seq dataset from *T. ni* larvae infected by AcMNPV (Shrestha et al. 2018). This absence may be due to the fact that the AcMNPV genomes used to infect *T. ni* larvae bore no TE and that no TE transposed *de novo* into AcMNPV during the experiment. Another possibility is that TEs carried by AcMNPV genomes used for these experiments were not expressed. However, we previously found that while a substantial proportion of AcMNPV genomes carry moth TEs, the vast majority of individual TE insertions segregate at extremely low frequency (Gilbert et al. 2016). For example, 99% of the 1,983 different TE insertions found in the AcMNPV-infecting *T. ni* G0 dataset (the most deeply sequenced dataset) were at a frequency lower than 0.1% and the highest insertion frequency in this dataset was 1.4% (Gilbert et al. 2016; Loiseau et al. in prep.).

Furthermore, the likelihood that these TE insertions may be co-transcribed with their neighboring gene may be low. Thus, the absence of TE-virus chimeras in these data might not necessarily reflect absence of TE-borne TEs but rather indicate that the expression level of such TEs might generally be too low to be detected with our approach. In this context, the short read length (51 bp) might have further hampered our ability to detect TE-virus chimeras, as the blastn command we used does not allow finding alignments shorter than 28 bp. In addition, the average sequencing depth did not exceed 2,550 X in this study. Though sufficient to detect TE insertions in principle, deeper sequencing would have undoubtedly increased the likelihood to detect expressed TEs.

By contrast, we were able to detect a large number of TE-virus chimeras in the RNA-seq dataset from the AcMNPV-infected *T. ni* cell line (Chen et al., 2014). Considering the seven time points and the three biological replicates at each time step, 11,914 chimeric reads were identified. Among the eleven TEs involved in these chimeras, three Class II piggybac and one Harbinger TEs were found in different replicates at different time steps. The eight other TEs (seven Class II and one Class I) were found in a single or a just a few replicates or time steps (Table 1.2). Importantly, a single TE (TFP3) accounted for the vast majority of the chimeric reads (11,533 out of 11,914), with 5,580 and 5,953 reads aligning at its 5' and 3' extremity, respectively. Among the other chimeras, 64 aligned at the 5' end of piggybac 2105, 22 reads aligned at the 3'end of piggybac 22360 and insertions of Harbinger Hichhiker TE were supported by 16 reads (7 at the 5' extremity and 9 at the 3' extremity). Among all 11,914 TE-virus chimeras, only 1.93% did not align at the TE tips but on their internal part, indicating that the vast majority of chimeras correspond to expression of TEs that were generated by *bona fide* transposition. Further supporting the biological nature of the chimeras detected in this analysis, we found target site duplications (TSDs) for TFP3 and Harbinger TEs. For example, for Harbinger, two chimeric reads were found to align on the viral genome 3 bp apart from each other, separated by a TTA motif, known to be typically duplicated during Harbinger transposition (Sinzelle et al. 2008). For TFP3, 19,491 reads were identified supporting TSDs: 4,940 reads at 12 hpi, 3,984 at 18 hpi, 2,784 at 24 hpi, 2,826 at 36 hpi and 4,958 at 48hpi. These reads indicated the expression of 202 different TFP3 insertions among which 44 were expressed at 12 hpi, 38 at 18 hpi, 24 hpi and 36 hpi, and 45 at 48 hpi. The two TSD motifs flanking these insertions (TTAA and ATAA) corresponded to those typically generated upon transposition of piggybac elements (Figure 1.6; Bouallègue et al. 2017).

Regarding the dynamics of virus-borne TEs during infection, we observed a sharp increase in the number of chimeric reads from 12 hpi followed by relatively steady counts afterwards. Three chimeric reads were detected at 0 hpi, 107 at 6 hpi, 2,458 at 12 hpi, 2,077 at 18 hpi, 2,093 à 24hpi, 2,078 at 36 hpi and 2,868 at 48 hpi (Table 1.2). The peak of TE-virus chimeras detected at 12 hpi was in agreement with the results of Chen et al. (2014), who showed that expression of AcMNPV genes reaches its highest levels at this time of the infection.

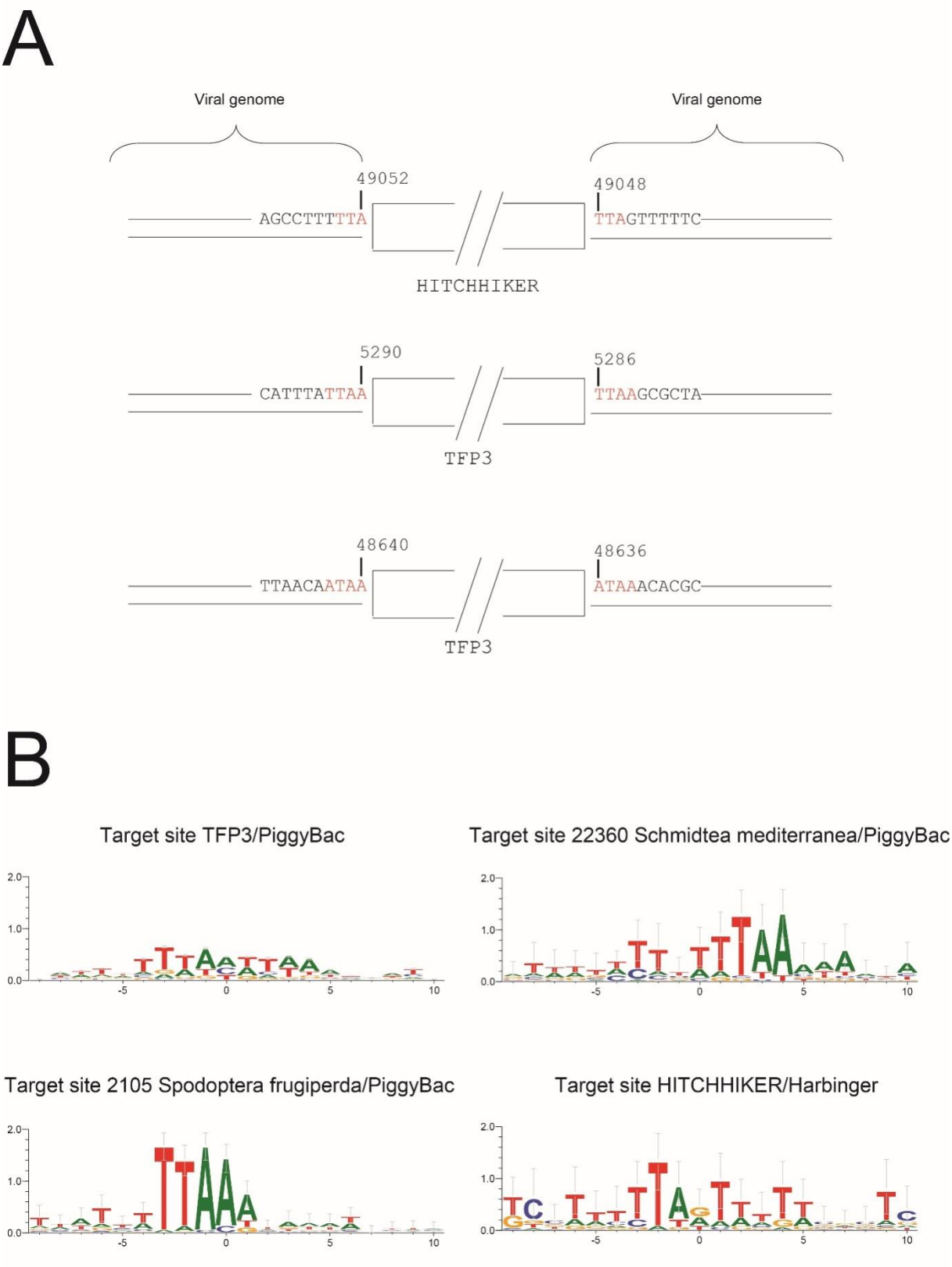



**Figure 1.6: Insertion mechanism of TEs in AcMNPV. A:** Examples of transposition events with TSD (in red). **B:** Logo sequences of TE insertions. The « 0 » corresponds to the TE insertion point in the viral genome.

We then mapped the distribution of TE-virus chimeras along the viral genome for each time point pooling all replicates, only focusing on TFP3, which is by far the most expressed virus-borne TE. Figure 1.7 further illustrates the sharp increase followed by steady expression of virus-borne TFP3 insertions at 12 hpi. It also reveals the presence of three highly expressed TFP3 copies, integrated at positions 4,856, 48,732 and 59,176 of the AcMNPV genome, in three different viral genes: *PH* (polyhedrin), *FP* (few polyhedra) and *Ac-Orf78*. The three genes are known to be involved in the formation of occlusion bodies (OBs). Inactivation of *FP* or *PH* leads to a drop of the AcMNPV OB formation (Fraser et al., 1983; Hink & Vail, 1973) and *Ac-Orf78* is associated with a structural protein that is essential for infectious OB formation (Tao et al. 2013). Interestingly, OBs are not necessary for the virus to replicate in cell lines and viruses unable to make OBs have a replication advantage over OB-forming viruses (Wood 1980). This may explain why most TEs found integrated in AcMNPV genomes in early studies were located in the *FP* or *PH* genes (Fraser et al. 1983; Bauser et al. 1996). Therefore, the TFP3 insertions in *FP*, *PH* and *Ac-Orf78* increased in frequency during passage of the virus in the *T. ni* cell line because their fitness cost may be much lower in these genes than elsewhere in the AcMNPV genome, or because they may provide a replication advantage to the genomes bearing them. However, the presence of TFP3 copies integrated in these genes did not impede their expression, which generated many TE-virus transcripts, increasing our ability to detect TE-virus chimeras in this dataset. Importantly, the longer read length (101 bp) produced by Chen et al. (2014) probably contributed to more efficiently detect TE-virus chimeras than in the Shrestha et al. (2018) dataset (read length 51 bp).

To obtain further insight into the expression level of TFP3 copies inserted in *FP*, *PH* and *Ac-Orf78*, we compared their expression in RPKM to the overall expression of the three genes as reported by Chen et al. (2014). For each gene, the average expression was computed over the three replicates for each time step (Table 1.3). After 12 hpi, the expression level of TFP3 is always 10 to >600 times lower than that of the three genes in which they are inserted. This suggests that either the frequency of viral genomes bearing TFP3 is not that high, or that the presence of a TPF3 copy in a gene somewhat lowers its expression. However, it is noteworthy that the absolute expression level of virus-borne TFP3 copies is higher than 35% of AcMNPV genes after 12 hpi. These results are in agreement with the high upregulation of TFP3 during the course of the infection we observed above in our analysis of DE TEs. Indeed, this TE was found to be the second most upregulated and the first one in terms of expression level found in the cell line data (Figure 1.1). Interestingly, our results also suggest that the upregulation of

TFP3 upon viral infection may be in large part due to expression of viral-borne TFP3 copies rather than to enhanced expression of TFP3 copies located in the *T. ni* genome.

Together, our results show that at least 11 TEs from a *T. ni* cell line can be inserted in and transcribed from AcMNPV genomes. Canonical insertion and transcription are supported by the presence of expected TSD motifs. Importantly, our approach only allows the detection of TEs that are co-transcribed with the upstream or downstream viral gene. Although the expression level of virus-borne TFP3 copies is equivalent to that of some AcMNPV genes, it is possible that expression of these virus-borne TFP3s is here highly underestimated because we only measured expression of TFP3 at junctions with the virus. Yet, some TEs carry their own promoter. In piggyBac TEs, the promoter is located in the repeated sequence at the 5' end (Cadiñanos and Bradley 2007). Thus, transcripts of virus-borne TFP3 copies initiated in the promoter region and terminated before the next viral gene upstream the 3' TE-virus junction were not taken into account here. Moreover, potential TE transcripts co-encapsidated into virions but not inserted into the viral genome, as found in several RNA viruses (e.g. Routh et al., 2012), were not considered here. In this regard, it will be interesting to assess the capacity of large dsDNA viruses to encapsidate TEs not integrated in their genomes. Our study also suggests that analyses of DE TEs during a viral infection must be interpreted with caution as an increase in TE expression level could be in part caused by expression of viral-borne TE copies rather than overexpression of host-borne TE copies.

In conclusion, we characterized the *T. ni* TE expression midgut larvae and in Tnms42 cells facing a biotic stress in the form of an AcMNPV infection. We found 27 DE TEs in midgut larvae, 13 and 14 of which were up- and downregulated, respectively. Another 66 DE TEs were identified in Tnms42 cells, 30 and 36 of which were up- and downregulated, respectively. Six TEs were DE in both datasets (2 piggyBac, 2 Gypsy, 1 L2 and 1 Copia TE families). Among all DE TEs, only two were previously found inserted in AcMNPV genomes: TFP3 and piggyBac_1 (Fraser et al. 1983; Wang and Fraser 1993; Gilbert et al. 2016). PiggyBac_1 did not have a consistent expression pattern between midguts and cultured cells. TFP3 was upregulated in the cell line and it was one of the most expressed DE TEs, similar to canonical host gene expression. These results confirm that most TE families are repressed in somatic host genomes in normal conditions, and they show that a viral infection can provoke upregulation of specific TEs, rather than a global de-repression. No correlation was found between TE expression and TE copy number, divergence to consensus or distance to genes. We finally

provided evidence of TE expression after insertion into viral genomes, a step necessary in the horizontal transfer process of TEs if viruses are to act as vectors. Thus, altogether, these results contribute to support viruses as potential vectors of TEs between animals.

**Table 1.2: Number of transcribed biological chimeric reads of inserted TEs.** The term "Others" refers to the less frequent TEs found inserted into viral genomes.

| | | 0h | | 6h | | 12h | | 18h | | 24h | | 36h | | 48h | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 5' | 3' | 5' | 3' | 5' | 3' | 5' | 3' | 5' | 3' | 5' | 3' | 5' | 3' |
| Replicate 1 | TFP3 | 0 | 1 | 13 | 8 | 230 | 218 | 287 | 284 | 369 | 388 | 309 | 293 | 336 | 379 |
| | PiggyBac (2105_S. frugiperda) | 0 | | 0 | | 2 | 0 | 3 | 0 | 9 | 0 | 3 | 0 | 5 | 0 |
| | PiggyBac (22360_S. mediterranea) | | | | | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| | Harbinger (HITCHHIKER) | | | | | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| | Others | | | | | 1 | 0 | 1 | 1 | 1 | 2 | 5 | 0 | 3 | 3 |
| Replicate 2 | TFP3 | 0 | 2 | 16 | 8 | 440 | 519 | 241 | 229 | 347 | 390 | 394 | 423 | 379 | 373 |
| | PiggyBac (2105_S. frugiperda) | 0 | | 0 | | 4 | 0 | 1 | 0 | 3 | 0 | 4 | 0 | 6 | 0 |
| | PiggyBac (22360_S. mediterranea) | | | | | 0 | 2 | 0 | 1 | 0 | 3 | 0 | 1 | 0 | 1 |
| | Harbinger (HITCHHIKER) | | | | | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| | Others | | | | | 0 | 0 | 3 | 0 | 2 | 2 | 4 | 2 | 2 | 1 |
| Replicate 3 | TFP3 | 0 | | 32 | 30 | 467 | 553 | 474 | 541 | 282 | 289 | 332 | 303 | 632 | 722 |
| | PiggyBac (2105_S. frugiperda) | | | 0 | | 5 | 0 | 7 | 0 | 1 | 0 | 1 | 0 | 10 | 0 |
| | PiggyBac (22360_S. mediterranea) | | | | | 0 | 5 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 3 |
| | Harbinger (HITCHHIKER) | | | | | 1 | 1 | 0 | 2 | 1 | 0 | 1 | 0 | 2 | 2 |
| | Others | | | | | 7 | 0 | 0 | 0 | 2 | 0 | 2 | 0 | 4 | 1 |
| Total | | 0 | 3 | 61 | 46 | 1158 | 1300 | 1017 | 1060 | 1017 | 1076 | 1055 | 1023 | 1380 | 1488 |
| | | 3 | | 107 | | 2458 | | 2077 | | 2093 | | 2078 | | 2868 | |

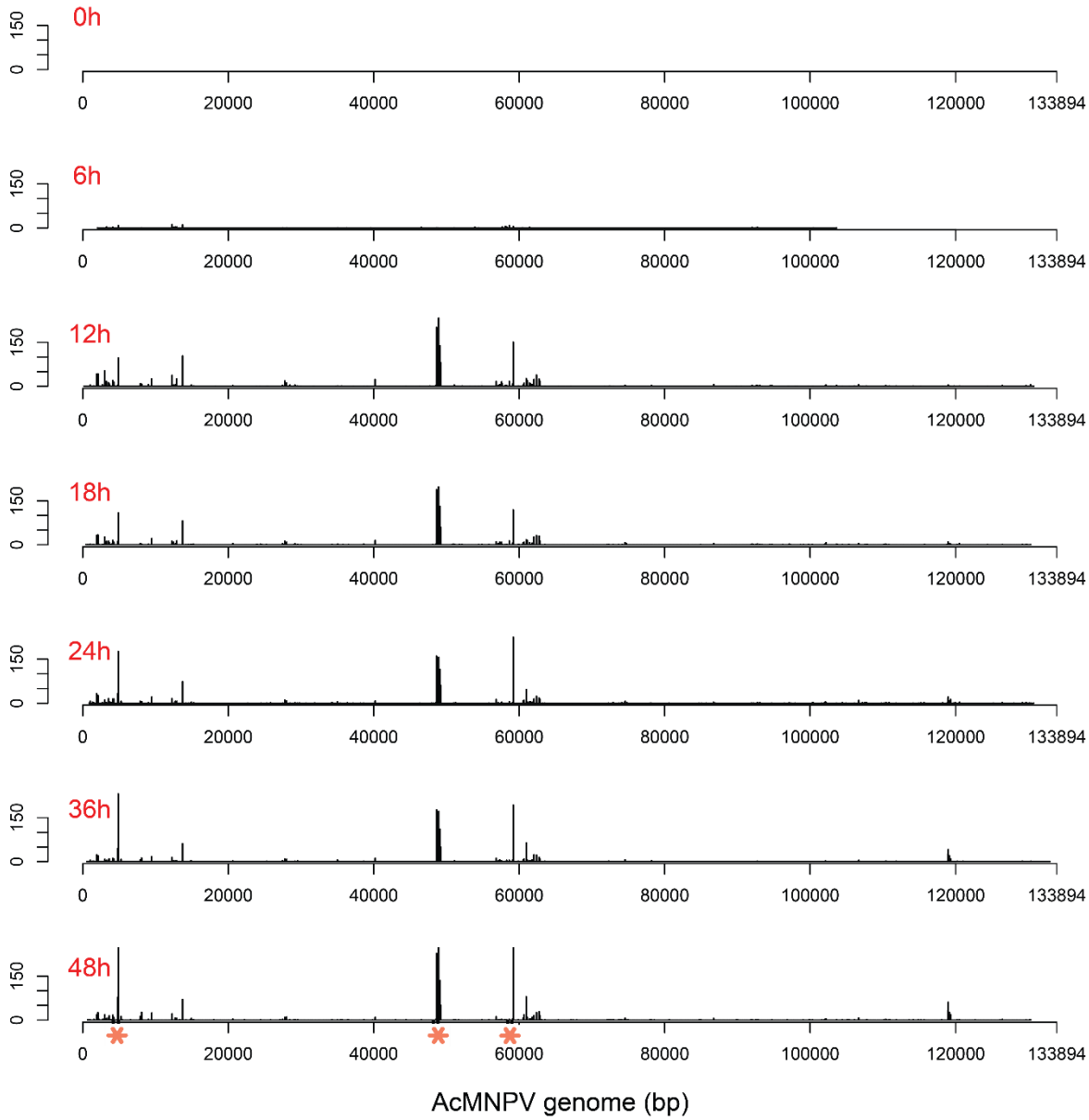



**Figure 1.7: Distribution of TE insertions for each time point.** The Y-axis corresponds to the read number. Insertions are binned into 50-bp windows. The three major insertion hotspots, shown by orange asterisks on the last distribution, correspond (from left to right) to *polyhedrin*, *FP* and *Orf78* genes, respectively.

**Table 1.3: Average transcriptomic expression of TFP3 and three viral genes (PH, FP and Orf78) of the three replicates for each time point.**

| Genes | Time points | TFP3 expression (RPKM) | Gene expression (RPKM) (Chen *et al.* 2014) |
|---|---|---|---|
| PH | 0h | 0 | 19 |
| | 6h | 34 | 11 |
| | 12h | 152 | 1124 |
| | 18h | 104 | 20444 |
| | 24h | 221 | 156323 |
| | 36h | 303 | 253928 |
| | 48h | 479 | 326276 |
| FP | 0h | 376 | 7 |
| | 6h | 7 | 4 |
| | 12h | 691 | 2605 |
| | 18h | 727 | 5414 |
| | 24h | 387 | 4907 |
| | 36h | 409 | 4678 |
| | 48h | 501 | 6050 |
| Orf78 | 0h | 116 | 44 |
| | 6h | 16 | 198 |
| | 12h | 92 | 7666 |
| | 18h | 88 | 9985 |
| | 24h | 116 | 6891 |
| | 36h | 114 | 5688 |
| | 48h | 187 | 4765 |

## References


Anwar, Sumadi, Wahyu Wulaningsih, and Ulrich Lehmann. 2017. "Transposable Elements in Human Cancer: Causes and Consequences of Deregulation." *International Journal of Molecular Sciences* 18 (5): 974. https://doi.org/10.3390/ijms18050974.

Arkhipova, Irina R. 2018. "Neutral Theory, Transposable Elements, and Eukaryotic Genome Evolution." Edited by Sudhir Kumar. *Molecular Biology and Evolution* 35 (6): 1332–37. https://doi.org/10.1093/molbev/msy083.

Barah, Pankaj, Naresh D. Jayavelu, John Mundy, and Atle M. Bones. 2013. "Genome Scale Transcriptional Response Diversity among Ten Ecotypes of Arabidopsis Thaliana during Heat Stress." *Frontiers in Plant Science* 4. https://doi.org/10.3389/fpls.2013.00532.

Barrett, John W., Andy J. Brownwright, Mark J. Primavera, and Subba Reddy Palli. 1998. "Studies of the Nucleopolyhedrovirus Infection Process in Insects by Using the Green Fluorescence Protein as a Reporter." *Journal of Virology* 72 (4): 3377–82. https://doi.org/10.1128/JVI.72.4.3377-3382.1998.

Barrón, Maite G., Anna-Sophie Fiston-Lavier, Dmitri A. Petrov, and Josefa González. 2014. "Population Genomics of Transposable Elements in *Drosophila*." *Annual Review of Genetics* 48 (1): 561–81. https://doi.org/10.1146/annurev-genet-120213-092359.

Bauser, Christopher A., Teresa A. Elick, and M.J. Fraser. 1996. "Characterization Ofhitchhiker,a Transposon Insertion Frequently Associated with Baculovirus FP Mutants Derived upon Passage in the TN-368 Cell Line." *Virology* 216 (1): 235–37. https://doi.org/10.1006/viro.1996.0053.

Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discoveryrate: a practical and powerful approach to multiple testing,Journal ofthe Royal Statistical Society Series B.85: 289–300.

Blumenstiel, Justin P. 2019. "Birth, School, Work, Death, and Resurrection: The Life Stages and Dynamics of Transposable Element Proliferation." *Genes* 10 (5): 336. https://doi.org/10.3390/genes10050336.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. "Trimmomatic: A Flexible Trimmer for Illumina Sequence Data." *Bioinformatics* 30 (15): 2114–20. https://doi.org/10.1093/bioinformatics/btu170.

Bouallègue, Maryem, Jacques-Deric Rouault, Aurélie Hua-Van, Mohamed Makni, and Pierre Capy. 2017. "Molecular Evolution of *PiggyBac* Superfamily: From Selfishness to Domestication." *Genome Biology and Evolution*, January, evw292. https://doi.org/10.1093/gbe/evw292.

Bourque, Guillaume, Kathleen H. Burns, Mary Gehring, Vera Gorbunova, Andrei Seluanov, Molly Hammell, Michaël Imbeault, et al. 2018. "Ten Things You Should Know about Transposable Elements." *Genome Biology* 19 (1): 199. https://doi.org/10.1186/s13059-018-1577-z.

Brookfield, John F. Y., and Richard M. Badge. 1997. "Population Genetics Models of Transposable Elements." In *Evolution and Impact of Transposable Elements*, edited by Pierre Capy, 6:281–94. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-011-4898-6_28.

Bucher, Etienne, Jon Reinders, and Marie Mirouze. 2012. "Epigenetic Control of Transposon Transcription and Mobility in Arabidopsis." *Current Opinion in Plant Biology* 15 (5): 503–10. https://doi.org/10.1016/j.pbi.2012.08.006.

Buchfink, Benjamin, Chao Xie, and Daniel H Huson. 2015. "Fast and Sensitive Protein Alignment Using DIAMOND." *Nature Methods* 12 (1): 59–60. https://doi.org/10.1038/nmeth.3176.

Burns, Kathleen H. 2017. “Transposable Elements in Cancer.” *Nature Reviews Cancer* 17 (7): 415–24. https://doi.org/10.1038/nrc.2017.35.

Cadiñanos, Juan, and Allan Bradley. 2007. “Generation of an Inducible and Optimized PiggyBac Transposon System†.” *Nucleic Acids Research* 35 (12): e87. https://doi.org/10.1093/nar/gkm446.

Capy, Pierre, Giuliano Gasperi, Christian Biémont, and Claude Bazin. 2000. “Stress and Transposable Elements: Co-Evolution or Useful Parasites?” *Heredity* 85 (2): 101–6. https://doi.org/10.1046/j.1365-2540.2000.00751.x.

Casacuberta, Elena, and Josefa González. 2013. “The Impact of Transposable Elements in Environmental Adaptation.” *Molecular Ecology* 22 (6): 1503–17. https://doi.org/10.1111/mec.12170.

Chen, Wenbo, Xiaowei Yang, Guillaume Tetreau, Xiaozhao Song, Cathy Coutu, Dwayne Hegedus, Gary Blissard, Zhangjun Fei, and Ping Wang. 2019. “A High-quality Chromosome-level Genome Assembly of a Generalist Herbivore, *Trichoplusia Ni*.” *Molecular Ecology Resources* 19 (2): 485–96. https://doi.org/10.1111/1755-0998.12966.

Chen, Y.-R., S. Zhong, Z. Fei, S. Gao, S. Zhang, Z. Li, P. Wang, and G. W. Blissard. 2014. “Transcriptome Responses of the Host Trichoplusia Ni to Infection by the Baculovirus Autographa Californica Multiple Nucleopolyhedrovirus.” *Journal of Virology* 88 (23): 13781–97. https://doi.org/10.1128/JVI.02243-14.

Chen, Y.-R., S. Zhong, Z. Fei, Y. Hashimoto, J. Z. Xiang, S. Zhang, and G. W. Blissard. 2013. “The Transcriptome of the Baculovirus Autographa Californica Multiple Nucleopolyhedrovirus in Trichoplusia Ni Cells.” *Journal of Virology* 87 (11): 6391–6405. https://doi.org/10.1128/JVI.00194-13.

Chenais, Benoit. 2015. “Transposable Elements in Cancer and Other Human Diseases.” *Current Cancer Drug Targets* 15 (3): 227–42. https://doi.org/10.2174/1568009615666150317122506.

Chuong, Edward B., Nels C. Elde, and Cédric Feschotte. 2017. “Regulatory Activities of Transposable Elements: From Conflicts to Benefits.” *Nature Reviews Genetics* 18 (2): 71–86. https://doi.org/10.1038/nrg.2016.139.

Cowley, Michael, and Rebecca J. Oakey. 2013. “Transposable Elements Re-Wire and Fine-Tune the Transcriptome.” Edited by Elizabeth M. C. Fisher. *PLoS Genetics* 9 (1): e1003234. https://doi.org/10.1371/journal.pgen.1003234.

Deniz, Özgen, Jennifer M. Frost, and Miguel R. Branco. 2019. “Regulation of Transposable Elements by DNA Modifications.” *Nature Reviews Genetics*, March. https://doi.org/10.1038/s41576-019-0106-6.

Dubin, Manu J, Ortrun Mittelsten Scheid, and Claude Becker. 2018. “Transposons: A Blessing Curse.” *Current Opinion in Plant Biology* 42 (April): 23–29. https://doi.org/10.1016/j.pbi.2018.01.003.

Engelhard, E. K., L. N. Kam-Morgan, J. O. Washburn, and L. E. Volkman. 1994. “The Insect Tracheal System: A Conduit for the Systemic Spread of Autographa Californica M Nuclear Polyhedrosis Virus.” *Proceedings of the National Academy of Sciences* 91 (8): 3224–27. https://doi.org/10.1073/pnas.91.8.3224.

Fouché, Simone, Thomas Badet, Ursula Oggenfuss, Clémence Plissonneau, Carolina Sardinha Francisco, and Daniel Croll. 2020. “Stress-Driven Transposable Element De-Repression Dynamics and Virulence Evolution in a Fungal Pathogen.” Edited by Irina Arkhipova. *Molecular Biology and Evolution* 37 (1): 221–39. https://doi.org/10.1093/molbev/msz216.

Fraser, M. J., Cary, L., Boonvisudhi, K. & Wang, H. G. 1995. Assay for movement of Lepidopteran transposon IFP2 in insect cells using a baculovirus genome as a target

DNA. Virology211, 397–407.
Fraser, M. J., Gale E. Smith, and Max D. Summers. 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. https://doi.org/10.1128/JVI.47.2.287-300.1983.
Fraser, M.J., John S. Brusca, Gale E. Smith, and Max D. Summers. 1985. “Transposon-Mediated Mutagenesis of a Baculovirus.” *Virology* 145 (2): 356–61. https://doi.org/10.1016/0042-6822(85)90172-2.
Fu, Yu, Yujing Yang, Han Zhang, Gwen Farley, Junling Wang, Kaycee A Quarles, Zhiping Weng, and Phillip D Zamore. 2018. “The Genome of the Hi5 Germ Cell Line from Trichoplusia Ni, an Agricultural Pest and Novel Model for Small RNA Biology.” *ELife* 7 (January): e31628. https://doi.org/10.7554/eLife.31628.
Galindo-González, Leonardo, Corinne Mhiri, Michael K. Deyholos, and Marie-Angèle Grandbastien. 2017. “LTR-Retrotransposons in Plants: Engines of Evolution.” *Gene* 626 (August): 14–25. https://doi.org/10.1016/j.gene.2017.04.051.
Gilbert, Clément, Aurélien Chateigner, Lise Ernenwein, Valérie Barbe, Annie Bézier, Elisabeth A. Herniou, and Richard Cordaux. 2014. “Population Genomics Supports Baculoviruses as Vectors of Horizontal Transfer of Insect Transposons.” *Nature Communications* 5 (1): 3348. https://doi.org/10.1038/ncomms4348.
Gilbert, Clément, Jean Peccoud, Aurélien Chateigner, Bouziane Moumen, Richard Cordaux, and Elisabeth A. Herniou. 2016. “Continuous Influx of Genetic Material from Host to Virus Populations.” Edited by Harmit S. Malik. *PLOS Genetics* 12 (2): e1005838. https://doi.org/10.1371/journal.pgen.1005838.
Grabherr, Manfred G, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn A Thompson, Ido Amit, Xian Adiconis, et al. 2011. “Full-Length Transcriptome Assembly from RNA-Seq Data without a Reference Genome.” *Nature Biotechnology* 29 (7): 644–52. https://doi.org/10.1038/nbt.1883.
Granados, Robert R., Anja C.G. Derksen, and Kathleen G. Dwyer. 1986. “Replication of the Trichoplusia Ni Granulosis and Nuclear Polyhedrosis Viruses in Cell Cultures.” *Virology* 152 (2): 472–76. https://doi.org/10.1016/0042-6822(86)90150-9.
Granados, Robert R., Li Guoxun, Anja C.G. Derksen, and Kevin A. McKenna. 1994. “A New Insect Cell Line from Trichoplusia Ni (BTI-Tn-5B1-4) Susceptible to Trichoplusia Ni Single Enveloped Nuclear Polyhedrosis Virus.” *Journal of Invertebrate Pathology* 64 (3): 260–66. https://doi.org/10.1016/S0022-2011(94)90400-6.
Hink, W. F., and P. V. Vail. 1973. A plaque assay for titration of alfalfa looper nuclear polyhedrosis virus in a cabbage looper (TN-368) cell line. J.Invertebr.Pathol. 22: 168 -174.
Horváth, Vivien, Miriam Merenciano, and Josefa González. 2017. “Revisiting the Relationship between Transposable Elements and the Eukaryotic Stress Response.” *Trends in Genetics* 33 (11): 832–41. https://doi.org/10.1016/j.tig.2017.08.007.
Huang, Jianhua, Yushuai Wang, Wenwen Liu, Xu Shen, Qiang Fan, Shuguang Jian, and Tian Tang. 2017. “EARE-1, a Transcriptionally Active Ty1/Copia-Like Retrotransposon Has Colonized the Genome of Excoecaria Agallocha through Horizontal Transfer.” *Frontiers in Plant Science* 8 (January). https://doi.org/10.3389/fpls.2017.00045.
Hummel, Barbara, Erik C Hansen, Aneliya Yoveva, Fernando Aprile-Garcia, Rebecca Hussong, and Ritwick Sawarkar. 2017. “The Evolutionary Capacitor HSP90 Buffers the Regulatory Effects of Mammalian Endogenous Retroviruses.” *Nature Structural & Molecular Biology* 24 (3): 234–42. https://doi.org/10.1038/nsmb.3368.

Jang, Hyo Sik, Nakul M. Shah, Alan Y. Du, Zea Z. Dailey, Erica C. Pehrsson, Paula M. Godoy, David Zhang, et al. 2019. "Transposable Elements Drive Widespread Expression of Oncogenes in Human Cancers." *Nature Genetics* 51 (4): 611–17. https://doi.org/10.1038/s41588-019-0373-3.

Jehle, Johannes A., Antje Nickel, Just M. Vlak, and Horst Backhaus. 1998. "Horizontal Escape of the Novel Tc1-Like Lepidopteran Transposon TCp3.2 into Cydia Pomonella Granulovirus." *Journal of Molecular Evolution* 46 (2): 215–24. https://doi.org/10.1007/PL00006296.

Kale, Shubha, Lakisha Moore, Prescott Deininger, and Astrid Roy-Engel. 2005. "Heavy Metals Stimulate Human LINE-1 Retrotransposition." *International Journal of Environmental Research and Public Health* 2 (1): 14–23. https://doi.org/10.3390/ijerph2005010014.

Kidwell, Margaret G., and Damon R. Lisch. 2001. "Perspective: transposable elements, parasitic DNA, and genome evolution." *Evolution* 55 (1): 1–24. https://doi.org/10.1111/j.0014-3820.2001.tb01268.x.

Lanciano, Sophie, and Marie Mirouze. 2018. "Transposable Elements: All Mobile, All Different, Some Stress Responsive, Some Adaptive?" *Current Opinion in Genetics & Development* 49 (April): 106–14. https://doi.org/10.1016/j.gde.2018.04.002.

Langmead, Ben, and Steven L Salzberg. 2012. "Fast Gapped-Read Alignment with Bowtie 2." *Nature Methods* 9 (4): 357–59. https://doi.org/10.1038/nmeth.1923.

Le Rouzic, A., T. S. Boutin, and P. Capy. 2007. "Long-Term Evolution of Transposable Elements." *Proceedings of the National Academy of Sciences* 104 (49): 19375–80. https://doi.org/10.1073/pnas.0705238104.

Lerat, Emmanuelle, Marie Fablet, Laurent Modolo, Hélène Lopez-Maestre, and Cristina Vieira. 2016. "TEtools Facilitates Big Data Expression Analysis of Transposable Elements and Reveals an Antagonism between Their Activity and That of PiRNA Genes." *Nucleic Acids Research*, October, gkw953. https://doi.org/10.1093/nar/gkw953.

Loiseau, Vincent, Elisabeth A Herniou, Yannis Moreau, Nicolas Lévêque, Carine Meignin, Laurent Daeffler, Brian Federici, Richard Cordaux, and Clément Gilbert. 2020. "Wide Spectrum and High Frequency of Genomic Structural Variation, Including Transposable Elements, in Large Double-Stranded DNA Viruses." *Virus Evolution* 6 (1): vez060. https://doi.org/10.1093/ve/vez060.

Love, Michael I, Wolfgang Huber, and Simon Anders. 2014. "Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2." *Genome Biology* 15 (12): 550. https://doi.org/10.1186/s13059-014-0550-8.

Macchietto, Marissa G, Ryan A Langlois, and Steven S Shen. 2020. "Virus-Induced Transposable Element Expression upregulation in Human and Mouse Host Cells." *Life Science Alliance* 3 (2): e201900536. https://doi.org/10.26508/lsa.201900536.

McClintock, B. 1950. "The Origin and Behavior of Mutable Loci in Maize." *Proceedings of the National Academy of Sciences* 36 (6): 344–55. https://doi.org/10.1073/pnas.36.6.344.

McClintock, B. 1984. "The Significance of Responses of the Genome to Challenge." *Science* 226 (4676): 792–801. https://doi.org/10.1126/science.15739260.

Menees, T. M., and S. B. Sandmeyer. 1996. "Cellular Stress Inhibits Transposition of the Yeast Retrovirus-like Element Ty3 by a Ubiquitin-Dependent Block of Virus-like Particle Formation." *Proceedings of the National Academy of Sciences* 93 (11): 5629–34. https://doi.org/10.1073/pnas.93.11.5629.

Miousse, Isabelle R., Marie-Cecile G. Chalbot, Annie Lumen, Alesia Ferguson, Ilias G. Kavouras, and Igor Koturbash. 2015. "Response of Transposable Elements to Environmental Stressors." *Mutation Research/Reviews in Mutation Research* 765 (July): 19–39. https://doi.org/10.1016/j.mrrev.2015.05.003.

Mita, Paolo, and Jef D Boeke. 2016. “How Retrotransposons Shape Genome Regulation.” *Current Opinion in Genetics & Development* 37 (April): 90–100. https://doi.org/10.1016/j.gde.2016.01.001.

Modolo, Laurent, and Emmanuelle Lerat. 2015. “UrQt: An Efficient Software for the Unsupervised Quality Trimming of NGS Data.” *BMC Bioinformatics* 16 (1): 137. https://doi.org/10.1186/s12859-015-0546-8.

Mortazavi, Ali, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. 2008. “Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq.” *Nature Methods* 5 (7): 621–28. https://doi.org/10.1038/nmeth.1226.

Negi, Pooja, Archana N. Rai, and Penna Suprasanna. 2016. “Moving through the Stressed Genome: Emerging Regulatory Roles for Transposons in Plant Stress Response.” *Frontiers in Plant Science* 7 (October). https://doi.org/10.3389/fpls.2016.01448.

Niederhuth, Chad E., Adam J. Bewick, Lexiang Ji, Magdy S. Alabady, Kyung Do Kim, Qing Li, Nicholas A. Rohr, et al. 2016. “Widespread Natural Variation of DNA Methylation within Angiosperms.” *Genome Biology* 17 (1): 194. https://doi.org/10.1186/s13059-016-1059-0.

Peccoud, Jean, Sébastian Lequime, Isabelle Moltini-Conclois, Isabelle Giraud, Louis Lambrechts, and Clément Gilbert. 2018. “A Survey of Virus Recombination Uncovers Canonical Features of Artificial Chimeras Generated During Deep Sequencing Library Preparation.” *G3: Genes|Genomes|Genetics* 8 (4): 1129–38. https://doi.org/10.1534/g3.117.300468.

R Core Team. 2019. R: A language and environment for statistical computing. R Foundation or Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rahman, Md.Masmudur, and Karumathil P Gopinathan. 2004. “Systemic and in Vitro Infection Process of Bombyx Mori Nucleopolyhedrovirus.” *Virus Research* 101 (2): 109–18. https://doi.org/10.1016/j.virusres.2003.12.027.

Rognes, Torbjørn, Tomáš Flouri, Ben Nichols, Christopher Quince, and Frédéric Mahé. 2016. “VSEARCH: A Versatile Open Source Tool for Metagenomics.” *PeerJ* 4 (October): e2584. https://doi.org/10.7717/peerj.2584.

Romero-Soriano, Valèria, and Maria Pilar Garcia Guerreiro. 2016. “Expression of the Retrotransposon Helena Reveals a Complex Pattern of TE Deregulation in Drosophila Hybrids.” Edited by Pawel Michalak. *PLOS ONE* 11 (1): e0147903. https://doi.org/10.1371/journal.pone.0147903.

Routh, A., T. Domitrovic, and J. E. Johnson. 2012. “Host RNAs, Including Transposons, Are Encapsidated by a Eukaryotic Single-Stranded RNA Virus.” *Proceedings of the National Academy of Sciences* 109 (6): 1907–12. https://doi.org/10.1073/pnas.1116168109.

Roy, Marlène, Barbara Viginier, Édouard Saint-Michel, Frédérick Arnaud, Maxime Ratinier, and Marie Fablet. 2020. “Viral Infection Impacts Transposable Element Transcript Amounts in *Drosophila*.” *Proceedings of the National Academy of Sciences* 117 (22): 12249–57. https://doi.org/10.1073/pnas.2006106117.

Ryan, Calen P., Jeremy C. Brownlie, and Steve Whyard. 2017. “*Hsp90 and Physiological Stress Are Linked to Autonomous Transposon Mobility and Heritable Genetic Change in Nematodes*.” *Genome Biology and Evolution*, January, evw284. https://doi.org/10.1093/gbe/evw284.

Saksouk, Nehmé, Elisabeth Simboeck, and Jérôme Déjardin. 2015. “Constitutive Heterochromatin Formation and Transcription in Mammals.” *Epigenetics & Chromatin* 8 (1): 3. https://doi.org/10.1186/1756-8935-8-3.

Schnable, P. S., D. Ware, R. S. Fulton, J. C. Stein, F. Wei, S. Pasternak, C. Liang, et al. 2009. "The B73 Maize Genome: Complexity, Diversity, and Dynamics." *Science* 326 (5956): 1112–15. https://doi.org/10.1126/science.1178534.

Shrestha, Anita, Kan Bao, Wenbo Chen, Ping Wang, Zhangjun Fei, and Gary W. Blissard. 2019. "Transcriptional Responses of the *Trichoplusia Ni* Midgut to Oral Infection by the Baculovirus Autographa Californica Multiple Nucleopolyhedrovirus." Edited by Joanna L. Shisler. *Journal of Virology* 93 (14): e00353-19, /jvi/93/14/JVI.00353-19.atom. https://doi.org/10.1128/JVI.00353-19.

Shrestha, Anita, Kan Bao, Yun-Ru Chen, Wenbo Chen, Ping Wang, Zhangjun Fei, and Gary W. Blissard. 2018. "Global Analysis of Baculovirus Autographa Californica Multiple Nucleopolyhedrovirus Gene Expression in the Midgut of the Lepidopteran Host *Trichoplusia Ni*." Edited by Joanna L. Shisler. *Journal of Virology* 92 (23): e01277-18, /jvi/92/23/e01277-18.atom. https://doi.org/10.1128/JVI.01277-18.

Sinzelle, L., V. V. Kapitonov, D. P. Grzela, T. Jursch, J. Jurka, Z. Izsvak, and Z. Ivics. 2008. "Transposition of a Reconstructed Harbinger Element in Human Cells and Functional Homology with Two Transposon-Derived Cellular Genes." *Proceedings of the National Academy of Sciences* 105 (12): 4715–20. https://doi.org/10.1073/pnas.0707746105.

Slotkin, R. Keith, and Robert Martienssen. 2007. "Transposable Elements and the Epigenetic Regulation of the Genome." *Nature Reviews Genetics* 8 (4): 272–85. https://doi.org/10.1038/nrg2072.

Song, Michael J., and Sarah Schaack. 2018. "Evolutionary Conflict between Mobile DNA and Host Genomes." *The American Naturalist* 192 (2): 263–73. https://doi.org/10.1086/698482.

Sotero-Caio, Cibele G., Roy N. Platt, Alexander Suh, and David A. Ray. 2017. "Evolution and Diversity of Transposable Elements in Vertebrate Genomes." *Genome Biology and Evolution* 9 (1): 161–77. https://doi.org/10.1093/gbe/evw264.

Tao, X. Y., J. Y. Choi, W. J. Kim, J. H. Lee, Q. Liu, S. E. Kim, S. B. An, et al. 2013. "The Autographa Californica Multiple Nucleopolyhedrovirus ORF78 Is Essential for Budded Virus Production and General Occlusion Body Formation." *Journal of Virology* 87 (15): 8441–50. https://doi.org/10.1128/JVI.01290-13.

Trojer, Patrick, and Danny Reinberg. 2007. "Facultative Heterochromatin: Is There a Distinctive Molecular Signature?" *Molecular Cell* 28 (1): 1–13. https://doi.org/10.1016/j.molcel.2007.09.011.

Van Meter, Michael, Mehr Kashyap, Sarallah Rezazadeh, Anthony J. Geneva, Timothy D. Morello, Andrei Seluanov, and Vera Gorbunova. 2014. "SIRT6 Represses LINE1 Retrotransposons by Ribosylating KAP1 but This Repression Fails with Stress and Age." *Nature Communications* 5 (1): 5011. https://doi.org/10.1038/ncomms6011.

Venner, Samuel, Vincent Miele, Christophe Terzian, Christian Biémont, Vincent Daubin, Cédric Feschotte, and Dominique Pontier. 2017. "Ecological Networks to Unravel the Routes to Horizontal Transposon Transfers." *PLOS Biology* 15 (2): e2001536. https://doi.org/10.1371/journal.pbio.2001536.

Volff, Jean-Nicolas. 2006. "Turning Junk into Gold: Domestication of Transposable Elements and the Creation of New Genes in Eukaryotes." *BioEssays* 28 (9): 913–22. https://doi.org/10.1002/bies.20452.

Voronova A, Belevich V, Jansons A, Rungis D. 2014. Stress-induced transcriptional activation of retrotransposon-like sequences in the Scots pine (*Pinus sylvestris L.*) genome. Tree Genet.Genomes 10:937–951.

Wagih, Omar. 2017. "Ggseqlogo: A Versatile R Package for Drawing Sequence Logos." Edited by John Hancock. *Bioinformatics* 33 (22): 3645–47. https://doi.org/10.1093/bioinformatics/btx469.

Walsh, A. M., R. D. Kortschak, M. G. Gardner, T. Bertozzi, and D. L. Adelson. 2013. "Widespread Horizontal Transfer of Retrotransposons." *Proceedings of the National Academy of Sciences* 110 (3): 1012–16. https://doi.org/10.1073/pnas.1205856110.

Wang, Hwei-gene Heidi, and M. J. Fraser. 1993. "TTAA Serves as the Target Site for TFP3 Lepidopteran Transposon Insertions in Both Nuclear Polyhedrosis Virus and *Trichoplusia Ni* Genomes." *Insect Molecular Biology* 1 (3): 109–16. https://doi.org/10.1111/j.1365-2583.1993.tb00111.x.

Wicker, Thomas, François Sabot, Aurélie Hua-Van, Jeffrey L. Bennetzen, Pierre Capy, Boulos Chalhoub, Andrew Flavell, et al. 2007. "A Unified Classification System for Eukaryotic Transposable Elements." *Nature Reviews Genetics* 8 (12): 973–82. https://doi.org/10.1038/nrg2165.

Wood, H.A. 1980. "Isolation and Replication of an Occlusion Body-Deficient Mutant of the Autographa Californica Nuclear Polyhedrosis Virus." *Virology* 105 (2): 338–44. https://doi.org/10.1016/0042-6822(80)90035-5.

Yoder, J. A., Walsh, C. P. & Bestor, T. H. 1997. Cytosine methylation and the ecology of intragenomic parasites. Trends Genet. 13: 335–340.

Zilberman, Daniel, Mary Gehring, Robert K Tran, Tracy Ballinger, and Steven Henikoff. 2007. "Genome-Wide Analysis of Arabidopsis Thaliana DNA Methylation Uncovers an Interdependence between Methylation and Transcription." *Nature Genetics* 39 (1): 61–69. https://doi.org/10.1038/ng1929.

Zovoilis, Athanasios, Catherine Cifuentes-Rojas, Hsueh-Ping Chu, Alfredo J. Hernandez, and Jeannie T. Lee. 2016. "Destabilization of B2 RNA by EZH2 Activates the Stress Response." *Cell* 167 (7): 1788-1802.e13. https://doi.org/10.1016/j.cell.2016.11.041.

**Supplementary data**

The supplementary file containing fasta sequences will be provided in the article. The Table S1 will also be provided in the article, it is only useful to understand names in the fasta file.

# Chapitre 2

# *Chapitre 2 : Diversité des systèmes dans lesquels les éléments transposables d'insectes s'intègrent dans les génomes viraux*

Après avoir étudié l'activité des éléments transposables (ET) de l'hôte suite au stress provoqué par une infection virale, le présent chapitre est consacré à explorer la diversité des systèmes hôte-virus pour lesquels les ET de l'hôte transposent dans les génomes viraux au cours de l'infection. Cette étape d'acquisition d'ADN hôte par les génomes viraux est essentielle pour mettre en lumière le rôle des virus comme vecteur d'ADN entre différents hôtes. Tous les couples hôte-virus présentés ici impliquent de grands virus à ADN double-brin (baculovirus, granulovirus et iridovirus, dont la taille est supérieure à 123 kb) et des hôtes invertébrés (insectes lépidoptères, diptères, et crustacés isopodes terrestres).

Chaque jeu de données de séquençage de génomes viraux a été produit après infection d'un ou plusieurs individus hôtes suivie d'une purification du virus et d'une extraction ADN. L'analyse des lectures générées permet de détecter les insertions d'ET dans les génomes de virus. L'étude de 35 jeux de données de séquençage correspondant à 11 systèmes hôte-virus différents a permis de mettre en avant 37 familles d'ET intégrés dans les génomes viraux dans 14 jeux de données différents, soit neuf systèmes hôte-virus différents. Les données analysées nous ont permis de calculer une fréquence d'insertion d'ET dans les génomes viraux. Cette fréquence varie de 0,01% à 26,33% de génomes viraux porteurs d'une insertion d'ET en moyenne. Cette étude a permis de découvrir trois nouveaux virus capables de porter des ET dans leur génome, à savoir *Agrotis segetum* Granulovirus, *Agrotis segetum* Nucleopolyhedrovirus et Invertebrate Iridescent virus 6. De même, cinq nouveaux hôtes sont désormais connus pour avoir des ET capables de transposer dans des génomes viraux, la noctuelle baignée *Agrotis ipsilon*, la noctuelle des moissons *Agrotis segetum*, le foreur ponctué de graminées *Chilo partellus*, la noctuelle du maïs *Sesamia nonagrioides* et la drosophile *Drosophila melanogaster*.

Ces différents systèmes hôte-virus ont révélé différentes dynamiques d'insertion d'ET dans les génomes viraux. Cette dynamique semble dépendre du virus et aussi de l'hôte, certains

systèmes semblant particulièrement favorables à la transposition dans les génomes viraux (e.g. système AcMNPV-*S. nonagrioides*). De plus, ces analyses ont permis de montrer que les ET portés par les génomes viraux étaient capables de transposer d'un génome viral vers un ou d'autres génomes viraux, ce qui n'avait jamais observé jusqu'à présent. Cette transposition de virus à virus soutient non seulement l'hypothèse des virus comme vecteurs d'ET entre différents hôtes, mais elle apporte aussi un éclairage nouveau sur la persistance possible d'ET dans des populations virales. Les populations virales pourraient ainsi constituer un véritable écosystème génomique pour certains ET et ne pas jouer le simple rôle d'intermédiaire entre deux hôtes.

Les expériences réalisées dans cette étude ont mené à de nombreuses collaborations qui ont contribué à l'enrichir. L'article scientifique qui en résulte et dont je suis le premier auteur sera soumis pour publication dans *Genome Biology and Evolution*.

**Monitoring insect transposable elements in large double-stranded DNA viruses supports virus-to-virus transposition**

Vincent Loiseau[1], Clémence Bouzar[1], Sandra Guillier[1], Jörg Wennmann[2], Laurent Daeffler[3], Carine Meignin[3], Elisabeth Herniou[4], Brian Federici[5], Johannes Jehle[2], Jean Peccoud[6], Richard Cordaux[6], Clément Gilbert[1*]

[1] Laboratoire Evolution, Génomes, Comportement, Écologie, Unité Mixte de Recherche 9191 Centre National de la Recherche Scientifique et Unité Mixte de Recherche 247 Institut de Recherche pour le Développement, Université Paris-Saclay, 91198, Gif-sur-Yvette, France.

[2] Institute for Biological Control, Julius Kühn-Institut, Darmstadt, Germany

[3] Modèles Insectes d'Immunité antivirale (M3i), Université de Strasbourg, IBMC CNRS-UPR9022, F-67000, France.

[4] Institut de Recherche sur la Biologie de l'Insecte, UMR7261 CNRS - Université de Tours, 37200 Tours, France.

[5] Department of Entomology, University of California, Riverside, CA 92521, USA.

[6] Université de Poitiers, Laboratoire Ecologie et Biologie des Interactions, Equipe Ecologie Evolution Symbiose, 5 Rue Albert Turpain, TSA 51106, 86073, Poitiers Cedex 9, France.

*correspondence: clement.gilbert@egce.cnrs-gif.fr

**Abstract**

Mechanisms leading to horizontal transfers (HTs) in animals are still poorly known. The role of viruses as vectors of DNA between hosts was proposed as one possibility to explain HT between animals. Previous studies have revealed the presence of host transposable elements (TEs) inserted into baculovirus genomes after infection of some lepidopteran hosts. To expand the narrow spectrum of host-virus systems studied so far to detect host-to-virus HTs, we analyzed 35 datasets encompassing 11 different host-virus systems with ten different hosts among which five lepidopterans, two crustaceans and three dipterans, and six different double-stranded DNA viruses among which two nucleopolyhedroviruses (NPVs), two granuloviruses (GVs) and two iridoviruses. A total of 37 TE families were found inserted into 14 viral datasets corresponding to nine different host-virus systems. TE insertion frequencies ranged from 0.01% to 26.33% of viral genomes, each being affected by one TE insertion on average. Our results expand the range of viruses able to carry host TEs to *Agrotis segetum* GV, *A. segetum* NPV and Invertebrate Iridovirus 6 (IIV6) and the range of hosts having TEs inserted into viruses to the black cutworm (*Agrotis ipsilon*) cells, turnip moth (*A. segetum*) larvae, spotted stalk borer (*Chilo partellus*) larvae, corn borer (*Sesamia nonagrioides*) larvae and fruit flies (*Drosophila melanogaster*). The infection of spotted stalk borer larvae, corn borer larvae and two fly species (*D. melanogaster* and *D. hydei*) with the same IIV6 parental strain that infected *D. melanogaster* S2 cells allowed us to point out different host TE dynamics. In all hosts infected with the parental IIV6 population, *D. melanogaster* TEs coming from this population were detected inserted. The difficulty was how to disentangle between subsampling biases during viral DNA sequencing or selection or genetic drift acting on host or remaining *D. melanogaster* TEs inserted into IIV6 genomes. However, it seems inter-viral TE transposition can occur during infection. Strikingly, a corn borer piggybac element was found to be inserted into >26% of *Autographa californica* multiple NPV (AcMNPV) genomes. Such a high frequency has never been described before; it raises the question of the role of TEs to delay or prevent host death during a viral infection. All these results highlight the complexity of host TE dynamics carried by viruses and reinforce the role of viruses as potential vectors of DNA between animals.

## Introduction

Much like any other component of genomes, transposable elements (TEs) are vertically transmitted from one generation to the next through reproduction. TEs can also cross species boundaries through a process not involving reproduction, called horizontal transfer (HT). The inference of HT of TEs (HTT) derives from the observation that in many instances, the low genetic distance measured between TE sequences extracted from different host organisms is largely incompatible with the divergence time of the hosts (Peccoud et al., 2018). Since the report of the P element transfer between *Drosophila willistoni* and *D. melanogaster* (Daniels et al., 1990), dozens of studies have characterized HTTs involving many branches of the eukaryote tree (Dotto et al., 2018; Schaack et al., 2010). Large-scale surveys of HTTs in plants, insects and vertebrates revealed that these transfers occurred recurrently, seeding a large fraction of the TE copies found today in these taxa (Bartolomé et al., 2009; Ivancevic et al., 2018; Reiss et al., 2019; Zhang et al., 2020). Given the strong impact TEs have on genome structure and dynamics (Bourque et al., 2018; Cordaux and Batzer, 2009), HTT can be seen as an important process shaping eukaryote genomes (Gilbert and Feschotte, 2018).

Several important questions about HTT remain to be answered, perhaps first and foremost that of the factors facilitating these transfers. Large-scale studies have shown that HTT are more likely to occur between species that are closely related and living in the same biogeographical realm than between more distantly related species living in different realms (Bartolomé et al., 2009; Peccoud et al., 2017). Interestingly, some host taxa such as teleost fish among vertebrates and moths and butterflies among arthropods seem to be more prone to HTT than others, a trend that remains to be explained (Reiss et al., 2019; Zhang et al., 2020). Furthermore, a number of HTT events have been reported that involve parasites and their hosts, suggesting host-parasite relationships may facilitate these transfers (Gilbert et al., 2010; Guo et al., 2014; Kuraku et al., 2012; Suh et al., 2016; Walsh et al., 2013). However, the molecular processes underlying HTT remain largely unknown.

Several scenarios have been proposed to explain how a TE can escape from a donor organism and enter the germline of a recipient one (Loreto et al., 2008; Schaack et al., 2010; Silva et al., 2004; Wallau et al., 2012). Two possible HTT routes are currently supported by some experimental observations. The first posits that extracellular vesicles (EVs) could act as vectors of HTT between animals. These 50-500 nm membrane-derived vesicles are secreted by most cell types, they may carry proteins, lipids or genetic material, and they are naturally present in

biological fluids (van Niel et al., 2018). It was recently shown that EVs can shuttle retrotransposons and mediate their horizontal transfer in laboratory conditions between different human cell lines or between cell culture media and mouse cell lines or embryos (Kawamura et al., 2019; Ono et al., 2019). The extent to which EVs may also shuttle TEs between species in natural conditions remains to be evaluated.

The second route of transfer receiving some experimental support involves viruses (Gilbert and Cordaux, 2017). Early studies using low-throughput targeted approaches identified TEs integrated in the genomes of several large double-stranded DNA viruses belonging to the Baculoviridae family that were passaged in moth cell cultures (Fraser et al., 1995, 1985; Miller and Miller, 1982) or whole larvae (Jehle et al., 1998, 1995). These pioneering works demonstrated that during the course of an infection, TEs can jump from the host genome to the virus genome, and that baculoviruses can receive and potentially carry a foreign genetic load from host to host. More recent works using high throughput sequencing showed that in addition to viral genomes, multiple host RNAs including TEs could be packaged in capsids of RNA viruses (Eckwahl et al., 2016; Ghoshal et al., 2015; Routh et al., 2012; Telesnitsky and Wolin, 2016). These results further emphasized the potential role of some viruses as vectors of HTT and suggested that TEs do not necessarily have to be integrated into viral genomes to be shuttled by viruses. Using an ultra-deep sequencing approach, we revisited early works on baculoviruses and characterized the whole spectrum and frequency of host TEs integrated in genomes of the Autographa californica multiple nucleo-polyhedrosis virus (AcMNPV) purified from infected moth larvae (Gilbert et al., 2016, 2014). Our results revealed that a large diversity of TEs are able to jump from the moth genome to that of the virus at each infection cycle, with an average of 4.8% of sequenced AcMNPV genomes carrying at least one host TE.

Studies of TEs segregating in baculovirus populations raised a number of outstanding questions. First, only a limited number of virus-host systems have been surveyed, such that it is still unclear whether the capacity of viral populations to carry host TEs is widespread among many viruses. Second, we showed that each individual TE insertion segregates at low frequency in AcMNPV and is purged out of the viral population over less than ten replication cycles. Indeed, we were unable to detect any shared TE insertion between an initial AcMNPV population (called G0) replicated on the cabbage looper moth (*Trichoplusia ni*) and populations purified after ten successive infection cycles of the G0 on ten lines of the beet armyworm (*Spodoptera exigua*). Thus, the extent to which virus-borne TEs can persist in viral populations remains

unclear. Third, the genome of *T. ni* and *S. exigua* were not available at the time we characterized TEs integrated in AcMNPV genomes purified from these two species. Thus, we could not exclude that some of the AcMNPV-borne TEs originated from hosts other than *T. ni* or *S. exigua* on which the virus replicated before we conducted our study. Finally, resequencing an AcMNPV population using the PacBio technology unveiled many full length TE copies integrated into viral genomes, suggesting that such copies have the capacity to encode the entire machinery necessary to transpose from the virus to the genome of another host (Loiseau et al., 2020). Yet, direct evidence supporting transposition of virus-borne TEs is still lacking.

Here, we monitored the presence, nature and frequency of TEs in 35 deep-sequenced viral genomes obtained from 11 virus-host systems involving two iridovirus and four baculovirus species. The finding of moth TEs in non-AcMNPV baculoviruses and in the iridescent virus 6 suggests that the capacity to carry host TEs may be widespread among large double-stranded DNA viruses. Importantly, our results also demonstrate that virus-borne TEs originating from a given species (e.g. *Drosophila melanogaster*) can be retrieved in the same virus after infection of another species (e.g. another fly species or a noctuid moth). Finally, we show that virus-borne TEs are able to transpose into other viral genomes during the course of an infection cycle.

## Materials & Methods

### General approach

To identify TEs integrated into viral genomes, a total of 35 Illumina sequencing datasets were analyzed in this study. Six datasets were produced by sequencing genomes of the invertebrate iridescent virus 31 (IIV31) from two pillbug species (*Armadillidium vulgare* and *Porcellio dilatatus*) (Figure 2.1B). Another dataset was produced by sequencing genomes of the invertebrate iridescent virus 6 (IIV6) purified from *Drosophila melanogaster* S2 cells. IIV6 particles purified from S2 cells were also used to infect whole *D. melanogaster* flies, as well as *D. hydei* flies and two species of moths, the spotted stalk borer (*Chilo partellus*) and the maize corn borer (*Sesamia nonagrioides*). This analysis allowed us to monitor TE gain and loss in viral populations after one replication of the same parental IIV6 population in several hosts (Figure 2.1A). In addition, we monitored the dynamics of gain and loss in a population of the AcMNPV baculovirus initially replicated on *T. ni* (Chateigner et al., 2015) and here used to infect the maize corn borer (Figure 2.2A). Finally, we surveyed the presence of TEs in 22 baculovirus sequencing datasets produced as part of other studies (Alletti et al., 2017; Fan et al., 2019, 2019; Gueli Alletti et al., 2018, 2017). These included one dataset of the Agrotis segetum granulovirus (AgseGV) purified from the turnip moth (*Agrotis segetum*), 6 datasets of the Agrotis segetum nucleopolyhedrovirus (AgseNPV) purified from a cell line of the black cutworm (*A. ipsilon*), and 15 datasets of the Cydia pomonella granulovirus (CpGV) purified from the codling moth (*Cydia pomonella*) (Figure 2.2B). Details of the infection protocols and search for TEs integrated into viral genomes are given below.

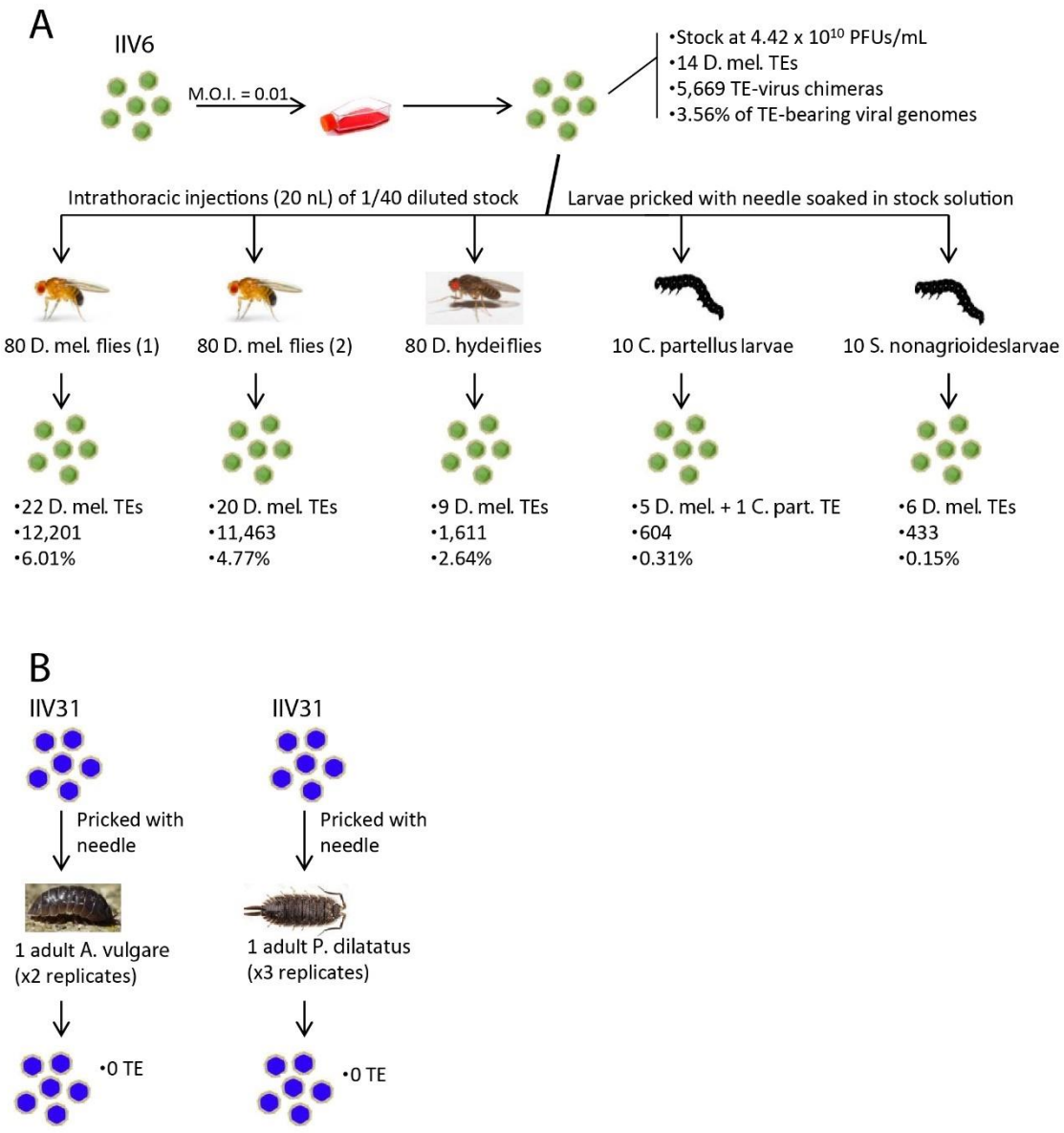



**Figure 2.1. Origin of the iridovirus sequencing datasET. A.** An IIV6 isolate was first replicated onto *Drosophila melanogaster* S2 cells and the resulting viral population was used to infect whole flies and moths. **B.** An IIV31 isolate was used to infect two pillbug species. The mode of infection, the number of transposable elements (TEs) found integrated into viral genomes, the number of chimeric reads and the percent of viral genomes carrying a TE are given for each experiment. M.O.I: multiplicity of infection.

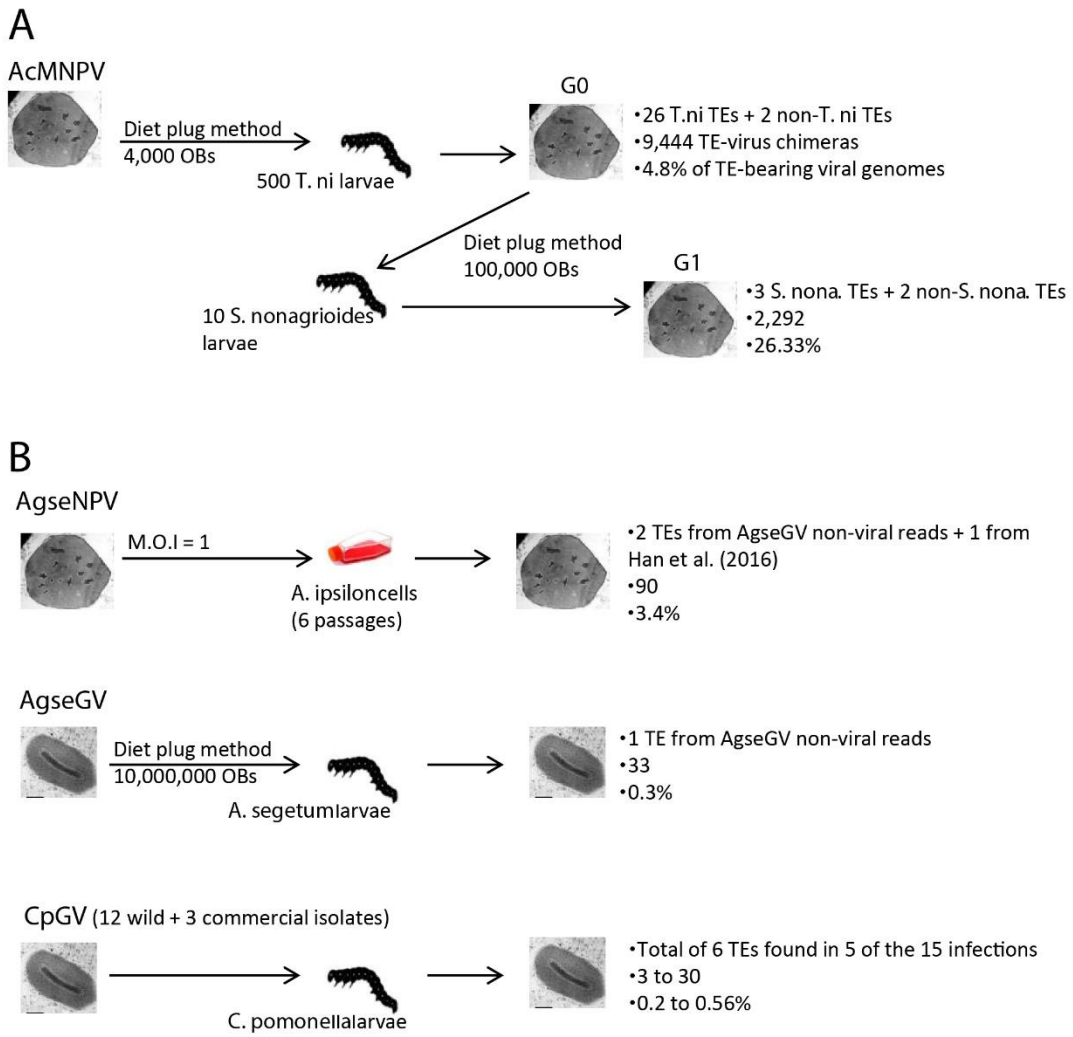



**Figure 2.2. Origin of the baculovirus sequencing datasET. A.** An AcMNPV isolate called "G0" purified from *Trichoplusia ni* larvae (Gilbert et al. 2016) was used to infect Sesamia nonagrioides larvae. The AcMNPV population purified from these larvae was called "G1". **B.** Illumina reads produced as part of other studies for three other baculoviruses were surveyed for the presence of transposable elements (TEs) integrated into viral genomes (Alletti et al., 2017; Fan et al., 2019, 2019; Gueli Alletti et al., 2018, 2017). The mode of infection (when appropriate), the number of transposable elements

(TEs) found integrated into viral genomes, the number of chimeric reads and the percent of viral genomes carrying a TE are given for each experiment. M.O.I: multiplicity of infection.

**Infection of *Drosophila melanogaster* S2 cells with invertebrate iridescent virus 6 (IIV6)**

The IIV6 viral strain used to infect *D. melanogaster* S2 cells is the one originally described in Fukaya and Nasu (1966). S2 cells were collected from two confluent T75 flasks and transferred into 50 mL tubes. The cells were counted and $2x10^8$ cells were transferred to a fresh 50 mL tube. The cells were pelleted at 500 g for 10 min at room temperature. Then, 5mL of an IIV6 solution were added to the cell pellet at a multiplicity of infection of 0.01. Cells were kept under the hood for one hour with gentle inversions every 10 min so that the cells stay in suspension. After 1 hour, 1.25 mL (i.e, $0.5 \times 10^8$ cells) was added into each of the 4 T75 flasks containing 13.5 mL of complete medium. The cells were incubated at 25°C for 5 days. Cells were detached by pipetting up and down with a 10 mL pipet and transferred into a 50 mL tube. The cells were broken down by three cycles of freezing at –20°C and thawing at 37°C. Cell debris were pelleted by centrifugation at 5,000 g for 30 min at 4°C. Supernatant was poured into an ULTRACLEAR tube (Beckman #361706) and underlaid with 1.5 mL of 30% (wt/wt) sucrose in 50 mM HEPES, pH 7, 0.1% BSA. Viral particles were pelleted at 35,000 g for 90 min at 4°C. The resulting virus pellet, showing the characteristic blue opalescence of insect iridoviruses, was resuspended in 300-400 µL 10 mM Tris pH 7.2, transferred to an Eppendorf tube and debris were pelleted by centrifugation for 5 min at 3,800 g at 4 °C. The supernatant was aliquoted, the viral titer determined ($4.42 \times 10^{10}$ PFUs/mL) and this stock solution was stored at –80 °C. Viral DNA was then extracted from an aliquot of this solution using the QIAamp DNA Mini kit (Qiagen).

**Infection of *Drosophila* flies with IIV6**

*D. melanogaster* and *D. hydei* flies were infected by intrathoracic injection of 20 nL of a 1/40 dilution of the IIV6 stock solution purified from Drosophila S2 cells (see above). Injections were performed with a nanoject II nano injector. The flies were monitored for two weeks. The abdomen of flies in which the virus successfully replicated typically turned iridescent blue 10 to 15 days post infection (Figure S2.1). Infected flies were frozen in 1.5 mL Eppendorf tubes. For each fly species, viral particles were purified from a pool of 80 infected individuals. The 80 individuals were first grinded with a plastic pestle in a Tris solution and two 5-min centrifugation steps at 500 g were performed to eliminate most of host cells and tissues. Then, an ultra-centrifugation step on sucrose cushion was performed at 35,000 g for 90 min at 4°C as

described above. The pellet was resuspended in 100 µL of Tris and viral DNA was extracted using the QIAamp DNA Mini kit (Qiagen).

**Infections of *Chilo partellus* with IIV6**

Ten fourth instar larvae of *C. partellus* were infected with the IIV6 stock solution purified from Drosophila S2 cells (see above). Larvae were pricked using a thin needle soaked in the viral solution. Fourteen days later, the larvae presented a purple iridescence and they finally died about four weeks after infection. Upon host death, viral particles were filtered through cheesecloth and two centrifugation steps were performed to eliminate most of host cells and tissues. Purification of the virus and DNA extraction were performed as above for IIV6 *Drosophila* flies.

**Infections of *Armadillidium vulgare* and *Porcellio dilatatus* with invertebrate iridescent virus 31 (IIV31)**

The IIV31 virus used to infect *A. vulgare* and *P. dilatatus* pillbugs was obtained through grinding a piece of cuticle from a naturally infected *A. vulgare* individual collected on the campus of the University of California Riverside (the same virus as in Loiseau et al. 2020). Three *P. dilatatus* and two *A. vulgare* individuals were pricked with a thin needle soaked in the viral solution. Four weeks post infection, bluish dead pillbugs were individually placed and crushed in a 1.5 ml tube in a Tris solution. Purification of the virus and DNA extraction were performed as above for IIV6 in *Drosophila* flies

**Infection of *Sesamia nonagrioides* with AcMNPV**

The AcMNPV-WP10 isolate (Chateigner et al., 2015) was used to infect 10 fourth instar larvae of *S. nonagrioides* using the diet plug method (Sparks et al., 2008). Each moth larva was fed ≈ 100,000 occlusion bodies (OBs) per 5 $mm^3$ diet plug. Upon host death, which occurred 2–5 days post-infection, OBs were first filtered through cheesecloth, purified twice by centrifugation (10 min at 7,000 rpm) with 0.1% sodium dodecyl sulfate, then distilled water, and finally resuspended in water. Approximately $1.5x10^{10}$ OBs were treated as described in Gilbert et al. (2014) to provide about 5 µg of high-quality dsDNA.

**Infection of *Cydia pomonella* larvae with Cydia pomonella granulovirus (CpGV)**

To access the diversity of TEs within a wide range of worldwide collected isolates CpGV, 15 sequenced isolates from the CpGV collection of the Institute for Biological Control, Julius

Kühn-Institut, Federal Research Centre for Cultivated Plants, Germany, were included in this study. The datasets included field isolates from Mexico (CpGV-M) and Canada (CpGV-S) (Wennmann et al., 2020), England (CpGV-E2) and Iran (CpGV-I12 and -I0X) (Fan et al., in prep.), as well as from China (CpGV-ALE, -JQ, -KS1, -KS2, -WW, -ZY and –ZY2) (Fan et al., 2020, 2019). The isolates CpGV-0006 and CpGV-V15 were derived from commercial products MadexMAX and MadexTOP (both Andermatt Biocontrol, Stahlermatten, Switzerland), respectively (Alletti et al., 2017; Zingg et al., 2011). Isolate CpGV-R5 originated from product Carpovirusine EVO2 (Arysta Lifescience, Nogueres, France) (Alletti et al., 2017; Graillot et al., 2014). Stocks of CpGV OBs were obtained either from propagation in third instar codling moth larvae for CpGV-M, -S, -E2, -I12 and -I0X or from deceased and field collected larvae for all Chinese isolates (Fan et al., 2019).

**Infection of *Agrotis segetum* larvae with Agrotis segetum granulovirus (AgseGV-DA)**

For this study, the Agrotis segetum granulovirus DA (AgseGV-DA) was also included, which had previously been propagated in *Agrotis segetum* larvae (Gueli Alletti et al., 2017). In brief, third to fourth instar larvae were starved overnight individually in and were fed the subsequent day with small cube of artificial diet that contained $10^6$ OBs of AgseGV-DA (Gueli Alletti et al., 2017). Larvae that consumed the entire piece within 12 h were transferred to virus free diet and were checked daily for symptoms of viral infection. Larvae deceased by viral infection were collected and stored at -20 °C until virus purification. The purification of OBs of AgseGV-DA and DNA extraction were performed as described previously (Gueli Alletti et al., 2017).

**Serial infections of the Agrotis segetum nucleopolyhedrovirus B (AgseNPV-B) in *Agrotis ipsilon* cell line**

To investigate TE occurrence in the AgseNPV-B virus serially passaged in *A. ipsilon* AiE1611T cell culture (Harrison and Lynn, 2008), six passages of the virus (Gueli Alletti et al., 2018) were performed. The isolate AgseNPV-B PP2 was generated as a plaque purified (PP2) clone from an *A. segetum* larvae *in vivo* propagated AgseNPV-B stock (Gueli Alletti et al., 2018). For each serial infection, 4 x $10^4$ cell/$cm^2$ were infected with a multiplicity of infection (MOI) of 1 pfu per cell starting with AgseNPV-B PP2 as the initial passage (PP2 #0). The virus treated cells were incubated for one week at 26 °C. After the initial passage, the virus was used for another ten subsequent infections (PP2 #1 to #10). After each serial infection, the tissue culture infective dose ($TCID_{50}$) was determined as described in (O'Reilly et al., 1992). Viruses from passages

#1, #3, #5, #7 and #10 OBs were harvested, purified and had their DNA extracted (Gueli Alletti et al., unpublished data).

**Sequencing of viral genomes**

The genomes of the viruses were sequenced in three batches. The first batch included six IIV31 samples purified from three *P. dilatatus* and three *A. vulgare* individuals and the IIV6 samples purified from *Drosophila* S2 cells. One µg of human DNA was added to the 2 µg of IIV6 DNA before sequencing to characterize artificial chimeras involving human TEs and the IIV6 genome (see below). These samples were sequenced by Génome Québec. A library was constructed for each sample with the NEB ultra II kit (average insert size was 260 bp), which was sequenced on a HiSeqX machine in 2 x 150 bp paired-end mode. The second batch included the two IIV6 samples purified from *Drosophila* flies and from *C. partellus* as well as the AcMNPV sample purified from *S. nonagrioides*. These samples were sequenced by Novogen. A library was constructed for each sample with the NEBnext kit (average size was 350 bp), which was sequenced on a HiSeqX machine in 2 x 150 bp mode. The third batch included the 15 isolates of CpGV, AgseGV-DA, AgseNPV-B PP2 (= #0) and its 5 serial infections (#1, #3, #5, #7 and #10). Sequencing was performed on 50 to 100 ng purified viral DNA. Libraries were generated using a NexteraXT library preparation kit and sequencing was conducted with an Illumina NextSeq500 system (StarSEQ Ltd., Mainz, Germany) generating 0.3 to 10.9 million paired-end reads of 150 nt in length (Alletti et al., 2017; Fan et al., 2020; Wennmann et al., 2020).

**Detection of transposable elements integrated in viral genomes**

The aim of this study was to characterize as comprehensively as possible the diversity of TEs that can become integrated into viral genomes during the course of an infection. To this end, we searched for TE-virus junctions in sequencing reads using the approach developed by Gilbert et al. (2016) to identify TEs integrated into the AcMNPV genome after replication of this virus into larvae of *Trichoplusia ni* and *Spodoptera exigua*. Reads were aligned separately on TE sequences and the viral genome using blastn (-task megablast). Chimeric reads for which a portion aligned on a TE *only* and the other portion aligned on the viral genome *only* were then identified based on alignment coordinates. This approach was previously used by Loiseau et al. (2020) to identify TEs from *Spodoptera exigua* integrated in genomes of another AcMNPV population purified from *S. exigua* larvae, as well as by Peccoud et al. (2018) to characterize artificial chimeras generated during the construction of sequencing libraries. Though artificial chimeras are certainly present in the various datasets analyzed here, several lines of evidence

indicate that chimeras retained for counting and calculating the frequencies of TEs integrated into viral genomes are not artificial and result from *bona fide* integration that took place during replication of the virus. First, we considered that a TE was integrated into a viral genome only if we found at least three chimeric reads between a given TE and the virus (they were not required to map to the same position along the viral genome). Second, canonical transposition of a TE into a viral genome is expected to generate junction mapping in most cases at the very extremities of the TE sequences, i.e., not anywhere along the TE sequence (Craig et al., 2015). Thus, if transposition into viral genomes occurred, chimeras covering the very extremities of the TE sequences should outnumber chimeras mapping internally to the TE sequences. This prediction was verified for all TEs (Figure 2.1, Figure S2.2, Table S2.1). Third, the addition of human DNA to the IIV6 DNA extracted from S2 cells as a control to estimate the presence of artificial chimeras allowed us to validate our approach by showing that contrary to *Drosophila* TEs, chimeras involving human TEs mapped almost exclusively internally to human TE sequences (see below). Fourth, many TEs are known to duplicate a small target sequence motif upon integration called target site duplication (TSD). As in our earlier study (Gilbert et al., 2016), we were able to identify TSD motifs for several TEs found integrated into viral genomes (Figure 2.1). Viral and TE sequences on which sequencing reads were aligned derive from a number of whole genome sequences (WGS) and TE libraries that we fully describe below.

**Viral genomes used for the blast searches**

The genome of four out of six viruses used in this study were retrieved from NCBI under accession number KR584663 (AgseGV-DA), KM102981 (AgseNPV-B), KM217575 (CpGV-M), KM667940.1 (AcMNPV strain E2). For IIV31, we used the genome sequenced in Loiseau et al. (2020) after replication of the virus in an *Armadillidium vulgare* pillbug. The sequence of the IIV31 is provided in the supplementary data of Loiseau et al. (2020). The IIV6 genome sequence available in NCBI was sequenced after replication on CF-124 cells of the *Chilo fumiferana* moth (Fukaya and Nasu, 1966). Since moths are distantly related to *Drosophila* flies, we reasoned that adaptation of IIV6 to S2 cells may have resulted in substantial changes in its genome. We thus assembled the IIV6 de novo with the tadpole program (https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/tadpole-guide/, version of December 2018, options used: ‘k=17’; ‘k=31’; ‘k=60’; ‘k=90’ with ‘mincov=100’). The different assemblies obtained with 17, 31, 60 and 90 mers with 500X and 1,500X depth coverage were then fused with Geneious version 11.0.2 (https://www.geneious.com, options: de novo assembly, Geneious assembler, high sensitivity). The final assembly was annotated

based on the available IIV6 (NC_003038.1) genome with the General Annotation Transfer Utility program (Tcherepanov et al., 2006).

**Library of transposable elements**

Our TE library included all TE consensus sequences downloaded from Repbase in October 2017 (Bao et al., 2015), all those identified in 195 insect genomes by Peccoud et al. (2017) as well as all TEs identified integrated in the AcMNPV genome after replication of the virus onto *T. ni* and *S. exigua* moths (Gilbert et al., 2016). The library also included all *D. melanogaster* TE sequences downloaded from Flybase and all *A. vulgare* TE consensus sequences characterized in (Chebbi et al., 2019). In addition, our library was completed with the consensus sequence of TEs constructed using RepeatModeler2 (Flynn et al., 2020) from the WGS of *A. ipsilon* (NCBI accession number: PNFC00000000.1), *Chilo suppressalis* (PNFC00000000.1), *C. pomonella* (QFTL00000000.2), *D. hydei* (QMEQ00000000.2) and our local assembly of *S. nonagrioides*. No WGS was available for *A. segetum*, *C. partellus* moths and for the *P. dilatatus* pillbug. To increase our chances to identify TEs in viral genomes, the sequencing reads obtained from *A. segetum* and *C. partellus* infections were also aligned on all WGS of lepidopterans available in LepBase (as of May 17th, 2018) as well as on all Noctuidae WGS and transcriptomes available in NCBI (as of October 15th, 2017). In addition for *A. segetum*, *C. partellus* and *P. dilatatus*, we aligned reads on contigs that we *de novo* assembled using all sequencing reads not mapping on viral genomes, i.e. reads resulting from sequencing residual host and non-viral/non-host DNA that remained in the viral solution after purification. Non-viral contig assembly was done with the Tadpole assembler from the BBMap tools. Residual viral chunks were filtered out from the resulting contigs. All 113,862 TEs included in our library or characterized in partial assemblies of *A. segetum*, *C. partellus* and *P. dilatatus* are provided in Dataset S1.

**Inferring horizontal transfer of TEs between insects**

To assess whether moth TEs found integrated into viral genomes underwent HT at some stage, we used them as queries to perform blastn searches (option megablast) on all insect WGS available in GenBank as of March 2020. Blastn hits showing >90% nucleotide identity to the TE over >200 bp and >80% of its length were considered further. To assess whether such high levels of TE identity were due to vertical or horizontal transmission, we compared TE synonymous distances (dS) to the distribution of dS expected to occur under vertical transmission between each species pair of interest. The dS between TEs were calculated over

the longest open reading frame (ORF) found in the two copies involved in the best alignment for each species pair of interest. The distribution of dS expected under vertical transmission was generated for each species pair of interest by calculating dS between all single-copy and complete genes extracted from the WGS of each species using the BUSCO program version 4.0.2 (Simão et al., 2015; Waterhouse et al., 2018). All genes showing homology to the Insecta set of genes (i.e., from 259 to 1,029 genes depending on the species pair) were used. The MAFFT alignment program version 7.310 (Katoh and Standley, 2013) was used to align nucleotide sequences of genes that are shared (orthologous) between species. Gene and TE alignments were analyzed in R version 3.6.1 (R Core Team., 2019) with the 'seqinr' package (Charif et al., 2005) to compute dS values.

**Frequency of TEs in viral populations and expected number of shared insertions between two sequencing datasets**

The frequency of genomes carrying TE insertions was computed following the method described in Gilbert et al. (2016). Briefly, we considered insertion frequencies to be the ratio of the number of all chimeric reads (including PCR duplicates) over the number of all viral reads (also including PCR duplicates). The ratios were normalized by the probability to detect a chimera in the data, taking into account the length of the viral genome over the read length adjusted by alignment criteria (minimum overlap needed by blast to output an alignment) and the overlap sequence length chosen in the analysis for the junction to be slightly closer to the read end. Finally, ratios had to be divided by two because we considered that one insertion event involved two junctions. We assumed the number of inserted TEs into viral genomes follows a Poisson distribution.

## Results

### Moth transposable elements in baculoviruses other than AcMNPV

We began by searching TEs integrated in genomes of six strains of a nucleopolyhedrovirus other than AcMNPV, i.e. AgseNPV replicated in cells of *A. ipsilon* and sequenced at depths varying from 839 to 2,930 X. We found TEs integrated in genomes of one strain only (AgseNPV-pp7), with a total of 90 TE-virus chimeras covering 13 different positions in the viral genome and corresponding to a frequency of 3.4% of viral genomes carrying an insertion (Table 2.1). The three TEs involved in TE-virus chimeras were class 2 DNA transposons from the piggybac and Sola superfamilies, the third one being an unclassified non-autonomous element (Table S2.1). Terminal inverted repeats (TIRs) were identified for all three TEs (Table S2.2) and all chimeric reads mapped to their extremities (Table S2.1, Figure 2.3, Figure S2.2).

**Table 2.1: Virus coverage, TE insertion frequency and number of chimeric reads for the 35 host-virus systems.** TE insertion frequency was computed considering the host TE insertion into viral genomes follows a Poisson distribution.

| Host | Virus | Average read depth over virus genome | Average read depth over host genome | Frequency of viral genomes carrying a TE (%) | Number of chimeric reads with PCR duplicates | Number of chimeric reads without PCR duplicates | Number of insertion points |
|---|---|---|---|---|---|---|---|
| *Agrotis segetum* larvae | AgseGV-DA | 7,997 | 0.0002 | 0.3 | 33 | 33 | 28 |
| *Agrotis ipsilon* cells | AgseNPV-pp0 | 839 | 0.15 | 0 | 0 | 0 | 0 |
| | AgseNPV-pp1 | 1,939 | 0.0052 | 0 | 0 | 0 | 0 |
| | AgseNPV-pp3 | 2,348 | 0.043 | 0 | 0 | 0 | 0 |
| | AgseNPV-pp5 | 956 | 0.12 | 0 | 0 | 0 | 0 |
| | AgseNPV-pp7 | 2,930 | 0.0002 | 3.4 | 90 | 58 | 13 |
| | AgseNPV-pp10 | 1,245 | 0.018 | 0 | 0 | 0 | 0 |
| *Cydia pomonella* larvae | CpGV-006 | 1,857 | 0.0002 | 0 | 0 | 0 | 0 |
| | CpGV-ALE | 1,253 | 0.00005 | 0 | 0 | 0 | 0 |
| | CpGV-E2 | 3,790 | 0.007 | 0.01 | 6 | 6 | 6 |
| | CpGV-I07 | 3,342 | 0.016 | 0 | 0 | 0 | 0 |
| | CpGV-I12 | 3,441 | 0.003 | 0.009 | 5 | 5 | 5 |
| | CpGV-JQ | 1,092 | 0.00007 | 0 | 0 | 0 | 0 |
| | CpGV-KS1 | 1,446 | 0.00008 | 0 | 0 | 0 | 0 |
| | CpGV-KS2 | 942 | 0.00009 | 0.2 | 3 | 3 | 2 |
| | CpGV-M | 3,809 | 0.001 | 0.02 | 13 | 12 | 10 |
| | CpGV-R5 | 784 | 0.035 | 0 | 0 | 0 | 0 |
| | CpGV-S | 3,192 | 0.01 | 0.56 | 30 | 30 | 5 |
| | CpGV-V15 | 2,380 | 0.001 | 0 | 0 | 0 | 0 |
| | CpGV-WW | 523 | 0.0001 | 0 | 0 | 0 | 0 |
| | CpGV-ZY | 744 | 0.00004 | 0 | 0 | 0 | 0 |
| | CpGV-ZY2 | 1,152 | 0.00009 | 0 | 0 | 0 | 0 |
| Adult *Armadillidium vulgare* individuals | IIV31 (1) | 133,086 | 0.19 | 0 | 0 | 0 | 0 |
| | IIV31 (2) | 163,802 | 0.002 | 0 | 0 | 0 | 0 |
| | IIV31 (3) | 188,230 | 0.002 | 0 | 0 | 0 | 0 |
| | IIV31 (1) | 112,750 | NA | 0 | 0 | 0 | 0 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Adult *Porcellio dilatatus* individuals | IIV31 (2) | 220,537 | NA | 0 | 0 | 0 | 0 |
| | IIV31 (3) | 184,610 | NA | 0 | 0 | 0 | 0 |
| *Drosophila melanogaster* S2 cells | IIV6 | 123,574 | 3.45 | 3.56 | 6572 | 5169 | 2761 |
| Adult *Drosophila melanogaster* individuals | IIV6 (1) | 170,859 | 6.46 | 6.01 | 16117 | 12054 | 4088 |
| | IIV6 (2) | 211,039 | 12.41 | 4.77 | 15709 | 11233 | 4018 |
| Adult *Drosophila hydei* individuals | IIV6 | 52,931 | 0.46 | 2.64 | 2048 | 1515 | 431 |
| *Chilo partellus* larvae | IIV6 | 325,206 | 0.24 | 0.31 | 1558 | 599 | 29 |
| *Sesamia nonagrioides* larvae | IIV6 | 245,899 | 26.77 | 0.15 | 555 | 437 | 55 |
| *Sesamia nonagrioides* larvae | AcMNPV | 82,103 | 0.06 | 26.33 | 37952 | 2282 | 613 |

The three TEs are absent from the WGS of *A. ipsilon*. Instead, the piggybac and Sola TEs were identified in our assembly of non-viral reads obtained from the AgseGV/*A. segetum* larvae system. The absence of these TEs from the WGS of the host on which the virus was replicated suggests they may originate from integration events that occurred during an earlier replication cycle of the virus in its original host (*A. segetum*, for which WGS are not available) or another host. Regarding the non-autonomous element, it was identified by (Han et al., 2016) in the plutellid moth *Plutella xylostella*, which diverged 156 myrs ago from *Agrotis* moths. Furthermore, one copy of an element 98% identical to the piggybac over 91% of its length (2,166 bp) was found in the WGS of another noctuid moth (*Mamestra configurata*) and six copies of an element 95% identical to the Sola over 100% of its length (4,466 bp) were found in the WGS of yet another noctuid moth (*S. exigua*). *M. configurata* and *S. exigua* diverged 42 and 56 myrs ago from *Agrotis* moths, respectively (Kumar et al., 2017). Synonymous distances calculated between the piggybac and Sola copies identified in the AgseGV/*A. segetum* non-viral reads and those found in *M. configurata* or *S. exigua* fall below the 0.5% quantile of the gene dS distribution calculated between *Agrotis* and *M. configurata* or *S. exigua* (Figure 2.4). Thus, the two TEs have been and/or are currently circulating between moths through HT. The fact that these TEs can be carried by an NPV, coupled with the known susceptibility of *Mamestra* and *Spodoptera* moths to NPVs (Goulson, 2003), further suggests that these viruses might be involved as vectors in these transfers, as previously proposed for AcMNPV (Fraser et al., 1995; Gilbert et al., 2016; Miller and Miller, 1982).

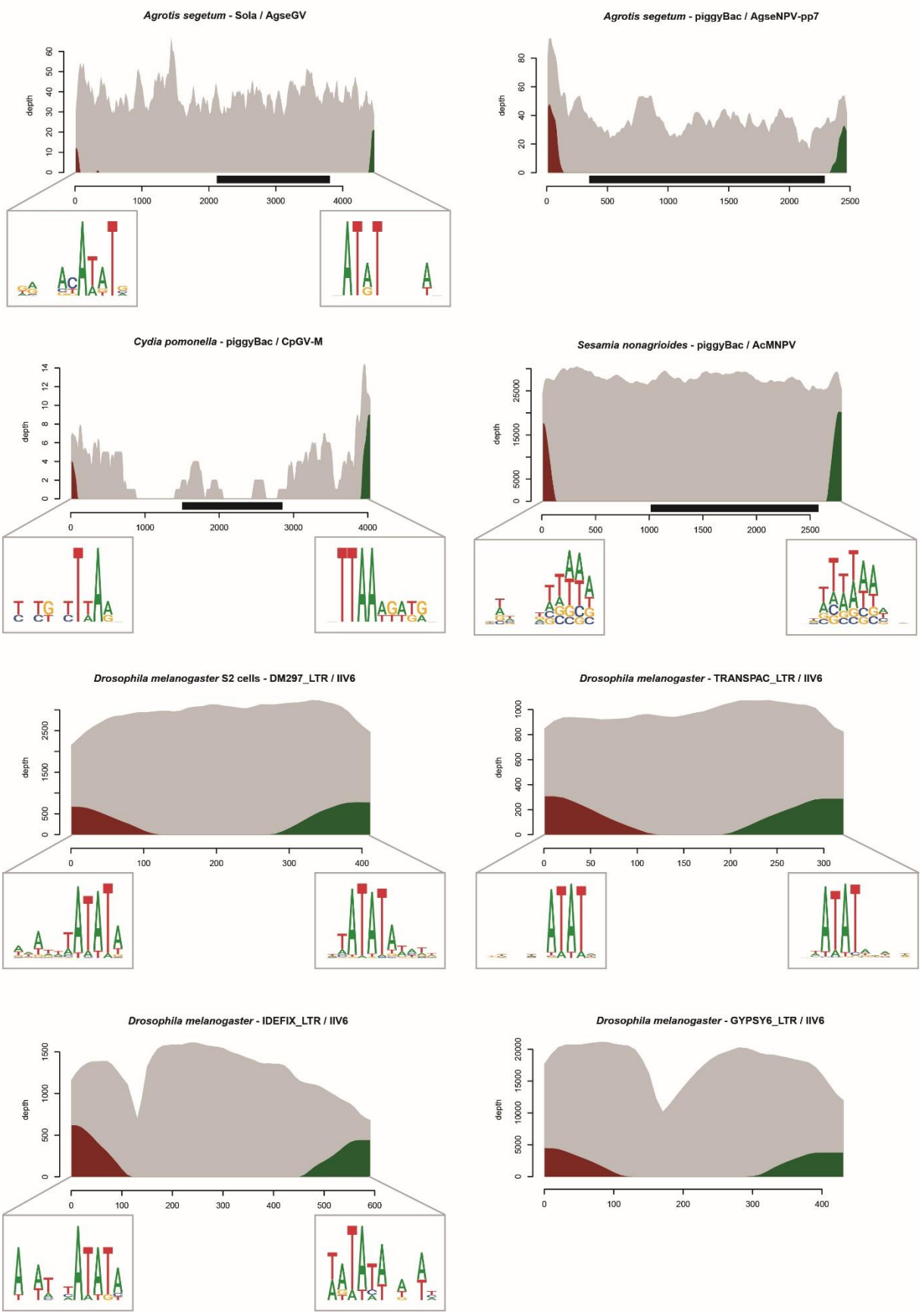



**Figure 2.3. Sequencing depth of some of the transposable elements (TEs) found integrated into viral genomes.** Sequencing depth by reads mapping entirely on the TEs is shown in grey. Sequencing depth by chimeric reads is shown in red on the 5' end of the TEs and in green for the 3' end of the TEs.

Black rectangles represent TE genes annotated in autonomous TE sequences. Target site duplications (TSDs) are shown on each side of the elements using sequence logos for all TEs for which conserved TSDs were observed (sequencing depth of the other TEs uncovered in this study is shown in Figure S2.2).

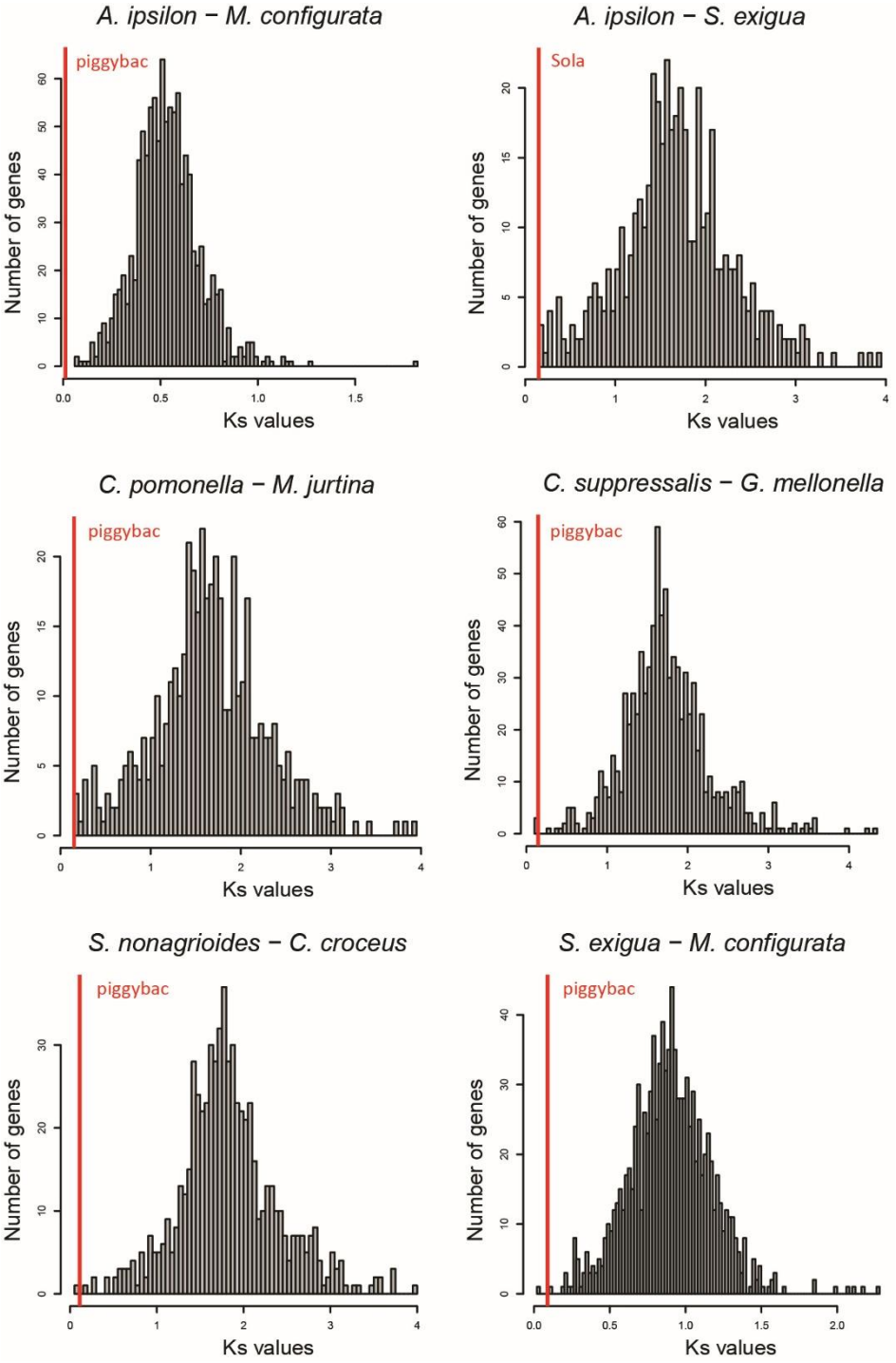



**Figure 2.4. Comparison of gene synonymous distance (Ks values) and transposable element (TE) Ks between six pairs of species.** The distribution of gene Ks is shown in grey and the TE Ks is shown by the red bar.

We then searched for TEs integrated in genomes of the AgseGV granulovirus replicated in *A. segetum* larvae and sequenced at 7,997 X. A total of 33 TE-virus chimeras were detected in the

virus genome, covering 28 different positions and corresponding to a frequency of 0.3% of viral genomes carrying an insertion (Table 2.1). All chimeras mapped to the Sola TE described above, which was found integrated into the AgseNPV (Table S2.1).

We also detected TE-virus chimeras in five of the 15 CpGV granulovirus strains replicated in *C. pomonella* larvae and sequenced at depths varying from 523 to 3,809 X. The number of chimeras varied from three (two different positions along the viral genome) to 30 (ten different positions) and the frequency of viral genomes carrying a TE varied from 0.009% to 0.56% (Table 2.1). In three of the five CpGV strains (CpGV-E2, CpGV-I12, CpGV-M), a single piggybac DNA TE, different from that detected into AgseNPV genomes, was integrated into viral genomes (Table S2.1). This piggybac was also found, together with a SHALINE-like element (non-LTR retrotransposon) in the CpGV-S dataset (Table S2.1). Finally, a Tc1/mariner DNA TE was found integrated in the CpGV-KS2 genome. All three TEs are present in the *C. pomonella* WGS. TIRs were identified in chimeric reads for the piggybac and Tc1/mariner elements (Table S2.2) and all chimeric reads mapped only to the extremities of the three elements found integrated into CpGVs (Figure 2.3, Figure S2.2, Table S2.1).

TEs closely related (>90% identical) to the SHALINE-like and Tc1/mariner element were not found in non-*Cydia* lepidopterans. The SHALINE-like is a partial, 419-bp long element, containing a truncated reverse transcriptase domain. There are more than 100 such copies in the *C. pomonella* WGS, which likely result from truncations during non-canonical reverse-transcription, as is common for non-LTR retrotransposons (Szak et al., 2002). Twenty-one of these copies are 99-100% identical to each other, consistent with this element being currently active. Regarding the Tc1/mariner, there are 10 copies of this element that are 92.5 to 94.4% identical to each other over their entire length (1,653 bp) in the *C. pomonella* WGS. Four copies of the piggybac element that are >99% identical to each other were found in the *C. pomonella* WGS, and we identified another four copies that are 90% identical to the *C. pomonella* piggybac over 88% of its length (1,734 bp) in the meadow brown (*Maniola jurtina*; Nymphalidae). This butterfly diverged 156 myrs ago from *C. pomonella*. The TE dS again falls below the 0.5% quantile distribution of the gene dS calculated between *C. pomonella* and *M. jurtina* suggesting that the piggybac has been horizontally transferred between the two species (Figure 2.4). It is unclear however, whether nymphalid butterflies are susceptible to granuloviruses and whether such viruses could have been involved as vectors in this HT event.

### *D. melanogaster* LTR retrotransposons in iridoviruses

To assess the potential of large double-stranded DNA viruses other than baculoviruses to shuttle TEs between arthropods, we searched for TEs integrated in genomes of the IIV31 iridovirus purified from three *P. dilatatus* and two *A. vulgare* individuals and in genomes of the IIV6 iridovirus after a passage on *D. melanogaster* S2 cells. Sequencing depths varied from 112,750 to 220,537 X for these datasets (Table 2.1). A third IIV31/*A. vulgare* dataset was previously analyzed in an earlier study and it was found to be devoid of TE-virus chimeras (Loiseau et al., 2020). The two other IIV31/*A. vulgare* as well as the three IIV31/*P. dilatatus* datasets were also found here to be entirely devoid of TE-virus chimeras (Table 2.1).

By contrast, we found 6,572 TE-virus chimeras in IIV6 genomes purified from *D. melanogaster* S2 cells, corresponding to 2,761 different positions along the viral genome (Figure 2.5) and we calculated that 3.56% of sequenced IIV6 genomes carried at least one TE (Table 2.1). TE-virus junctions involved 14 different *D. melanogaster* LTR retrotransposons including 6 (MDG1_LTR, DM297_LTR, DM176, IDEFIX_LTR, TRANSPAC_LTR, BLOOD_LTR) that account for 414 or more (up to 2,917) chimeras (Table S2.1). Given that this is the first report of TEs segregating in IIV6 genome populations, we set out to validate the biological nature of TE-virus chimeras by comparing these chimeras with those involving human DNA added to the IIV6 DNA sample prior to constructing the sequencing library. We focused on the structure of chimeras occurring between human TEs and the IIV6 genome. These can only be technical as IIV6 has not been in contact with human cells. Given such chimeras are not generated by transposition, they are not expected to preferentially involve the extremities of TEs. In agreement with this, chimeras involving human TEs almost all (554 out of 560) map internally to TE sequences (Table S2.3). By contrast, the vast majority of chimeras involving *D. melanogaster* TEs (5,627 out of 5,833) map at the extremities of the elements (Figure 2.3, Figure S2.2, Table S2.1). This is true for all TEs found integrated in viral genomes uncovered in this study (Figure 2.3, Figure S2.2, Table S2.1). In addition to the fact that we were able to identify TSDs for some TEs (some examples are given in Figure 2.3), these results confirm the biological nature of the TE-virus chimeras we have identified. Transposition of *D. melanogaster* LTR retrotransposons into IIV6 genomes is consistent with population genetics studies showing that most of these elements are of recent origin and currently actively transposing in natural fly populations (Kofler et al., 2015). Further supporting the recent origin and current activity of these *D. melanogaster* LTR retrotransposons, 8 of the 14 TEs we found integrated into IIV6 genomes have undergone HT between *D. melanogaster* and its sister

species *D. simulans*, which diverged less than 5 million years (myr) ago (Table S2.4, (Bartolomé et al., 2009)).

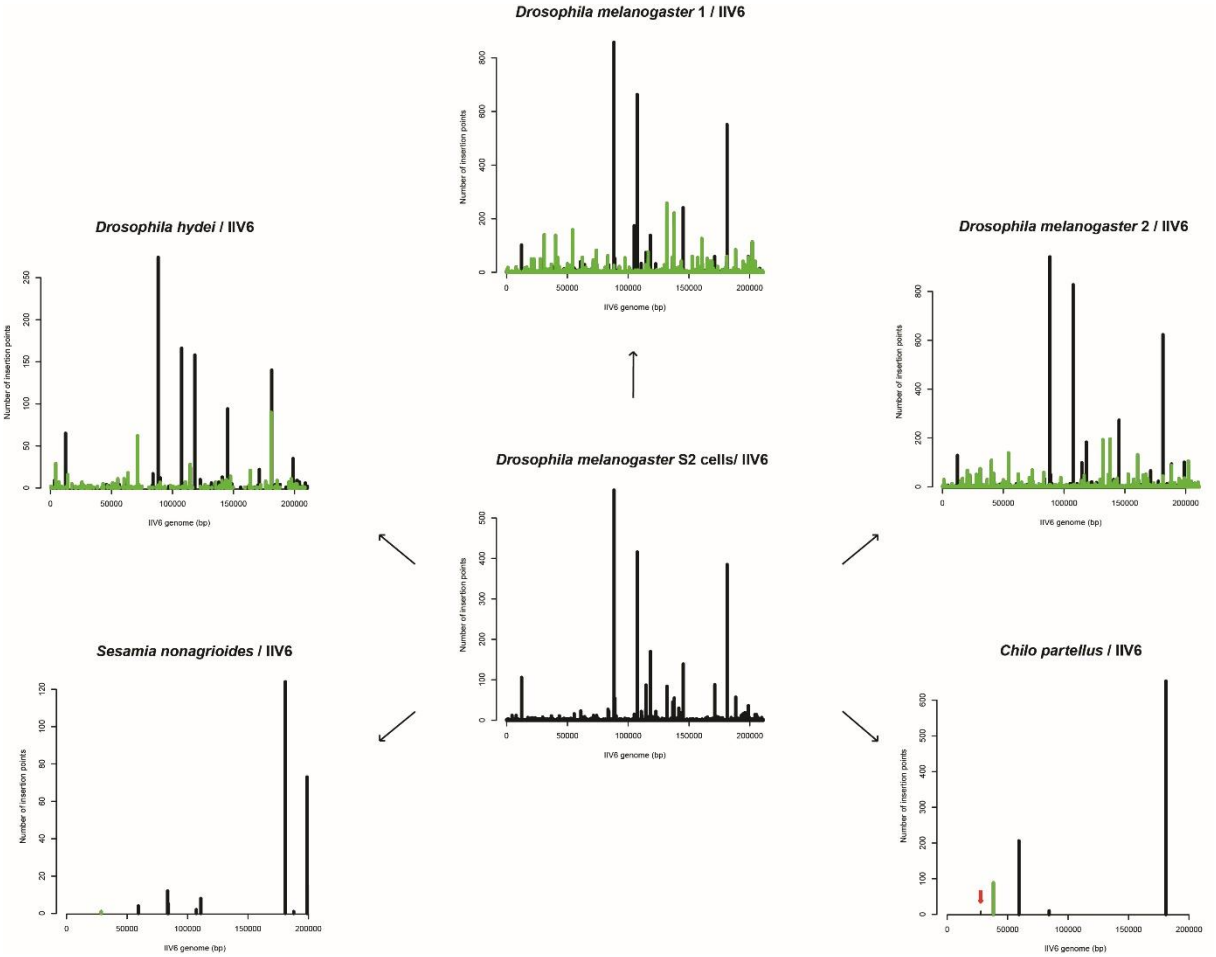



**Figure 2.5. Number of different insertions of transposable elements (TEs) in along the IIV6 genome.** The central diagram illustrates TE integrations in the parental IIV6 purified from *D. melanogaster* S2 cells. The other diagrams illustrate TE integrations in the daughter IIV6 purified from whole flies and moths. Black bars correspond to TE insertion points shared by the different IIV6, i.e. insertions of the same TE at the exact same position than in the parental IIV6 population. Green bars correspond to the TE insertion points not shared with the parental IIV6 population, i.e. which only transposed into IIV6 during infection of *Drosophila* whole flies or moth larvae. Red arrow along the IIV6 genome sequenced from *Chilo partellus* indicates the position of *C. partellus* TEs. Numbers of insertions are given in Table S2.1 for each TE.

To assess whether TEs can also transpose from whole insects into IIV6 genomes, we infected two batches of 80 *D. melanogaster* adult flies with the IIV6 population purified from S2 cells and sequenced IIV6 viral genomes purified from these infections at 170,859 and 211,039 X (Figure 2.1). We found more TE-virus chimeras in the two IIV6/adult flies batches (12,201 and 11,463 in batches 1 and 2, respectively) than in the parental IIV6/S2 cells dataset (5,669) and a higher frequency of viral genomes carrying at least one TE (6.01% and 4.77% in batches 1 and

2, respectively than in the parental IIV6/S2 cells population (3.56%) (Table 2.1, Table S2.1). All TEs except LTRMDG3_DM among the 14 present in the parental IIV6 purified from S2 cells were recovered integrated at 1,660 and 1,719 different positions in IIV6 from whole flies (batches 1 and 2, respectively) (Figure 2.5; Figure 2.6; Table S2.4). In principle, these shared TE insertions could have two sources. First, they could correspond to insertions that were present in the parental IIV6 purified from S2 cells and that persisted in the viral population during replication in whole flies, implying that IIV6 genomes bearing TE insertions were encapsidated again in whole flies. Alternatively, the TE insertions could result from transposition of TE copies located in the genomes of whole flies. Given that the TE content of *D. melanogaster* S2 cells and whole flies genomes is likely very similar, it is difficult to assess whether and what portion of insertions of the shared TEs come from the parental IIV6 population or from *de novo* transposition from whole flies. Interestingly however, it turns out that the majority of TE-virus chimeras found in IIV6 purified from whole flies involved GYPSY6, an LTR retrotransposon not present in the parental IIV6 purified from S2 cells (Figure 2.3, Figure 2.6, Table S2.1, Table S2.4). Another 8 and 6 TEs not present in IIV6 purified from S2 cells were found in batches 1 and 2, respectively, including one (GYPSY1) involved in more than 100 TE-virus chimeras in the two whole flies batches (Figure 2.6, Table S2.1, Table S2.4). Given their absence in the parental IIV6/S2 cell dataset, *de novo* integration of these TEs in IIV6 genomes can only be explained by transposition of TE copies located in the genomes of whole flies.

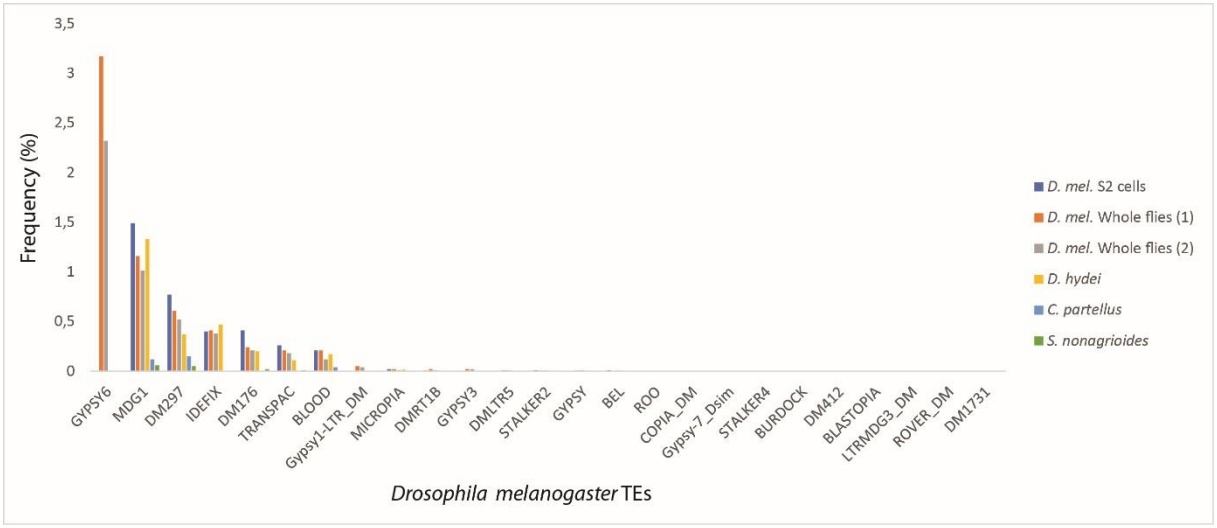



**Figure 2.6**. **Frequency of IIV6 genomes carrying each *Drosophila melanogaster* transposable elements in IIV6 populations purified from flies and moths.** More details can be found in Table S2.1.

**Transposable elements from other hosts in IIV6**

To assess whether TEs from hosts other than *D. melanogaster* can transpose into IIV6 genomes, we again used IIV6 particles purified from S2 cells to infect whole flies from another *Drosophila* species (*D. hydei*) as well as larvae from the *C. partellus* moth (Figure 2.1). We also reanalyzed with a higher precision level another IIV6 sequencing dataset from Loiseau et al. (2020) which we generated by infecting larvae from the *S. nonagrioides* moth with the IIV6 particles purified from S2 cells. Sequencing depths varied from 52,931 to 325,206 X for these datasets (Table 2.1). We found no integration of *D. hydei* TE into IIV6 genomes, nor did we find any integration of *S. nonagrioides* TE (Loiseau et al. 2020). The only non-*D. melanogaster* TE we found integrated in IIV6 is a piggybac element involved in 6 TE-virus chimeras in the *C. partellus* dataset (Table S2.1). This element was identified in assembly of non-viral reads present in this dataset. There is one copy 100% identical to this element over 100% of its length in the *Chilo suppressalis* WGS and we validated its presence in the *C. partellus* genome by PCR/Sanger sequencing (not shown). Thus, we conclude that this element transposed from *C. partellus* to IIV6 during replication of the virus in the moth larvae. Finally, we found TEs highly similar to this piggybac in six other lepidopteran species (*Megathymus ursus*, Hesperidae; *Lymantria dispar*, Erebidae; *Hyles vespertilio*, Sphingidae; *Eumeta japonica*, Psychidae; *Galleria mellonella*, Pyralidae and *Melitea cinxia*, Nymphalidae) and one trichopteran species (*Stenopsyche tienmushanensis*, Stenopsychidae). The best hit for each species genome was 92% to 95% identical the the *C. partellus* element over between 1,341 bp for the trichopteran species *S. tienmushanensis* (53% of the piggybac length) and 2,477 bp for the lepidopteran species *M. ursus* (100% of the TE length). These species diverged from *C. partellus* between 90 Mya (Pyralidae; Kawahara *et al.*, 2019) and 232 Mya (Stenopsychidae trichopteran family). TE dS between *C. partellus* and all species all fall below the 0.5% quantile of the distribution of gene dS (Figure 2.4, Figure S2.3), strongly suggesting the presence of this element in the various species is due to HT. IIV6 was originally obtained from a *Chilo fumerana* moth and is known to be able to infect several other moth species (Fukaya and Nasu, 1966), in agreement with the possibility that this virus could serve as vector of HTT.

**Persistence of fly TEs during IIV6 replication**

Contrasting with the absence of *D. hydei* and *S. nonagrioides* TEs and the very low number of *C. partellus* TEs integrated into IIV6 genomes, we found 1,611, 598 and 433 TE-virus chimeras involving *D. melanogaster* TEs in IIV6 genomes purified from the three hosts, respectively. All *D. melanogaster* TEs recovered in these IIV6 populations were present in the parental IIV6 purified from S2 cells and the 6 most frequent TEs were also the most frequent ones in the

parental IIV6 population (Figure 2.6; Table S2.4). We verified using blastn that none of these *D. melanogaster* TEs are present in the genome of *D. hydei*, *S. nonagrioides* or *C. partellus*. Thus, we conclude that *D. melanogaster* TEs integrated into IIV6 genomes purified from these three hosts come from the parental IIV6 population purified from S2 cells. The various chimeric reads involving *D. melanogaster* TEs found in IIV6 purified from *D. hydei*, *S. nonagrioides* and *C. partellus* correspond to 344, ten and four different positions along the IIV6 genome, respectively (Figure 2.5). Consistent with persistence of *D. melanogaster* TEs in IIV6 purified from the three other hosts, 126, nine and three of these different insertion positions were shared with the parental IIV6 population purified from S2 cells, respectively. This implies that many TE-bearing IIV6 genomes persisted through one replication cycle and were again encapsidated in three different hosts. By contrast, 218, one, and one insertions found in IIV6 purified from *D. hydei*, *S. nonagrioides* and *C. partellus* were not found in the parental IIV6 purified from S2 cells. A majority (94% in *D. hydei*) of these insertions are supported by one or less than ten reads in IIV6 purified from *D. hydei*, *S. nonagrioides* and *C. partellus*. Thus, insertions not shared with the parental IIV6 population purified from S2 cells may in fact have been present at very low frequency in this population but not sequenced.

**Persistence of TEs integrated in AcMNPV genomes**

Finally, we sought to assess whether as observed for IIV6, TEs integrated into AcMNPV genomes can be recovered after one replication cycle. For this, we infected larvae of the mediterranean corn borer (*S. nonagrioides*) with a viral population called "G0" purified from *T. ni* larvae (Chateigner et al., 2015) and known to contain thousands of TE copies belonging to at least 30 families and 13 superfamilies (Gilbert et al., 2016). Our search for TEs integrated into the AcMNPV population purified from *S. nonagrioides* (here called "G1") unveiled two TEs that were present in the G0 population, called Tni_contig_27 (piggybac) and Tni_contig_21 (mariner) in Gilbert et al. (2016) (Table S2.1). A blast search of these two TEs onto the *S. nonagrioides* genome did not reveal any significant hit and a PCR screening using three primer sets designed to amplify three regions of these elements supported their absence from *S. nonagrioides*. Thus, the presence of Tni_contig_27 and Tni_contig_21 in AcMNPV genomes purified from *S. nonagrioides* cannot be due to transposition of these TEs from the *S. nonagrioides* genome. We conclude that the two TEs were carried over from the G0 population, confirming that as observed for IIV6, some TEs integrated into AcMNPV genomes can persist over at least one replication cycle. Interestingly, aside from the shared presence of Tni_contig_27 and Tni_contig_21, the TE landscape of the AcMNPV G1 population markedly

differed from that of the G0. Indeed, none of the 18 other TEs present in the G0 population were recovered after replication on *S. nonagrioides* (Figure 2.7). This is in stark contrast with IIV6, whereby TE contents and frequencies were relatively similar between the populations purified from *D. melanogaster* S2 cells and *D. hydei*.

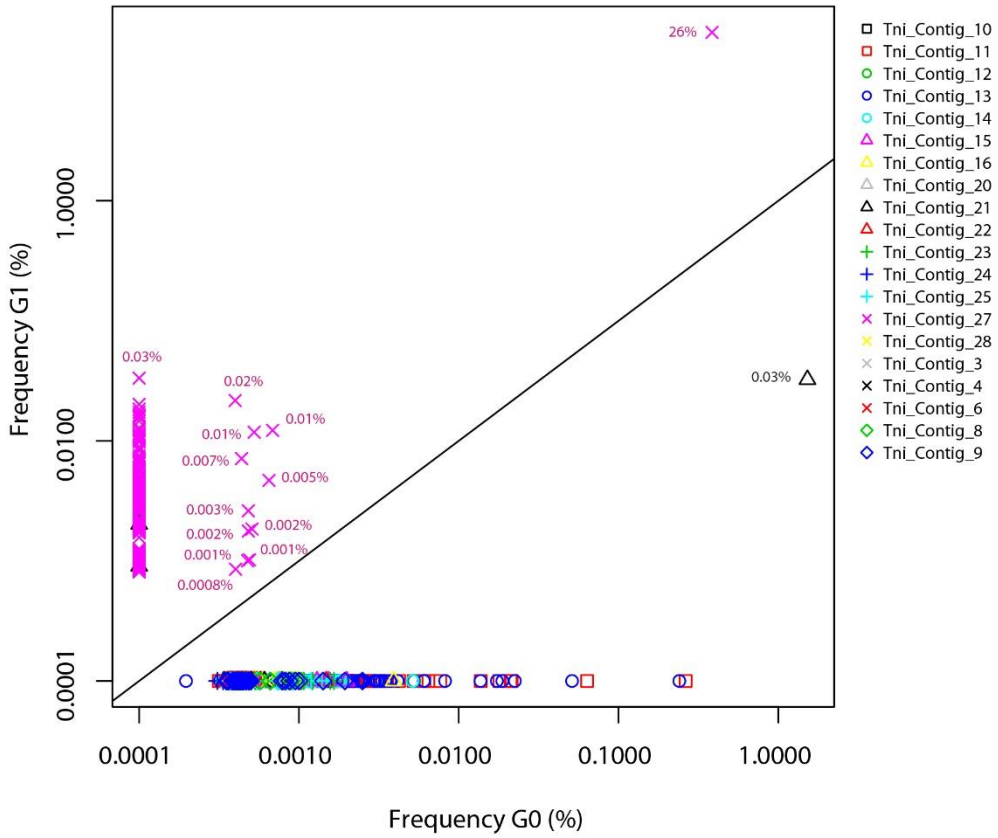



**Figure 2.7. Frequency of TE insertion points in the G0 and G1 populations.** Frequency is expressed in logarithm. Only two TEs, Tni_Contig_27 and Tni_Contig_21 have shared insertion points between both populations. Exact frequency in the G1 population corresponding at shared insertion points are shown near dots on the plot.

### AcMNPV-to-AcMNPV transposition in moths

A striking feature of the AcMNPV G1 population purified from *S. nonagrioides* is that 26.33% of its genomes carried at least one TE insertion. This high frequency was mainly explained by Tni_contig_27, as the frequency of AcMNPV genomes bearing other TEs is only 0.04%. Thus, starting from 0.38% in the G0 population purified from *T. ni* larvae, the frequency of genomes carrying a Tni_contig_27 insertion underwent a 69-fold increase during replication in *S. nonagrioides*. Similarly, the number of Tni_contig_27 insertions was much higher in the G1 (n

= 417) than in the parental G0 (n = 17) AcMNPV population (Figure 2.8). One could argue that all Tni_contig_27 insertions detected in G1 were in fact present at low frequency but were not sequenced in the G0 population. Fluctuations in insertion frequencies driven by genetic drift could then be invoked to explain the higher content in Tni_contig_27 insertion in the G1 versus G0 population. However, the G0 population was sequenced at a higher depth (124,221 X) than the G1 population (82,103 X), which does not support the hypothesis according to which many Tni_contig_27 insertions would have been missed in the G0 AcMNPV population. Furthermore, we tested whether the change in Tni_contig_27 content between the two AcMNPV populations could be explained by genetic drift. Under such a scenario the frequency of an insertion would be expected to have increased or decreased with the same probability (0.5) during replication in *S. nonagrioides* larvae. Using a binomial test, we found that the overall fluctuation in frequency of Tni_contig_27 insertions between the G0 and G1 population cannot be explained by drift (p-value = $2.2 \times 10^{-17}$). In other words, assuming that all Tni_contig_27 insertions found in the G1 population were in fact present in the G0 population would imply a much higher probability for their frequency to augment than to decrease, in the G1 population. Though inconsistent with drift alone, such a global increase could in principle be observed if all or most Tni_contig_27 insertions were beneficial to the virus, leading to an increase in their frequency during replication in *S. nonagrioides* through positive selection. Such a scenario, whereby several hundreds of Tni_contig_27 insertions would be beneficial to the virus irrespective of their location along the AcMNPV genome, is highly unlikely. In fact, current evidence strongly suggests that the vast majority of TE insertions are deleterious to viruses as they very rarely reach high frequencies during viral replication in natural hosts (Gilbert et al., 2016; Gilbert and Cordaux, 2017). This is best illustrated by the rare occurrence of TEs in consensus viral genomes (Filée, 2018; Gilbert and Cordaux, 2017; Sun et al., 2015). Thus, the nearly 25-fold increase in the number of Tni_contig_27 insertions observed in the G1 is unlikely to be explained by fluctuation in insertion frequency. Instead, the Tni_contig_27 most likely transposed from AcMNPV genomes to other AcMNPV genomes during replication in *S. nonagrioides* larvae.

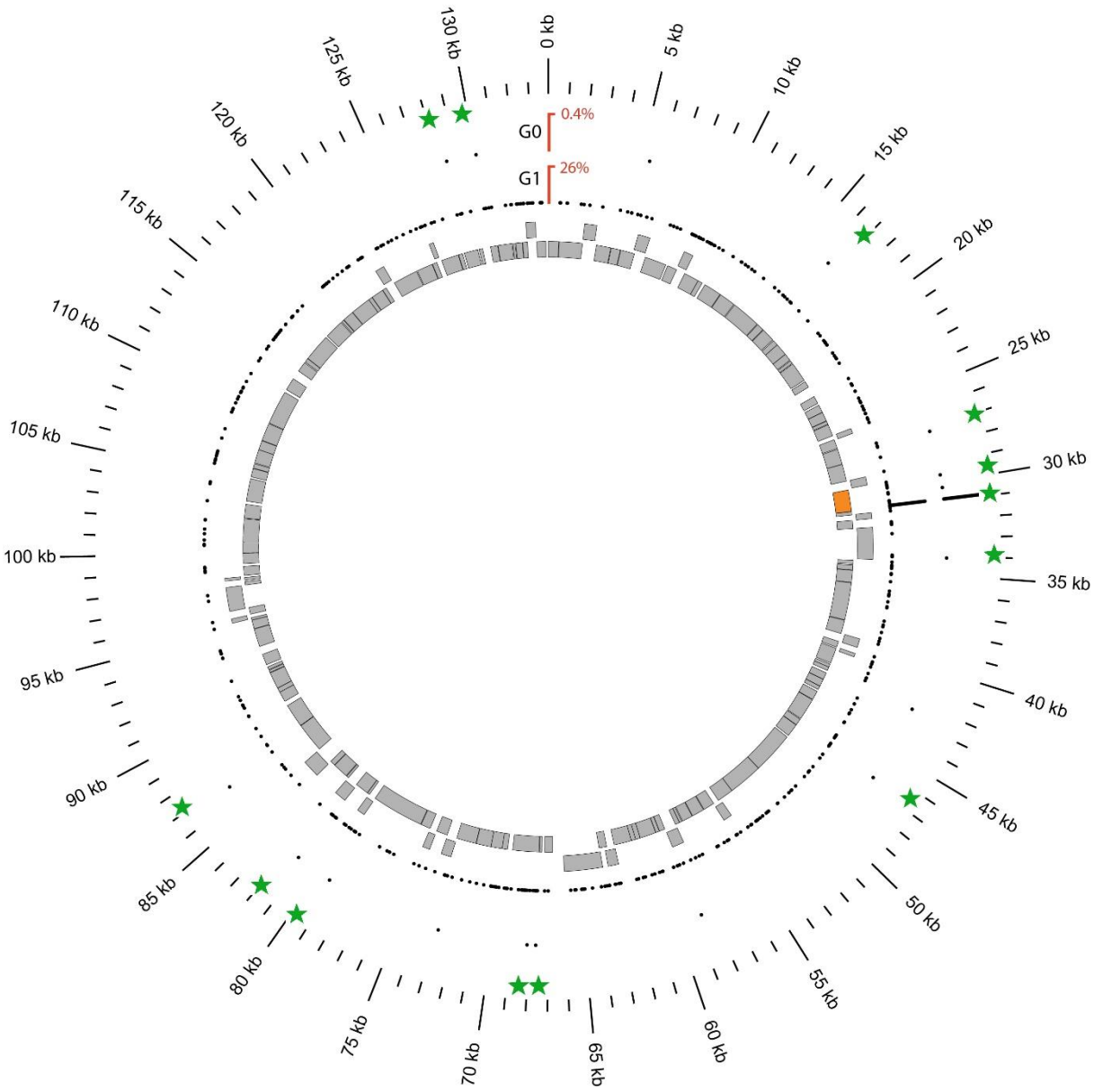



**Figure 2.8. Insertion map of Tni_contig_27 in G0 and G1 populations along the AcMNPV genome.** Green stars represent the thirteen shared insertion points between both AcMNPV populations. Grey rectangles represent genes along the viral genome. The orange rectangle represents the Ac-GTA gene. Most of Tni_Contig_27 insertions fall into the Ac-GTA gene for both AcMNPV populations although insertion points are scattered all along the viral genome.

### High frequency of a Tni_contig_27 insertion in the AcMNPV *GTA* gene

While virus-to-virus transposition likely explains the increase in number of Tni_contig_27 insertions in the AcMNPV G1 population, it does not account for the overall increase in AcMNPV genomes bearing a Tni_contig_27 insertion. This frequency increased from 0.38% in the G0 to 26.29% in the G1 population. In fact, the vast majority of TE-virus chimeras involving Tni_contig_27 in the G1 population map to a single insertion site, located at position 30918 of the WP10 AcMNPV genome, in the global transactivator (*Ac-GTA*) gene (Figure 2.8).

This insertion, together with 12 other Tni_contig_27 insertions, is present in the parental G0 population but at a much lower frequency (0.38%; Figure 2.7). The frequency increase of this insertion during replication in *S. nonagrioides* larvae may be due to repeated integration of Tni_contig_27 at the exact same site and/or to preferential replication and/or preferential encapsidation of viral genomes bearing this insertion. Our earlier analysis of 10 AcMNPV populations obtained after infection of *T.ni* larvae with the G0 population did not reveal any frequency increase of this insertion (Gilbert et al. 2016). Repeated integration of Tni_contig_27 at position 30918 of the AcMNPV genome may have occurred but it is unlikely to solely explain the sharp increase in frequency of this insertion during replication in *S. nonagrioides* larvae. Instead, the frequency of this insertion may have mainly increased through replication of the viral genomes bearing it. It is noteworthy that *Ac-GTA* is known to be involved in regulation of transcription, DNA recombination and repair, chromatin unwinding and other functions (Rohrmann, 2019). Deletion of this gene from the Bombyx mori NPV only led to mild effects on the infection outcome (i.e. delayed killing time), suggesting that it may not be essential for viral replication in some contexts (Katsuma et al., 2008). We reasoned that the increase in frequency of the Tni_contig_27 insertion in this gene may be due to low functional constraints inducing relaxation of purifying selection. If true, other mutations inactivating this gene may also be expected to have increased in frequency during replication in *S. nonagrioides*. However, we found no evidence of other high frequency TE insertions or high frequency indels in *Ac-GTA* (Figure 2.8, Figure S2.4). Thus, the strong increase in frequency of the Tni_contig_27 insertion in *Ac-GTA* is unlikely to be due to the absence of functional constraints acting on this gene in our experiment. The possibility that this insertion increased in frequency because of its positive effect on viral replication cannot be excluded at this stage. Another possibility worth testing in future experiments is that genomes bearing this insertion may have a higher propensity to be encapsidated compared to other genomes.

**Origin and horizontal transfer of transposable elements found in AcMNPV**

Though our earlier study identified Tni_conti_27 and Tni_contig_21 into AcMNPV genomes purified from *T. ni* larvae, we were, at the time, unable to trace the origin of these TEs because no WGS was available for *T. ni* (Gilbert et al. 2016). A blastn search for these TEs against the two available *T. ni* genomes (Chen et al., 2019; Talsania et al., 2019) revealed no hit. This result suggested the presence of Tni_contig_27 and Tni_contig_21 in the G0 AcMNPV population purified from *T. ni* larvae did not result from transposition of TE copies located in the *T. ni*

genome. Instead, the two TEs must originate from an earlier replication cycle of AcMNPV on another host.

In addition to Tni_contig_27 and Tni_contig_21, we found three *S. nonagrioides* DNA TEs (one piggybac and 2 sola) involved in 5, 6 and 5 TE-virus chimeras mapping at 4, 6 and 5 different positions along the AcMNPV genome, respectively (Table S2.1). Thus, TEs became integrated into AcMNPV genomes not only through virus-to-virus transposition but also via transposition of copies located in the *S. nonagrioides* genome. In turn, this shows that TEs from yet another host can transpose into AcMNPV genomes during the course of an infection, though at lower rates than previously reported in *T. ni* or *S. exigua*.

Finally, as for other TEs found integrated into viral genomes, we found evidence that at least some of the TEs integrated into AcMNPV genomes were involved in HT. For example, we found one TE copy 95.5% identical to the *S. nonagrioides* piggybac (Scf7180002741278) over 99% of its length in the WGS of *C. croceus* (Pieridae), which diverged 114 myrs ago from *S. nonagrioides*. There are also two copies 95% identical to this element in the WGS of the noctuid *S. frugiperda*, which diverged from *S. nonagrioides* 49 myrs ago. Furthermore, TEs >98% identical to Tni_contig_27 over >97% of its length were found in *S. exigua* and in *M. configurata*. Analyses of TE versus gene synonymous distances between these species again clearly indicate that these high levels of similarity are incompatible with vertical inheritance and that these TEs underwent one or more events of HT in lepidopterans (Figure 2.4 and Figure S2.3).

**Discussion**

The diversity and frequency of TEs integrated in baculovirus genomes had so far been assessed using ultra-deep sequencing in 11 AcMNPV populations purified from *T. ni* larvae in Gilbert et al. (2016) as well as in 10 and 1 AcMNPV populations purified from *S. exigua* larvae in Gilbert et al. (2016) and Loiseau et al. (2020), respectively. In these earlier studies, 29 and 40 different TEs were found integrated in AcMNPV replicated on *T. ni* and *S. exigua* respectively, with frequencies ranging from 1.1 to 14.3% of viral genomes carrying at least one host sequence depending on the dataset. Here, we show that TEs from another host (*S. nonagrioides*) can transpose into AcMNPV, though at lower rates than in *T. ni* and *S. exigua*. We further report measurable frequencies of viral genomes carrying TEs in another NPV (AgseNPV) replicated in *A. ipsilon* cells as well as in two granuloviruses (AgseGV and CpGV) replicated in *A. segetum* and *C. pomonella* larvae, respectively. We also show that large numbers of *D. melanogaster* TEs, as well as a few *C. partellus* TEs, can integrate in genomes of the IIV6 iridovirus during replication in *D. melanogaster* cells or in whole flies and moths. Interestingly, earlier works (Bartolomé et al., 2009) and our own inferences of HTT based on comparisons of synonymous distances between TEs and host genes show that many TEs we found integrated in viral genomes have been transmitted through HT between insect species. Altogether, these results extend our knowledge of host-virus systems for which TEs are able to jump into viral genomes. They are in line with previous works supporting a role for large dsDNA viruses as possible vectors of HTT between eukaryotes (Carstens, 1987; Fraser et al., 1995; Gilbert and Cordaux, 2017; Jehle et al., 1998; Miller and Miller, 1982; Sun et al., 2015).

While this study demonstrates the capacity of some viruses to receive and carry TEs, it also reveals important variation in occurrence and frequency of TE integration into viral genomes across host-virus systems. No TE integration was detected in IIV31 iridoviruses purified from two pillbug species (*A. vulgare* and *P. dilatatus*), confirming an earlier report in *A. vulgare* (Loiseau et al. 2020). While we retrieved many *D. melanogaster* TEs and a few from *C. partellus* in IIV6, no TE from *D. hydei* or *S. nonagrioides* transposed into genomes of this virus. We found moth TEs integrated in AgseNPVs and CpGVs purified from only one out of six and four out of 15 viral strains, respectively. Thus, while baculoviruses and iridoviruses have the capacity to encapsidate and shuttle TE-bearing viral genomes, transposition into viral genomes does not systematically occur during infection. It may not even occur at all in some host-virus systems. We cannot exclude that some TE insertions may have been missed, especially in some relatively low sequencing-depth datasets (e.g., lower than 1,000 X for six baculoviruses), or

because our TE library could not include all host TEs due to unavailable WGS (e.g., for *C. partellus*). Yet, we found no correlation between sequencing depth and frequency of viral genomes carrying a TE (Pearson correlation test, p-value=0.82). Also, we found no TE insertion in several viral genomes ultra-deeply sequenced that were purified from hosts with available WGS (i.e., *A. vulgare*, *S. nonagrioides*, *D. hydei*; Table 2.1). This strongly suggests the absence of TEs in these datasets has biological underpinnings.

It is likely that multiple host and virus factors are involved in shaping the numbers of host TEs found integrated in viral genomes. On the host side, the level of TE activity may be involved. For example, the absence of *A. vulgare*, *S. nonagrioides* and *D. hydei* TEs in IIV31, AcMNPV and IIV6 genomes, respectively, could be due to the absence of currently active TEs in the genome of the three species. However, the TE landscape of *A. vulgare* and *D. hydei* (and TE expression patterns for *A. vulgare*) does not support this hypothesis, as a relatively large fraction of TE copies are nearly identical to their cognate family consensus sequence, which is highly suggestive of current activity (Becking et al., 2020); Figure S2.5). In addition, the finding of three *S. nonagrioides* TEs integrated into AcMNPV genomes provides direct evidence that some TEs are indeed active in this moth. On the virus side, we are not aware of any major difference between baculovirus and iridovirus replication that could explain varying propensities of these viruses to receive and carry hosts TEs. In particular, replication involves transportation of viral DNA into the nucleus for both types of viruses (Rohrmann, 2019; Williams et al., 2005). However, among baculoviruses, the frequency of viral genomes carrying a TE is significantly lower for GVs than for NPVs (Wilcoxon test, W=0; p-value=$3.4*10^{-4}$). Complementation between genomes may be more likely to occur in NPVs, which viral particles contain many capsids (and thus genomes), than in GVs which encapsidate only one genome per viral particle (Rohrmann, 2019). Thus, defective TE-bearing NPV genomes may be more likely to be replicated than defective GVs, which may explain the trend we observe. In fact, NPVs are the only known viruses to be transmitted as multi-capsid particles, while all other large dsDNA viruses, including iridoviruses, are transmitted as mono-capsid particles. This may in part explain why no *A. vulgare* and *P. dilatatus* TEs and only few *C. partellus* TEs were found in iridovirus genomes. If true, NPVs would stand out as being more efficient HTT vectors than other viruses. In turn, given that NPVs mainly infect lepidopterans (Goulson, 2003), their high propensity to shuttle TEs may explain why lepidopterans are seemingly more prone to HTT than other insects, as proposed by Reiss et al. (2019). Yet, the frequency of viral genomes carrying a TE is as high as or even higher in IIV6 purified from *Drosophila* S2 cells or whole

flies (from 3.6 to 6%) than in NPVs, which contradicts the view that the mono-capsid nature of iridoviruses would dampen their ability to carry TEs. In this context, it is noteworthy that IIV6, which was originally sampled from the moth *C. fumerana* (Fukaya and Nasu, 1966) is not known to naturally infect flies. Thus, this artificial host-virus system involving a passage in cell culture and infection of whole flies by injection may have created conditions favorable to transposition and persistence of fly TEs in IIV6. In any case, our results call for future studies specifically dedicated to decipher the relative contribution of host and virus factors involved in shaping transposition of host TEs into viral genomes, as well as TE persistence in viral populations over multiple replication rounds.

Another important finding of this study is that it provides direct evidence that TEs from a given insect species integrated into genomes of a viral population can be recovered after subsequent purification of the same virus from another species. Our previous study showed that the number and frequency of TEs could be similar in a parental (G0) AcMNPV population replicated in one host (*T. ni*) and its daugther populations (G10) separated from the G0 by ten successive infection cycles on another host (*S. exigua*). However, no TE was shared between the two populations (Gilbert et al. 2016). This result revealed a continuous dynamics of gain (via transposition) and loss (via purifying selection) of TEs in AcMNPV populations, and showed that the persistence of a given TE community in AcMNPV populations was less than 10 infection cycles. Here, we found *D. melanogaster* TEs integrated in IIV6 genomes purified from two moths (*C. partellus* and *S. nonagrioides*) and another fly (*D. hydei*). We also recovered TEs present in the AcMNPV G0 dataset (and absent from the *S. nonagrioides* genome) in AcMNPV genomes purified from *S. nonagrioides* larvae. This provides direct evidence that TE-bearing viral genomes can be encapsidated and shuttled between different host species and that TEs can persist in viral populations for longer than one infection cycle. In this context, it is noteworthy that while the frequency of TE-bearing genomes in the IIV6 population purified from *D. hydei* (2.64%) is close to that of the parental IIV6 population purified from S2 cells (3.56%), it dropped down to 0.31% and 0.15% in IIV6 purified from *C. partellus* and *S. nonagrioides,* respectively. Furthermore, while many TEs originating from the genome of S2 cells were recovered in IIV6 purified from *D. hydei*, *C. partellus* and *S. nonagrioides*, only two out of 31 TEs initially found in AcMNPV purified from *T. ni* persisted during replication of this virus in *S. nonagrioides*. Such differences in TE diversity and number of viral genomes bearing TE insertions observed after one replication cycle of the same starting viral population in different hosts may be due to host-virus interactions other than host-to-virus

transposition. It will be interesting to further evaluate the role played by the host in shaping TE content and frequency in viral population in more controlled viral replication assays.

Another interesting result of this study is that the mere fraction of TE copies present in AcMNPV genomes purified from *S. nonagrioides* were generated by virus-to-virus transposition. This is remarkable because it shows that virus-borne TEs can transpose in a new host, even if this new host is distantly related to the host from which the TEs originated (*S. nonagrioides* and *T. ni* diverged 60 Myrs ago). It confirms and extends earlier findings based on long read sequencing showing that TEs were integrated as full-length copies into AcMNPV genomes and were thus likely to be able to transpose from the virus genome to another DNA molecule (Loiseau et al. 2020). Our results further indicate that only some of the TEs found in a given viral population may undergo virus-to-virus transposition during subsequent infection cycles. Indeed, strong evidence of virus-to-virus transposition in AcMNPV purified from *S. nonagrioides* was found for only one TE (Tni_contig_27) out of the 31 present in the AcMNPV G0 population purified from *T. ni*. It is noteworthy that this TE is the second most frequent autonomous TE in the AcMNPV G0 population, being present in 0.38% of the AcMNPV genomes (Gilbert et al. 2016). This suggests that the likelihood for a TE to undergo virus-to-virus transposition may in part depend on its frequency in the infecting viral population. Perhaps more importantly, Tni_contig_27 is the only TE present in the G0 population for which a single insertion underwent a sharp increase in frequency during replication in *S. nonagrioides* larvae. We speculate that the increase of the Tni_contig_27 insertion located in the Ac-GTA gene may be independent of transposition. Instead, the frequency of this insertion likely increased through viral replication that may be driven by positive selection, or by a higher propensity of viral genomes bearing this insertion to be encapsidated. These two hypotheses are not mutually exclusive. In turn, the large number of Tni_contig_27 copies present in the viral population have allowed this TE to reach a rate of virus-to-virus transposition high enough to be detected by our approach. In other words, while many TEs may have transposed from viral genomes to other viral genomes at low rates in our experiments, we might have only detected virus-to-virus transposition for Tni_contig_27 because of the sharp increase in frequency of one of its insertions. In any case, many TEs, host and virus factors are likely to shape the likelihood of a TE to undergo virus-to-virus transposition. It will be interesting to design new experiments to unveil these factors and the ways they interact.

To conclude, our study extends our knowledge of animal host-virus systems in which TEs can transpose from host to viral genomes. It also directly shows that the genome of viruses carrying animal TEs can indeed be encapsidated into viral particles and be transported into another host where transposition of virus-borne TEs can occur. Another important observation we made is that of a single TE undergoing a dramatic (69-fold) frequency increase in a viral population (from 0.38 to >25% of viral genomes) in a short timeframe (a single infection cycle). Perhaps even more remarkably, our study provides direct evidence for virus-to-virus transposition of animal TEs in a baculovirus. This suggests that viruses may not only serve as launching platforms for these genomic symbionts to colonize naive cellular genomes (a hypothesis that still needs to be formally tested), but that they may also be viewed as an alternative niche in which TEs can persist through time and evolve under different constraints than in cellular hosts. This is reminiscent of the molecular symbiosis hypothesis proposed by Filée (2018) to characterize the intricate interactions occurring between various types of mobile genetic elements and giant viruses infecting single-cell eukaryotes.

## References


Alletti, G., Sauer, A., Weihrauch, B., Fritsch, E., Undorf-Spahn, K., Wennmann, J., Jehle, J., 2017. Using Next Generation Sequencing to Identify and Quantify the Genetic Composition of Resistance-Breaking Commercial Isolates of Cydia pomonella Granulovirus. Viruses 9, 250. https://doi.org/10.3390/v9090250

Bao, W., Kojima, K.K., Kohany, O., 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob. DNA 6, 11. https://doi.org/10.1186/s13100-015-0041-9

Bartolomé, C., Bello, X., Maside, X., 2009. Widespread evidence for horizontal transfer of transposable elements across Drosophila genomes. Genome Biol. 10, R22. https://doi.org/10.1186/gb-2009-10-2-r22

Becking, T., Gilbert, C., Cordaux, R., 2020. Impact of transposable elements on genome size variation between two closely related crustacean species. Anal. Biochem. 600, 113770. https://doi.org/10.1016/j.ab.2020.113770

Bourque, G., Burns, K.H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H.L., Macfarlan, T.S., Mager, D.L., Feschotte, C., 2018. Ten things you should know about transposable elements. Genome Biol. 19, 199. https://doi.org/10.1186/s13059-018-1577-z

Carstens, E.B., 1987. Identification and nucleotide sequence of the regions of Autographa californica nuclear polyhedrosis virus genome carrying insertion elements derived from Spodoptera frugiperda. Virology 161, 8–17. https://doi.org/10.1016/0042-6822(87)90165-6

Charif, D., Thioulouse, J., Lobry, J.R., Perri re, G., 2005. Online synonymous codon usage analyses with the ade4 and seqinR packages. Bioinformatics 21, 545–547. https://doi.org/10.1093/bioinformatics/bti037

Chateigner, A., Bézier, A., Labrousse, C., Jiolle, D., Barbe, V., Herniou, E., 2015. Ultra Deep Sequencing of a Baculovirus Population Reveals Widespread Genomic Variations. Viruses 7, 3625–3646. https://doi.org/10.3390/v7072788

Chebbi, M.A., Becking, T., Moumen, B., Giraud, I., Gilbert, C., Peccoud, J., Cordaux, R., 2019. The Genome of Armadillidium vulgare (Crustacea, Isopoda) Provides Insights into Sex Chromosome Evolution in the Context of Cytoplasmic Sex Determination. Mol. Biol. Evol. 36, 727–741. https://doi.org/10.1093/molbev/msz010

Chen, W., Yang, X., Tetreau, G., Song, X., Coutu, C., Hegedus, D., Blissard, G., Fei, Z., Wang, P., 2019. A high-quality chromosome-level genome assembly of a generalist herbivore, *Trichoplusia ni*. Mol. Ecol. Resour. 19, 485–496. https://doi.org/10.1111/1755-0998.12966

Cordaux, R., Batzer, M.A., 2009. The impact of retrotransposons on human genome evolution. Nat. Rev. Genet. 10, 691–703. https://doi.org/10.1038/nrg2640

Craig, N.L., Chandler, M., Gellert, M., Lambowitz, A.M., Rice, P.A., Sandmeyer, S.B. (Eds.), 2015. Mobile DNA III. ASM Press, Washington, DC, USA. https://doi.org/10.1128/9781555819217

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, 339–355.

Dotto, B.R., Carvalho, E.L., da Silva, A.F., Dezordi, F.Z., Pinto, P.M., Campos, T. de L., Rezende, A.M., Wallau, G. da L., 2018. HTT-DB: new features and updates. Database 2018. https://doi.org/10.1093/database/bax102

Eckwahl, M.J., Telesnitsky, A., Wolin, S.L., 2016. Host RNA Packaging by Retroviruses: A Newly Synthesized Story. mBio 7, e02025-02015. https://doi.org/10.1128/mBio.02025-15

Fan, J., Wennmann, J.T., Wang, D., Jehle, J.A., 2020. Single nucleotide polymorphism (SNP) frequencies and distribution reveal complex genetic composition of seven novel natural isolates of Cydia pomonella granulovirus. Virology 541, 32–40. https://doi.org/10.1016/j.virol.2019.11.016

Fan, J., Wennmann, J.T., Wang, D., Jehle, J.A., 2019. Novel Diversity and Virulence Patterns Found in New Isolates of Cydia pomonella Granulovirus from China. Appl. Environ. Microbiol. 86, e02000-19, /aem/86/2/AEM.02000-19.atom. https://doi.org/10.1128/AEM.02000-19

Filée, J., 2018. Giant viruses and their mobile genetic elements: the molecular symbiosis hypothesis. Curr. Opin. Virol. 33, 81–88. https://doi.org/10.1016/j.coviro.2018.07.013

Flynn, J.M., Hubley, R., Goubert, C., Rosen, J., Clark, A.G., Feschotte, C., Smit, A.F., 2020. RepeatModeler2 for automated genomic discovery of transposable element families. Proc. Natl. Acad. Sci. 117, 9451–9457. https://doi.org/10.1073/pnas.1921046117

Fraser, M.J., Brusca, J.S., Smith, G.E., Summers, M.D., 1985. Transposon-mediated mutagenesis of a baculovirus. Virology 145, 356–361. https://doi.org/10.1016/0042-6822(85)90172-2

Fraser, M.J., Cary, L., Boonvisudhi, K., Wang, H.G., 1995. Assay for movement of Lepidopteran transposon IFP2 in insect cells using a baculovirus genome as a target DNA. Virology 211, 397–407. https://doi.org/10.1006/viro.1995.1422

Fukaya, M., Nasu, S., 1966. A Chilo Iridescent Virus (CIV) from the Rice Stem Borer, Chilo suppressalis WALKER (Lepidoptera : Pyralidae). Appl. Entomol. Zool. 1, 69–72. https://doi.org/10.1303/aez.1.69

Ghoshal, K., Theilmann, J., Reade, R., Maghodia, A., Rochon, D., 2015. Encapsidation of Host RNAs by Cucumber Necrosis Virus Coat Protein during both Agroinfiltration and Infection. J. Virol. 89, 10748–10761. https://doi.org/10.1128/JVI.01466-15

Gilbert, C., Chateigner, A., Ernenwein, L., Barbe, V., Bézier, A., Herniou, E.A., Cordaux, R., 2014. Population genomics supports baculoviruses as vectors of horizontal transfer of insect transposons. Nat. Commun. 5, 3348. https://doi.org/10.1038/ncomms4348

Gilbert, C., Cordaux, R., 2017. Viruses as vectors of horizontal transfer of genetic material in eukaryotes. Curr. Opin. Virol. 25, 16–22. https://doi.org/10.1016/j.coviro.2017.06.005

Gilbert, C., Feschotte, C., 2018. Horizontal acquisition of transposable elements and viral sequences: patterns and consequences. Curr. Opin. Genet. Dev. 49, 15–24. https://doi.org/10.1016/j.gde.2018.02.007

Gilbert, C., Peccoud, J., Chateigner, A., Moumen, B., Cordaux, R., Herniou, E.A., 2016. Continuous Influx of Genetic Material from Host to Virus Populations. PLoS Genet. 12, e1005838. https://doi.org/10.1371/journal.pgen.1005838

Gilbert, C., Schaack, S., Pace, J.K., Brindley, P.J., Feschotte, C., 2010. A role for host-parasite interactions in the horizontal transfer of transposons across phyla. Nature 464, 1347–1350. https://doi.org/10.1038/nature08939

Goulson, D., 2003. Can Host Susceptibility to Baculovirus Infection be Predicted from Host Taxonomy or Life History? Environ. Entomol. 32, 61–70. https://doi.org/10.1603/0046-225X-32.1.61

Graillot, B., Berling, M., Blachere-López, C., Siegwart, M., Besse, S., López-Ferber, M., 2014. Progressive Adaptation of a CpGV Isolate to Codling Moth Populations Resistant to CpGV-M. Viruses 6, 5135–5144. https://doi.org/10.3390/v6125135

Gueli Alletti, G., Carstens, E.B., Weihrauch, B., Jehle, J.A., 2018. Agrotis segetum nucleopolyhedrovirus but not Agrotis segetum granulovirus replicate in AiE1611T cell line of Agrotisipsilon. J. Invertebr. Pathol. 151, 7–13. https://doi.org/10.1016/j.jip.2017.10.005

Gueli Alletti, G., Eigenbrod, M., Carstens, E.B., Kleespies, R.G., Jehle, J.A., 2017. The genome sequence of Agrotis segetum granulovirus, isolate AgseGV-DA, reveals a new Betabaculovirus species of a slow killing granulovirus. J. Invertebr. Pathol. 146, 58–68. https://doi.org/10.1016/j.jip.2017.04.008

Guo, X., Gao, J., Li, F., Wang, J., 2014. Evidence of horizontal transfer of non-autonomous Lep1 Helitrons facilitated by host-parasite interactions. Sci. Rep. 4, 5119. https://doi.org/10.1038/srep05119

Han, M.-J., Zhou, Q.-Z., Zhang, H.-H., Tong, X., Lu, C., Zhang, Z., Dai, F., 2016. iMITEdb: the genome-wide landscape of miniature inverted-repeat transposable elements in insects. Database 2016, baw148. https://doi.org/10.1093/database/baw148

Harrison, R.L., Lynn, D.E., 2008. New cell lines derived from the black cutworm, Agrotis ipsilon, that support replication of the A. ipsilon multiple nucleopolyhedrovirus and several group I nucleopolyhedroviruses. J. Invertebr. Pathol. 99, 28–34. https://doi.org/10.1016/j.jip.2008.02.015

Ivancevic, A.M., Kortschak, R.D., Bertozzi, T., Adelson, D.L., 2018. Horizontal transfer of BovB and L1 retrotransposons in eukaryotes. Genome Biol. 19, 85. https://doi.org/10.1186/s13059-018-1456-7

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, 369–379. https://doi.org/10.1006/viro.1995.1096
Jehle, J.A., Nickel, A., Vlak, J.M., Backhaus, H., 1998. Horizontal escape of the novel Tc1-like lepidopteran transposon TCp3.2 into Cydia pomonella granulovirus. J. Mol. Evol. 46, 215–224. https://doi.org/10.1007/pl00006296
Katoh, K., Standley, D.M., 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol. Biol. Evol. 30, 772–780. https://doi.org/10.1093/molbev/mst010
Katsuma, S., Fujii, T., Kawaoka, S., Shimada, T., 2008. Bombyx mori nucleopolyhedrovirus SNF2 global transactivator homologue (Bm33) enhances viral pathogenicity in B. mori larvae. J. Gen. Virol. 89, 3039–3046. https://doi.org/10.1099/vir.0.2008/004887-0
Kawamura, Y., Sanchez Calle, A., Yamamoto, Y., Sato, T.-A., Ochiya, T., 2019. Extracellular vesicles mediate the horizontal transfer of an active LINE-1 retrotransposon. J. Extracell. Vesicles 8, 1643214. https://doi.org/10.1080/20013078.2019.1643214
Kofler, R., Nolte, V., Schlötterer, C., 2015. Tempo and Mode of Transposable Element Activity in Drosophila. PLoS Genet. 11, e1005406. https://doi.org/10.1371/journal.pgen.1005406
Kumar, S., Stecher, G., Suleski, M., Hedges, S.B., 2017. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Mol. Biol. Evol. 34, 1812–1819. https://doi.org/10.1093/molbev/msx116
Kuraku, S., Qiu, H., Meyer, A., 2012. Horizontal Transfers of Tc1 Elements between Teleost Fishes and Their Vertebrate Parasites, Lampreys. Genome Biol. Evol. 4, 929–936. https://doi.org/10.1093/gbe/evs069
Loiseau, V., Herniou, E.A., Moreau, Y., Lévêque, N., Meignin, C., Daeffler, L., Federici, B., Cordaux, R., Gilbert, C., 2020. Wide spectrum and high frequency of genomic structural variation, including transposable elements, in large double-stranded DNA viruses. Virus Evol. 6, vez060. https://doi.org/10.1093/ve/vez060
Loreto, E.L.S., Carareto, C.M.A., Capy, P., 2008. Revisiting horizontal transfer of transposable elements in Drosophila. Heredity 100, 545–554. https://doi.org/10.1038/sj.hdy.6801094
Miller, D.W., Miller, L.K., 1982. A virus mutant with an insertion of a copia-like transposable element. Nature 299, 562–564. https://doi.org/10.1038/299562a0
Ono, R., Yasuhiko, Y., Aisaki, K.-I., Kitajima, S., Kanno, J., Hirabayashi, Y., 2019. Exosome-mediated horizontal gene transfer occurs in double-strand break repair during genome editing. Commun. Biol. 2, 57. https://doi.org/10.1038/s42003-019-0300-2
O'Reilly, D.R., Miller, L., Luckow, V.A., 1992. Baculovirus expression vectors: a laboratory manual. W.H. Freeman, New York.
Peccoud, J., Cordaux, R., Gilbert, C., 2018. Analyzing Horizontal Transfer of Transposable Elements on a Large Scale: Challenges and Prospects. BioEssays 40, 1700177. https://doi.org/10.1002/bies.201700177
Peccoud, J., Loiseau, V., Cordaux, R., Gilbert, C., 2017. Massive horizontal transfer of transposable elements in insects. Proc. Natl. Acad. Sci. U. S. A. 114, 4721–4726. https://doi.org/10.1073/pnas.1621178114
R Core Team., 2019. R: A Language and Environment for Statistical Computing.
Reiss, D., Mialdea, G., Miele, V., de Vienne, D.M., Peccoud, J., Gilbert, C., Duret, L., Charlat, S., 2019. Global survey of mobile DNA horizontal transfer in arthropods reveals Lepidoptera as a prime hotspot. PLoS Genet. 15, e1007965. https://doi.org/10.1371/journal.pgen.1007965
Rohrmann, G.F., 2019. Baculovirus Molecular Biology, 4th ed. National Center for Biotechnology Information (US), Bethesda (MD).
Routh, A., Domitrovic, T., Johnson, J.E., 2012. Host RNAs, including transposons, are encapsidated by a eukaryotic single-stranded RNA virus. Proc. Natl. Acad. Sci. U. S. A. 109, 1907–1912. https://doi.org/10.1073/pnas.1116168109
Schaack, S., Gilbert, C., Feschotte, C., 2010. Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. Trends Ecol. Evol. 25, 537–546. https://doi.org/10.1016/j.tree.2010.06.001

Silva, J.C., Loreto, E.L., Clark, J.B., 2004. Factors that affect the horizontal transfer of transposable elements. Curr. Issues Mol. Biol. 6, 57–71.
Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., Zdobnov, E.M., 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinforma. Oxf. Engl. 31, 3210–3212. https://doi.org/10.1093/bioinformatics/btv351
Sparks, W., Li, H., Bonning, B., 2008. Protocols for Oral Infection of Lepidopteran Larvae with Baculovirus. J. Vis. Exp. JoVE. https://doi.org/10.3791/888
Suh, A., Witt, C.C., Menger, J., Sadanandan, K.R., Podsiadlowski, L., Gerth, M., Weigert, A., McGuire, J.A., Mudge, J., Edwards, S.V., Rheindt, F.E., 2016. Ancient horizontal transfers of retrotransposons between birds and ancestors of human pathogenic nematodes. Nat. Commun. 7, 11396. https://doi.org/10.1038/ncomms11396
Sun, C., Feschotte, C., Wu, Z., Mueller, R.L., 2015. DNA transposons have colonized the genome of the giant virus Pandoravirus salinus. BMC Biol. 13, 38. https://doi.org/10.1186/s12915-015-0145-1
Szak, S.T., Pickeral, O.K., Makalowski, W., Boguski, M.S., Landsman, D., Boeke, J.D., 2002. Molecular archeology of L1 insertions in the human genome. Genome Biol. 3, research0052. https://doi.org/10.1186/gb-2002-3-10-research0052
Talsania, K., Mehta, M., Raley, C., Kriga, Y., Gowda, S., Grose, C., Drew, M., Roberts, V., Cheng, K.T., Burkett, S., Oeser, S., Stephens, R., Soppet, D., Chen, X., Kumar, P., German, O., Smirnova, T., Hautman, C., Shetty, J., Tran, B., Zhao, Y., Esposito, D., 2019. Genome Assembly and Annotation of the Trichoplusia ni Tni-FNL Insect Cell Line Enabled by Long-Read Technologies. Genes 10, 79. https://doi.org/10.3390/genes10020079
Tcherepanov, V., Ehlers, A., Upton, C., 2006. Genome Annotation Transfer Utility (GATU): rapid annotation of viral genomes using a closely related reference genome. BMC Genomics 7, 150. https://doi.org/10.1186/1471-2164-7-150
Telesnitsky, A., Wolin, S.L., 2016. The Host RNAs in Retroviral Particles. Viruses 8. https://doi.org/10.3390/v8080235
van Niel, G., D'Angelo, G., Raposo, G., 2018. Shedding light on the cell biology of extracellular vesicles. Nat. Rev. Mol. Cell Biol. 19, 213–228. https://doi.org/10.1038/nrm.2017.125
Wallau, G.L., Ortiz, M.F., Loreto, E.L.S., 2012. Horizontal Transposon Transfer in Eukarya: Detection, Bias, and Perspectives. Genome Biol. Evol. 4, 801–811. https://doi.org/10.1093/gbe/evs055
Walsh, A.M., Kortschak, R.D., Gardner, M.G., Bertozzi, T., Adelson, D.L., 2013. Widespread horizontal transfer of retrotransposons. Proc. Natl. Acad. Sci. U. S. A. 110, 1012–1016. https://doi.org/10.1073/pnas.1205856110
Waterhouse, R.M., Seppey, M., Simão, F.A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E.V., Zdobnov, E.M., 2018. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. Mol. Biol. Evol. 35, 543–548. https://doi.org/10.1093/molbev/msx319
Wennmann, J.T., Fan, J., Jehle, J.A., 2020. Bacsnp: Using Single Nucleotide Polymorphism (SNP) Specificities and Frequencies to Identify Genotype Composition in Baculoviruses. Viruses 12, 625. https://doi.org/10.3390/v12060625
Williams, T., Barbosa-Solomieu, V., Chinchar, V.G., 2005. A Decade of Advances in Iridovirus Research, in: Advances in Virus Research. Elsevier, pp. 173–248. https://doi.org/10.1016/S0065-3527(05)65006-3
Zhang, H.-H., Peccoud, J., Xu, M.-R.-X., Zhang, X.-G., Gilbert, C., 2020. Horizontal transfer and evolution of transposable elements in vertebrates. Nat. Commun. 11, 1362. https://doi.org/10.1038/s41467-020-15149-4
Zingg, D., Züger, M., Bollhalder, F., Andermatt, M., 2011. Use of resistance overcoming CpGV isolates and CpGV resistance situation of the codling moth in Europe seven years after the first discovery of resistance to CpGV-M 3.

# Supplementary data

**Table S2.1. Characteristics of transposable elements found integrated into viral genomes.** *numbers in brackets correspond to replicates. **non-LTR: non-LTR retrotransposons; LTR: LTR retrotransposons. ***Most LTR TEs appear as two separate sequences in RepBase, one corresponding to the LTR region of the element and the other correspond to the internal region of the element. In such case, we report only the length of the sequence that has gathered the higher number of chimeric reads, which is always the LTR sequence, hence the many TEs with lengths lower than 1000 bp. ****For LTR retrotransposons appearing as two separate sequences in RepBase, the number of chimeric reads mapping internally to each TE was assessed using both the LTR and the internal portion of the elements.

| Host | Virus* | TE name | Type of TE** | TE length (bp)*** | Number of chimeric reads at the extremities of TEs | Number of chimeric reads at the extremities of TEs without PCR duplicates | Number of chimeric reads internal to the TEs without PCR duplicates**** | Frequency of viral genomes carrying each type of TE | Overall frequency of viral genomes carrying a TE | Number of insertion points |
|---|---|---|---|---|---|---|---|---|---|---|
| Agrotis segetum larvae | AgseGV-DA | Agrotis_segetum_Sola | Autonomous DNA | 4466 | 33 | 33 | 1 | 0.30 | 0.30 | 28 |
| Agrotis ipsilon cells | AgseNPV-pp0 | / | / | / | | / | / | / | / | / |
| | AgseNPV-pp1 | / | / | / | | / | / | / | / | / |
| | AgseNPV-pp3 | / | / | / | | / | / | / | / | / |
| | AgseNPV-pp5 | / | / | / | | / | / | / | / | / |
| | AgseNPV-pp7 | Agrotis_segetum_PiggyBac | Autonomous DNA | 2468 | 81 | 49 | 0 | 3.07 | 3.40 | 5 |
| | | Agrotis_segetum_Sola | Autonomous DNA | 4466 | 6 | 6 | 0 | 0.23 | | 6 |
| | | Plutella_xylostella_MITE | Non-autonomous DNA | 336 | 3 | 3 | 0 | 0.12 | | 2 |
| | AgseNPV-pp10 | / | / | / | | / | / | / | / | / |
| Cydia pomonella larvae | CpGV-006 | / | / | / | | / | / | / | / | / |
| | CpGV-ALE | / | / | / | | / | / | / | / | / |
| | CpGV-E2 | Cydia_pomonella_PiggyBac | Autonomous DNA | 4025 | 6 | 6 | 0 | 0.01 | 0.01 | 6 |
| | CpGV-I07 | / | / | / | | / | / | / | / | / |
| | CpGV-I12 | Cydia_pomonella_PiggyBac | Autonomous DNA | 4025 | 5 | 5 | 0 | 0.009 | 0.009 | 5 |
| | CpGV-JQ | / | / | / | | / | / | / | / | / |
| | CpGV-KS1 | / | / | / | | / | / | / | / | / |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | CpGV-KS2 | Cydia_pomonella_rnd-5_family-1654#DNA/TcMar-Tc1 | Autonomous DNA | 1653 | 3 | 3 | 0 | 0.20 | 0.20 | 2 |
| | CpGV-M | Cydia_pomonella_PiggyBac | Autonomous DNA | 4025 | 13 | 12 | 0 | 0.02 | 0.02 | 10 |
| | CpGV-R5 | / | / | / | | / | / | / | / | / |
| | CpGV-S | Cydia_pomonella_PiggyBac | Autonomous DNA | 4025 | 5 | 5 | 0 | 0.09 | 0.56 | 3 |
| | | Cydia_pomonella_SHALINE | Non-autonomous non-LTR | 419 | 25 | 25 | 0 | 0.47 | | 2 |
| | CpGV-V15 | / | / | / | 0 | 0 | / | / | / | / |
| | CpGV-WW | / | / | / | 0 | 0 | / | / | / | / |
| | CpGV-ZY | / | / | / | 0 | 0 | / | / | / | / |
| | CpGV-ZY2 | / | / | / | 0 | 0 | / | / | / | / |
| Armadillidium vulgare | IIV31 (1) | / | / | / | 0 | 0 | / | / | / | / |
| | IIV31 (2) | / | / | / | 0 | 0 | / | / | / | / |
| | IIV31 (3) | / | / | / | 0 | 0 | / | / | / | / |
| Porcellio dilatatus | IIV31 (1) | / | / | / | 0 | 0 | / | / | / | / |
| | IIV31 (2) | / | / | / | 0 | 0 | / | / | / | / |
| | IIV31 (3) | / | / | / | 0 | 0 | / | / | / | / |
| Drosophila melanogaster S2 cells | IIV6 | MDG1 | Autonomous LTR | 442 | 2917 | 2296 | 75 | 1.49 | 3.56 | 1397 |
| | | DM297 | Autonomous LTR | 414 | 1499 | 1237 | 22 | 0.77 | | 334 |
| | | DM176 | Autonomous LTR | 7439 | 760 | 659 | 42 | 0.40 | | 234 |
| | | IDEFIX | Autonomous LTR | 594 | 786 | 484 | 12 | 0.40 | | 71 |
| | | TRANSPAC | Autonomous LTR | 330 | 496 | 451 | 8 | 0.26 | | 336 |
| | | BLOOD | Autonomous LTR | 399 | 414 | 378 | 15 | 0.21 | | 301 |
| | | MICROPIA | Autonomous LTR | 476 | 31 | 30 | 4 | 0.02 | | 25 |
| | | BEL | Autonomous LTR | 361 | 17 | 21 | 4 | 0.009 | | 14 |
| | | STALKER2 | Autonomous LTR | 424 | 22 | 21 | 1 | 0.01 | | 16 |
| | | DMRT1B | Autonomous LTR | 428 | 12 | 10 | 6 | 0.006 | | 9 |
| | | ROO | Autonomous LTR | 276 | 10 | 18 | 9 | 0.005 | | 10 |
| | | COPIA_DM | Autonomous LTR | 267 | 3 | 11 | 8 | 0.001 | | 4 |
| | | LTRMDG3_DM | Autonomous LTR | 398 | 5 | 4 | 0 | 0.003 | | 4 |
| | | GYPSY3 | Autonomous LTR | 408 | 10 | 7 | 0 | 0.005 | | 6 |
| Drosophila melanogaster flies | IIV6 (1) | GYPSY6 | Autonomous LTR | 407 | 8299 | 6687 | 92 | 3.17 | 6.01 | 1739 |
| | | MDG1 | Autonomous LTR | 442 | 3040 | 2080 | 26 | 1.16 | | 1042 |
| | | DM297 | Autonomous LTR | 414 | 1584 | 1096 | 10 | 0.61 | | 268 |
| | | IDEFIX | Autonomous LTR | 594 | 1062 | 503 | 2 | 0.41 | | 49 |
| | | DM176 | Autonomous LTR | 7439 | 607 | 486 | 8 | 0.24 | | 152 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | TRANSPAC | Autonomous LTR | 330 | 557 | 484 | 2 | 0.21 | | 289 |
| | | BLOOD | Autonomous LTR | 399 | 548 | 457 | 4 | 0.21 | | 271 |
| | | Gypsy1-LTR_DM | Autonomous LTR | 408 | 142 | 124 | 7 | 0.05 | | 94 |
| | | DMRT1B | Autonomous LTR | 5183 | 55 | 50 | 27 | 0.02 | | 40 |
| | | GYPSY3 | Autonomous LTR | 398 | 50 | 41 | 0 | 0.02 | | 39 |
| | | MICROPIA | Autonomous LTR | 476 | 40 | 40 | 3 | 0.02 | | 17 |
| | | DMLTR5 | Autonomous LTR | 266 | 29 | 27 | 0 | 0.01 | | 22 |
| | | STALKER2 | Autonomous LTR | 424 | 19 | 22 | 3 | 0.007 | | 12 |
| | | GYPSY | Autonomous LTR | 482 | 18 | 17 | 2 | 0.007 | | 14 |
| | | BEL | Autonomous LTR | 361 | 14 | 17 | 7 | 0.005 | | 8 |
| | | ROO | Autonomous LTR | 428 | 10 | 20 | 12 | 0.004 | | 8 |
| | | COPIA_DM | Autonomous LTR | 276 | 7 | 9 | 4 | 0.003 | | 5 |
| | | BLASTOPIA | Autonomous LTR | 275 | 6 | 10 | 6 | 0.002 | | 4 |
| | | Gypsy-7_Dsim | Autonomous LTR | 398 | 6 | 6 | 1 | 0.002 | | 5 |
| | | STALKER4 | Autonomous LTR | 402 | 5 | 8 | 4 | 0.002 | | 3 |
| | | BURDOCK | Autonomous LTR | 275 | 4 | 4 | 0 | 0.002 | | 3 |
| | | DM412 | Autonomous LTR | 481 | 4 | 5 | 1 | 0.002 | | 4 |
| Drosophila melanogaster flies | IIV6 (2) | GYPSY6 | Autonomous LTR | 407 | 7486 | 5651 | 132 | 2.32 | 4.77 | 1578 |
| | | MDG1 | Autonomous LTR | 420 | 3271 | 2370 | 41 | 1.01 | | 1153 |
| | | DM297 | Autonomous LTR | 414 | 1669 | 1139 | 15 | 0.52 | | 288 |
| | | IDEFIX | Autonomous LTR | 594 | 1210 | 590 | 7 | 0.38 | | 61 |
| | | DM176 | Autonomous LTR | 7439 | 670 | 511 | 9 | 0.21 | | 166 |
| | | TRANSPAC | Autonomous LTR | 330 | 593 | 480 | 6 | 0.18 | | 302 |
| | | BLOOD | Autonomous LTR | 399 | 399 | 331 | 6 | 0.12 | | 232 |
| | | Gypsy1-LTR_DM | Autonomous LTR | 408 | 115 | 104 | 12 | 0.04 | | 77 |
| | | GYPSY3 | Autonomous LTR | 398 | 57 | 45 | 3 | 0.02 | | 32 |
| | | DMRT1B | Autonomous LTR | 5183 | 42 | 38 | 24 | 0.01 | | 34 |
| | | DMLTR5 | Autonomous LTR | 266 | 38 | 28 | 1 | 0.01 | | 18 |
| | | MICROPIA | Autonomous LTR | 476 | 30 | 26 | 3 | 0.009 | | 11 |
| | | GYPSY | Autonomous LTR | 482 | 28 | 25 | 3 | 0.009 | | 20 |
| | | STALKER2 | Autonomous LTR | 361 | 24 | 21 | 1 | 0.007 | | 14 |
| | | BEL | Autonomous LTR | 424 | 24 | 34 | 13 | 0.007 | | 12 |
| | | ROVER_DM | Autonomous LTR | 7318 | 14 | 11 | 4 | 0.004 | | 4 |
| | | ROO | Autonomous LTR | 276 | 8 | 26 | 19 | 0.002 | | 6 |
| | | COPIA_DM | Autonomous LTR | 428 | 8 | 9 | 2 | 0.002 | | 3 |
| | | Gypsy-7_Dsim | Autonomous LTR | 398 | 6 | 5 | 0 | 0.002 | | 5 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | DM1731 | Autonomous LTR | 336 | 4 | 10 | 6 | 0.001 | | 2 |
| Drosophila hydei flies | IIV6 | MDG1 | Autonomous LTR | 442 | 1079 | 778 | 3 | 1.33 | 2.64 | 194 |
| | | IDEFIX | Autonomous LTR | 594 | 382 | 267 | 0 | 0.47 | | 24 |
| | | DM297 | Autonomous LTR | 414 | 296 | 237 | 0 | 0.37 | | 60 |
| | | DM176 | Autonomous LTR | 7439 | 160 | 133 | 0 | 0.20 | | 54 |
| | | BLOOD | Autonomous LTR | 399 | 139 | 112 | 1 | 0.17 | | 50 |
| | | TRANSPAC | Autonomous LTR | 330 | 85 | 68 | 2 | 0.11 | | 42 |
| | | MICROPIA | Autonomous LTR | 476 | 12 | 10 | 0 | 0.015 | | 3 |
| | | STALKER2 | Autonomous LTR | 424 | 4 | 3 | 0 | 0.005 | | 2 |
| | | BEL | Autonomous LTR | 361 | 3 | 3 | 0 | 0.004 | | 2 |
| Chilo partellus larvae | IIV6 | DM297 | Autonomous LTR | 414 | 742 | 258 | 0 | 0.15 | 0.31 | 6 |
| | | MDG1 | Autonomous LTR | 442 | 578 | 210 | 5 | 0.12 | | 15 |
| | | BLOOD | Autonomous LTR | 399 | 207 | 108 | 1 | 0.04 | | 2 |
| | | DM176 | Autonomous LTR | 7439 | 19 | 16 | 0 | 0.004 | | 2 |
| | | TRANSPAC | Autonomous LTR | 330 | 7 | 6 | 0 | 0.001 | | 2 |
| | | C.partellus_piggyBac | Autonomous DNA | 2477 | 6 | 6 | 0 | 0.001 | | 2 |
| Sesamia nonagrioides larvae | IIV6 | MDG1 | Autonomous LTR | 442 | 215 | 168 | 3 | 0.06 | 0.15 | 27 |
| | | DM297 | Autonomous LTR | 414 | 206 | 152 | 0 | 0.05 | | 12 |
| | | BLOOD | Autonomous LTR | 399 | 4 | 3 | 0 | 0.001 | | 1 |
| | | DM176 | Autonomous LTR | 7439 | 70 | 57 | 0 | 0.02 | | 6 |
| | | TRANSPAC | Autonomous LTR | 330 | 43 | 40 | 0 | 0.01 | | 6 |
| | | IDEFIX | Autonomous LTR | 594 | 13 | 13 | 0 | 0.003 | | 3 |
| Sesamia nonagrioides larvae | AcMNPV | Tni_Contig_27(piggyBac) | Autonomous DNA | 2,773 | 37,897 | 2245 | 21 | 26.29 | 26.33 | 594 |
| | | Tni_contig_21(Mariner) | Autonomous DNA | 3,66 | 38 | 31 | 0 | 0.03 | | 4 |
| | | Scf7180002741278_Snona _PiggyBac (IIV6/Snona data) | Autonomous DNA | 1,763 | 5 | 5 | 0 | 0.004 | | 4 |
| | | Snona_rnd-5_family-1050 #DNA/Sola | Autonomous DNA | 1,278 | 7 | 6 | 0 | 0.005 | | 6 |
| | | Snona_rnd-5_family-1246 #DNA/Sola | Autonomous DNA | 2,093 | 6 | 5 | 0 | 0.004 | | 5 |

**Table S2.2. Characterization of Terminal Inverted Repeats (TIRs) for all inserted DNA TEs in AgseGV, AgseNPV, CpGV, IIV6 and AcMNPV datasET.** TIRs were identified for eight inserted TEs over ten, their size ranging from 12 to 62 bp.

| TE | 5' TIR coordinates | 3' TIR coordinates | TIR length (bp) | TE length (bp) |
|---|---|---|---|---|
| Agrotis_segetum_Sola | 1-45 | 4422-4465 | 45 | 4466 |
| Agrotis_segetum_PiggyBac | 1-18 | 2451-2468 | 18 | 2468 |
| Plutella_xylostella_MITE | 1-57 | 280-336 | 57 | 336 |
| Cydia_pomonella_PiggyBac | 1-12 | 4013-4024 | 12 | 4025 |
| Cydia_pomonella_rnd-5_family-1654#DNA/TcMar-Tc1 | 3-64 | 1591-1653 | 62 | 1653 |
| piggyBac_Chilo_partellus | 1-20 | 2458-2477 | 20 | 2477 |
| Tni_Contig_27 (piggybac) | 5-22 | 2759-2777 | 18 | 2773 |
| Tni_contig_21 (Mariner) | 1-21 | 3635-3655 | 21 | 3660 |
| Snona_PiggyBac_IIV6_dataset | ? | ? | ? | 1763 |
| Snona_rnd-5_family-1050#DNA/Sola | ? | ? | ? | 1278 |

**Table S2.3. Number of TE-IIV6 chimeras involving human TEs found in the IIV6/*D. melanogaster* dataset.** Almost all chimeras map internally to human TEs, in agreement with their artificial nature.

| Human TE | 5' reads | Internal reads | 3' reads |
|---|---|---|---|
| L1HS | 0 | 183 | 0 |
| L1 | 0 | 170 | 0 |
| L1PREC1 | 0 | 91 | 0 |
| AluSz | 3 | 36 | 2 |
| L1PA3 | 0 | 38 | 1 |
| L1PB2c | 0 | 36 | 0 |

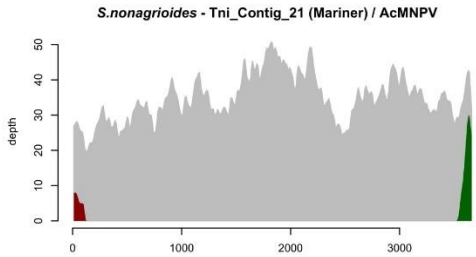
**Table S2.4. Frequency and number of chimeric reads without PCR duplicates between *D. melanogaster* transposable elements and IIV6 genomes purified from various host species.** *D. mel: Drosophila melanogaster; D. hydei: Drosophila hydei; C. partellus: Chilo partellus; S. nonagrioides: Sesamia nonagrioides*. HT indicates TEs that have been horizontally transferred between *D. melanogaster* and *D. simulans* according to Bartolomé et al. (2009).

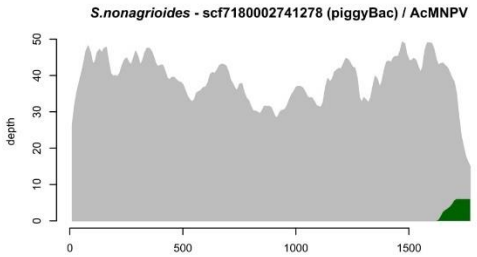
| | ***D. mel.* S2 cells** | | ***D. mel.* Whole flies (1)** | | ***D. mel.* Whole flies (2)** | | ***D. hydei*** | | ***C. partellus*** | | ***S. nonagrioides*** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ***D. mel.* TE name** | **Nb. chim.** | **Freq.** | **Nb. chim.** | **Freq.** | **Nb. chim.** | **Freq.** | **Nb. chim.** | **Freq.** | **Nb. chim.** | **Freq.** | **Nb. chim.** | **Freq.** |
| MDG1 (HT) | 2296 | 1.49 | 2080 | 1.16 | 2370 | 1.01 | 778 | 1.33 | 210 | 0.12 | 168 | 0.06 |
| DM297 (HT) | 1237 | 0.77 | 1096 | 0.61 | 1139 | 0.52 | 237 | 0.37 | 258 | 0.15 | 152 | 0.05 |
| DM176 | 701 | 0.41 | 494 | 0.24 | 520 | 0.21 | 133 | 0.20 | 16 | 0.004 | 57 | 0.02 |
| IDEFIX | 484 | 0.40 | 503 | 0.41 | 590 | 0.38 | 267 | 0.47 | 0 | 0 | 13 | 0.003 |
| TRANSPAC (HT) | 451 | 0.26 | 484 | 0.21 | 480 | 0.18 | 68 | 0.11 | 6 | 0.001 | 40 | 0.01 |
| BLOOD (HT) | 378 | 0.21 | 457 | 0.21 | 331 | 0.12 | 112 | 0.17 | 108 | 0.04 | 3 | 0.001 |
| MICROPIA (HT) | 30 | 0.02 | 40 | 0.02 | 26 | 0.009 | 10 | 0.015 | 0 | 0 | 0 | 0 |
| BEL | 21 | 0.009 | 17 | 0.005 | 34 | 0.007 | 3 | 0.004 | 0 | 0 | 0 | 0 |
| STALKER2 (HT) | 21 | 0.01 | 22 | 0.007 | 21 | 0.007 | 3 | 0.005 | 0 | 0 | 0 | 0 |
| DMRT1B | 10 | 0.006 | 50 | 0.02 | 38 | 0.01 | 0 | 0 | 0 | 0 | 0 | 0 |
| ROO | 18 | 0.005 | 20 | 0.004 | 26 | 0.002 | 0 | 0 | 0 | 0 | 0 | 0 |
| COPIA_DM (HT) | 11 | 0.001 | 9 | 0.003 | 9 | 0.002 | 0 | 0 | 0 | 0 | 0 | 0 |
| GYPSY3 | 7 | 0.005 | 41 | 0.02 | 45 | 0.02 | 0 | 0 | 0 | 0 | 0 | 0 |
| LTRMDG3_DM (HT) | 4 | 0.003 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| GYPSY6 | 0 | 0 | 6687 | 3.17 | 5651 | 2.32 | 0 | 0 | 0 | 0 | 0 | 0 |
| Gypsy1-LTR_DM | 0 | 0 | 124 | 0.05 | 104 | 0.04 | 0 | 0 | 0 | 0 | 0 | 0 |
| DMLTR5 | 0 | 0 | 27 | 0.01 | 28 | 0.01 | 0 | 0 | 0 | 0 | 0 | 0 |
| GYPSY | 0 | 0 | 17 | 0.007 | 25 | 0.009 | 0 | 0 | 0 | 0 | 0 | 0 |
| Gypsy-7_Dsim | 0 | 0 | 6 | 0.002 | 5 | 0.002 | 0 | 0 | 0 | 0 | 0 | 0 |
| STALKER4 (HT) | 0 | 0 | 8 | 0.002 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| BURDOCK (HT) | 0 | 0 | 4 | 0.002 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

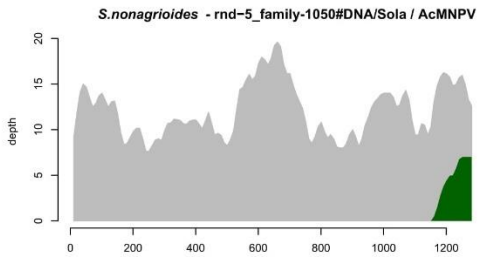
| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DM412 (HT) | 0 | 0 | 5 | 0.002 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| BLASTOPIA | 0 | 0 | 10 | 0.002 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ROVER_DM | 0 | 0 | 0 | 0 | 11 | 0.004 | 0 | 0 | 0 | 0 | 0 | 0 |
| DM1731 (HT) | 0 | 0 | 0 | 0 | 10 | 0.001 | 0 | 0 | 0 | 0 | 0 | 0 |

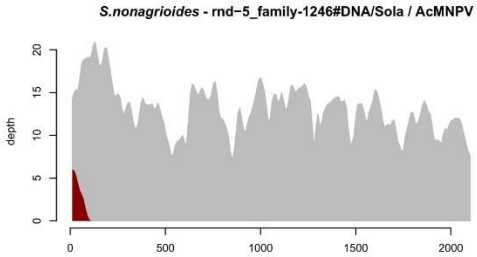
**Figure S2.1**. **Photographs of infected versus non-infected *Drosophila melanogaster* flies.** The abdomen of infected flies (photographs on the right) is iridescent blue due to the arrangement of viral particles in the cuticle of infected flies.

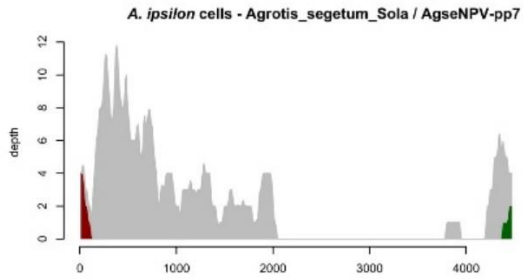
All the following coverage graphs are part of the figure S2.2.

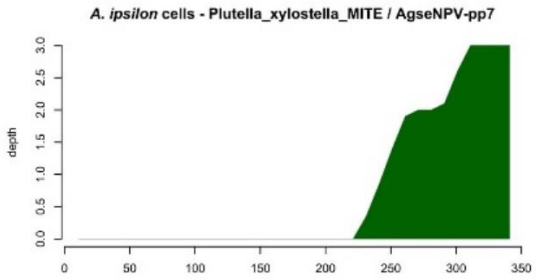

S.nonagrioides - Tni_Contig_21 (Mariner) / AcMNPV
depth
0
10
20
30
40
50
0
1000
2000
3000

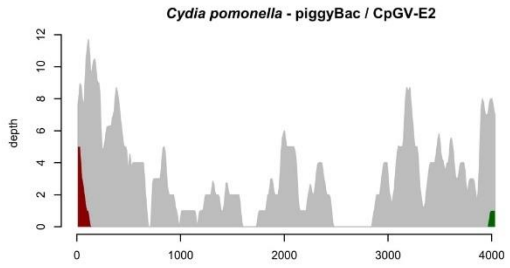

S.nonagrioides - scf7180002741278 (piggyBac) / AcMNPV
depth
0
10
20
30
40
50
0
500
1000
1500

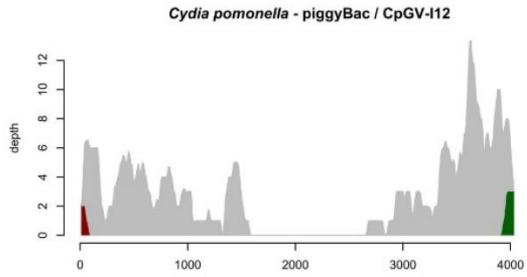

S.nonagrioides - rnd−5_family-1050#DNA/Sola / AcMNPV
depth
0
5
10
15
20
0
200
400
600
800
1000
1200

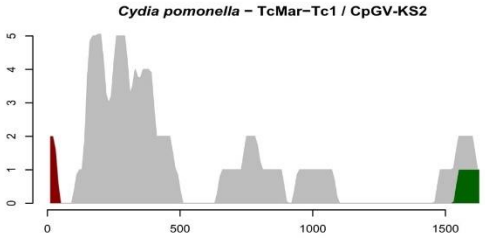

S.nonagrioides - rnd−5_family-1246#DNA/Sola / AcMNPV
depth
0
5
10
15
20
0
500
1000
1500
2000

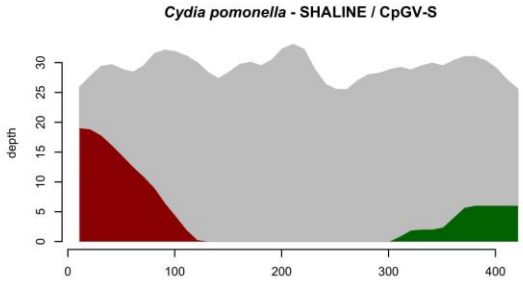

A. ipsilon cells - Agrotis_segetum_Sola / AgseNPV-pp7
depth
0
2
4
6
8
10
12
0
1000
2000
3000
4000

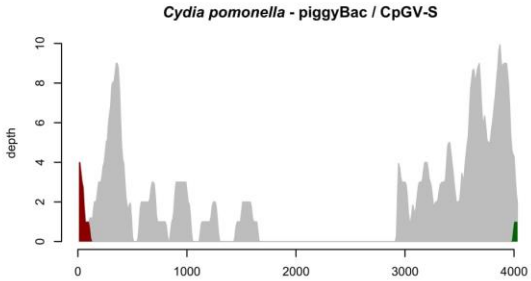

A. ipsilon cells - Plutella_xylostella_MITE / AgseNPV-pp7
depth
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
50
100
150
200
250
300
350

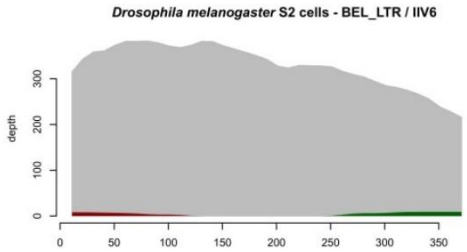

*Cydia pomonella* - piggyBac / CpGV-E2
depth
0
2
4
6
8
10
12
0
1000
2000
3000
4000

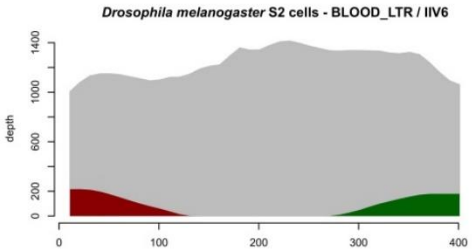

*Cydia pomonella* - piggyBac / CpGV-I12
depth
0
2
4
6
8
10
12
0
1000
2000
3000
4000

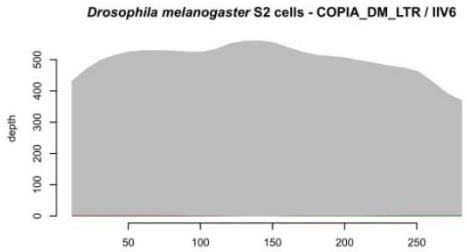

*Cydia pomonella* – TcMar–Tc1 / CpGV-KS2
0
1
2
3
4
5
0
500
1000
1500

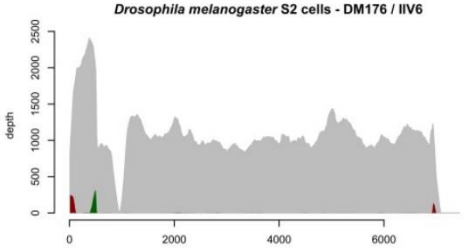

*Cydia pomonella* - SHALINE / CpGV-S
depth
0
5
10
15
20
25
30
0
100
200
300
400

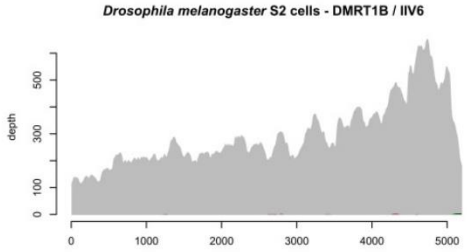

*Cydia pomonella* - piggyBac / CpGV-S
depth
0
2
4
6
8
10
0
1000
2000
3000
4000

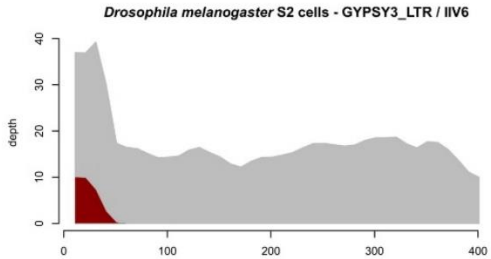

Drosophila melanogaster S2 cells - BEL_LTR / IIV6
depth
0
100
200
300
0
50
100
150
200
250
300
350

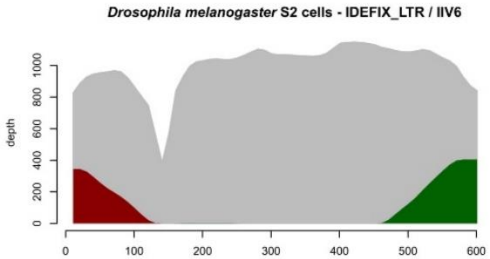

Drosophila melanogaster S2 cells - BLOOD_LTR / IIV6
depth
0
200
600
1000
1400
0
100
200
300
400

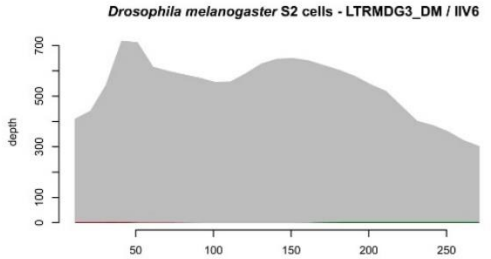

Drosophila melanogaster S2 cells - COPIA_DM_LTR / IIV6
depth
0
100
200
300
400
500
50
100
150
200
250

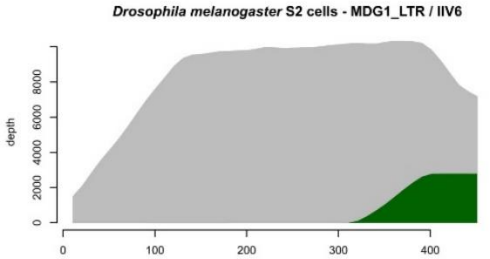

Drosophila melanogaster S2 cells - DM176 / IIV6
depth
0
500
1000
1500
2000
2500
0
2000
4000
6000

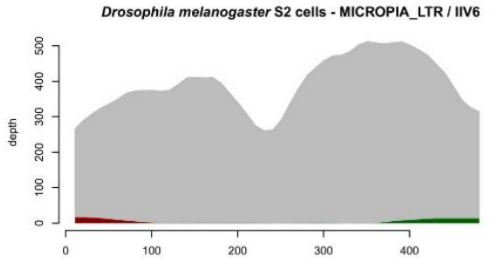

Drosophila melanogaster S2 cells - DMRT1B / IIV6
depth
0
100
300
500
0
1000
2000
3000
4000
5000

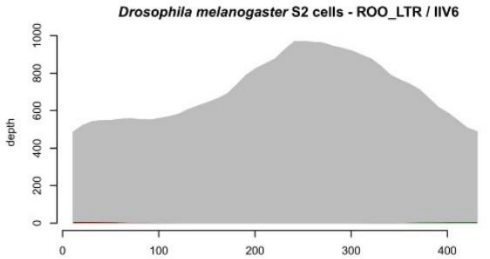

Drosophila melanogaster S2 cells - GYPSY3_LTR / IIV6
depth
0
10
20
30
40
0
100
200
300
400

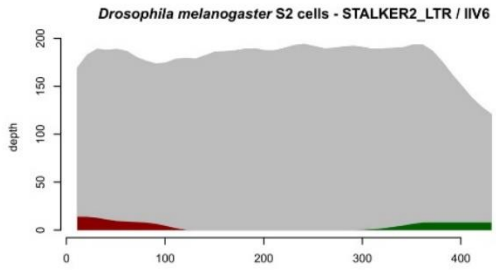

Drosophila melanogaster S2 cells - IDEFIX_LTR / IIV6
depth
0
200
400
600
800
1000
0
100
200
300
400
500
600

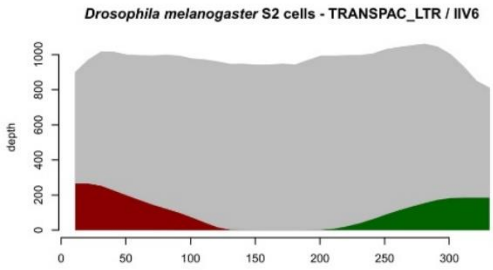

Drosophila melanogaster S2 cells - LTRMDG3_DM / IIV6
depth
0
100
300
500
700
50
100
150
200
250

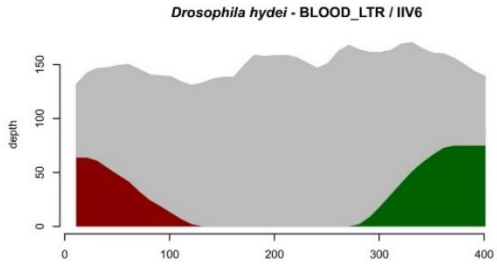

Drosophila melanogaster S2 cells - MDG1_LTR / IIV6
depth
0
2000
4000
6000
8000
0
100
200
300
400

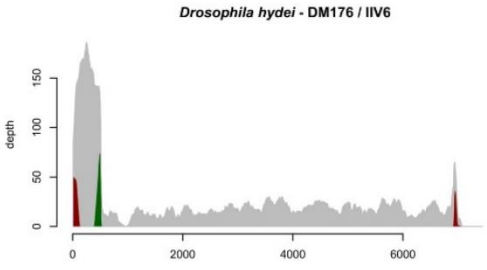

Drosophila melanogaster S2 cells - MICROPIA_LTR / IIV6
depth
0
100
200
300
400
500
0
100
200
300
400

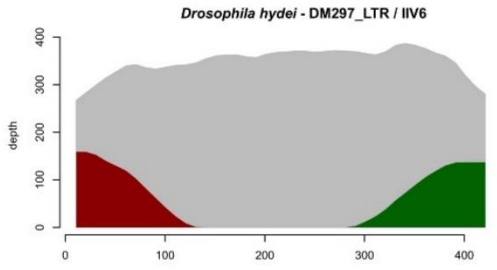

Drosophila melanogaster S2 cells - ROO_LTR / IIV6
depth
0
200
400
600
800
1000
0
100
200
300
400

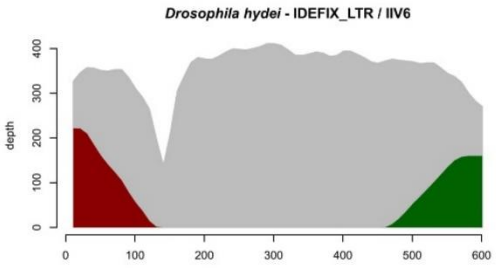

Drosophila melanogaster S2 cells - STALKER2_LTR / IIV6
depth
0
50
100
150
200
0
100
200
300
400

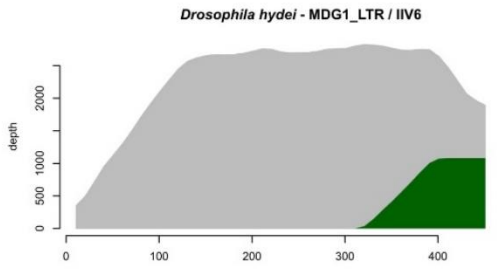

Drosophila melanogaster S2 cells - TRANSPAC_LTR / IIV6
depth
0
200
400
600
800
1000
0
50
100
150
200
250
300

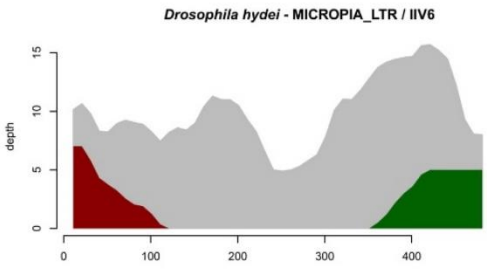

Drosophila hydei - BLOOD_LTR / IIV6
depth
0
50
100
150
0
100
200
300
400

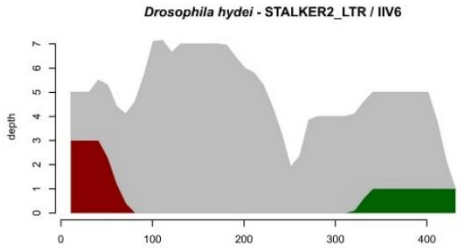

Drosophila hydei - DM176 / IIV6
depth
0
50
100
150
0
2000
4000
6000

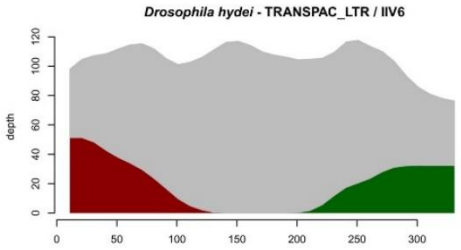

Drosophila hydei - DM297_LTR / IIV6
depth
0
100
200
300
400
0
100
200
300
400

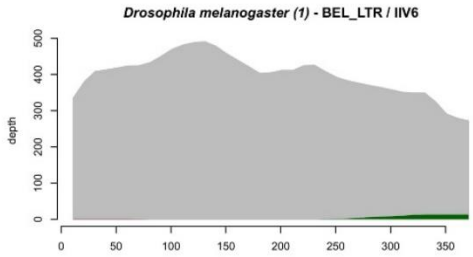

Drosophila hydei - IDEFIX_LTR / IIV6
depth
0
100
200
300
400
0
100
200
300
400
500
600

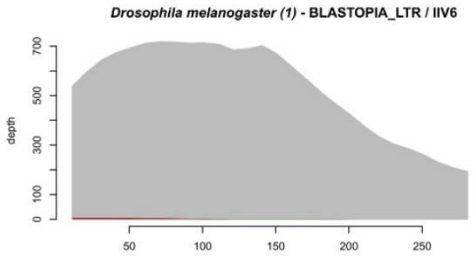

Drosophila hydei - MDG1_LTR / IIV6
depth
0
500
1000
2000
0
100
200
300
400

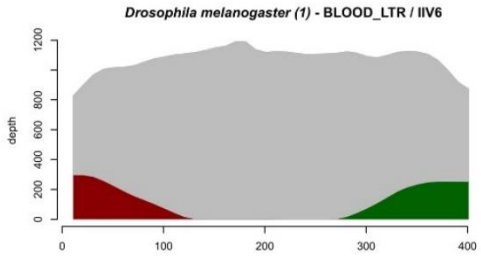

Drosophila hydei - MICROPIA_LTR / IIV6
depth
0
5
10
15
0
100
200
300
400

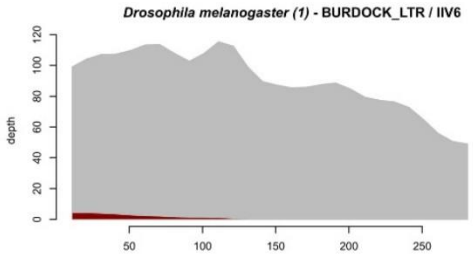

Drosophila hydei - STALKER2_LTR / IIV6
depth

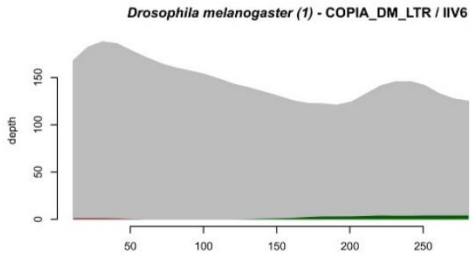

Drosophila hydei - TRANSPAC_LTR / IIV6
depth

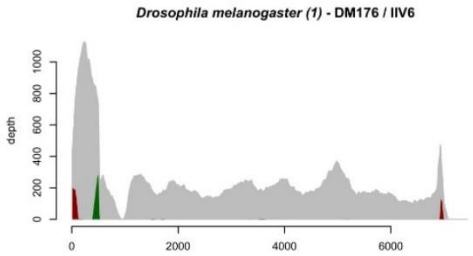

Drosophila melanogaster (1) - BEL_LTR / IIV6
depth

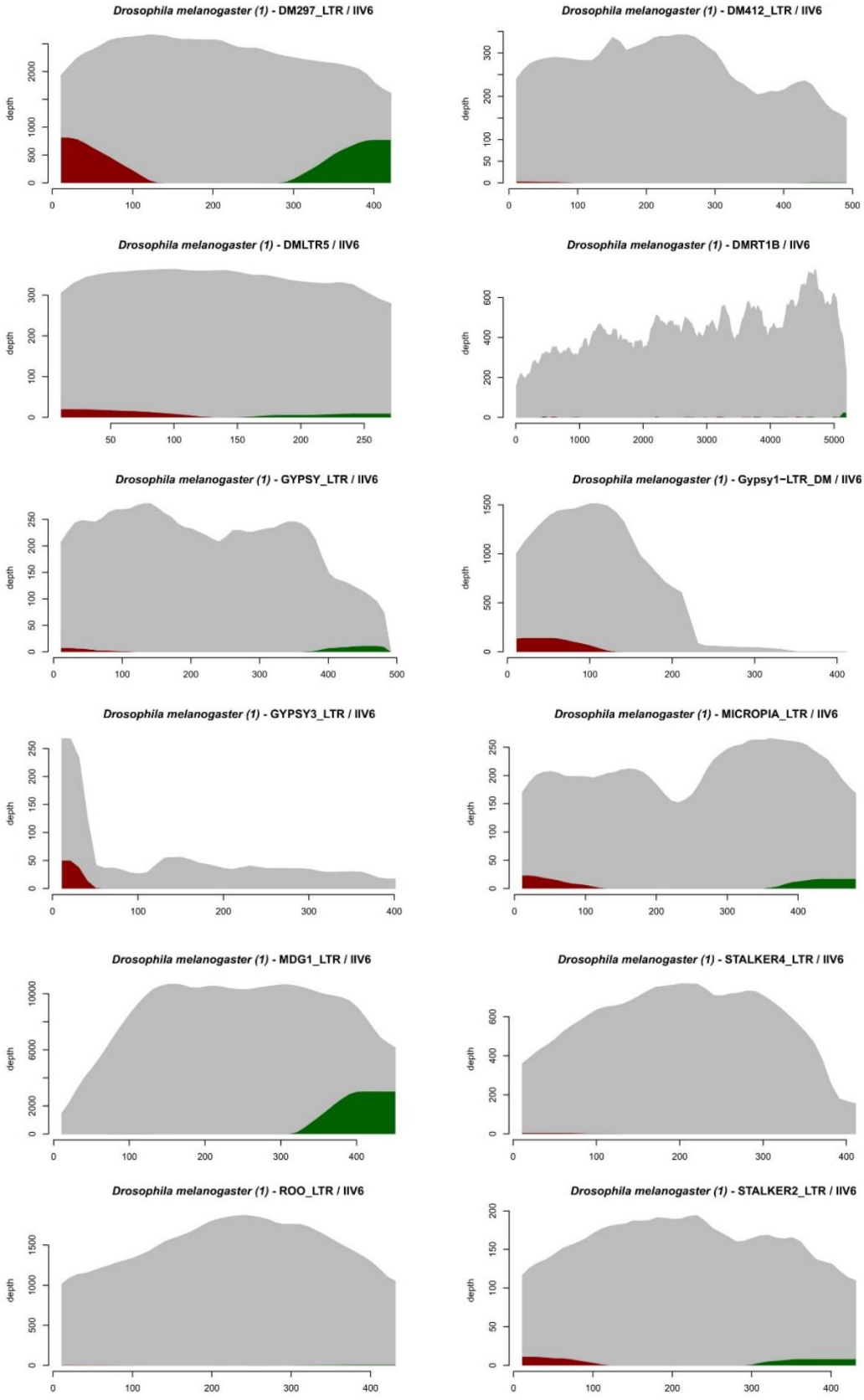

Drosophila melanogaster (1) - BLASTOPIA_LTR / IIV6
depth

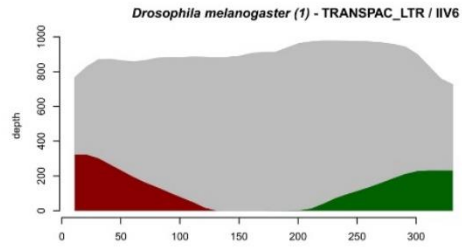

Drosophila melanogaster (1) - BLOOD_LTR / IIV6
depth

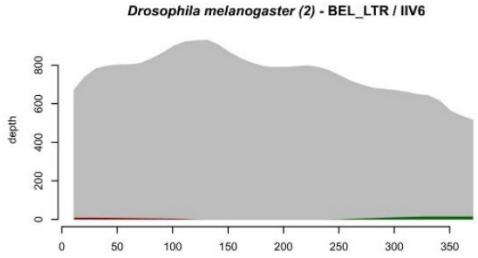

Drosophila melanogaster (1) - BURDOCK_LTR / IIV6
depth

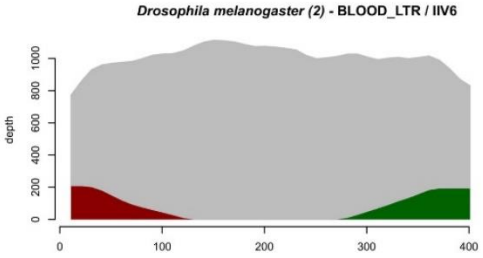

Drosophila melanogaster (1) - COPIA_DM_LTR / IIV6
depth

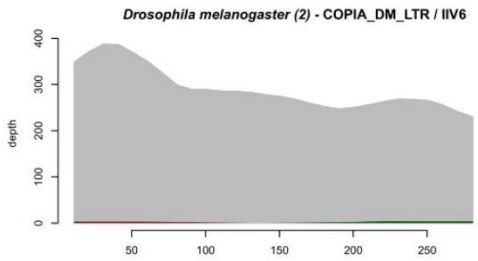

Drosophila melanogaster (1) - DM176 / IIV6
depth

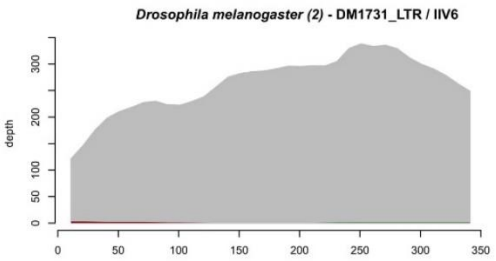

Drosophila melanogaster (1) - DM297_LTR / IIV6
Drosophila melanogaster (1) - DM412_LTR / IIV6
Drosophila melanogaster (1) - DMLTR5 / IIV6
Drosophila melanogaster (1) - DMRT1B / IIV6
Drosophila melanogaster (1) - GYPSY_LTR / IIV6
Drosophila melanogaster (1) - Gypsy1−LTR_DM / IIV6
Drosophila melanogaster (1) - GYPSY3_LTR / IIV6
Drosophila melanogaster (1) - MICROPIA_LTR / IIV6
Drosophila melanogaster (1) - MDG1_LTR / IIV6
Drosophila melanogaster (1) - STALKER4_LTR / IIV6
Drosophila melanogaster (1) - ROO_LTR / IIV6
Drosophila melanogaster (1) - STALKER2_LTR / IIV6
depth

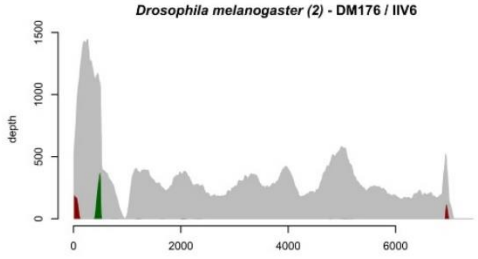

Drosophila melanogaster (1) - TRANSPAC_LTR / IIV6
depth
0
200
400
600
800
1000
0
50
100
150
200
250
300

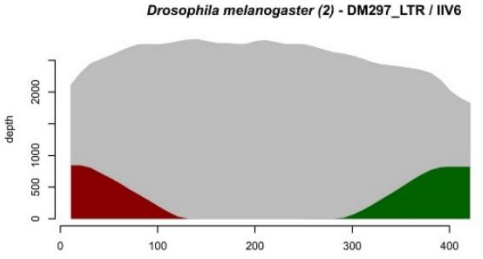

Drosophila melanogaster (2) - BEL_LTR / IIV6
depth
0
200
400
600
800
0
50
100
150
200
250
300
350

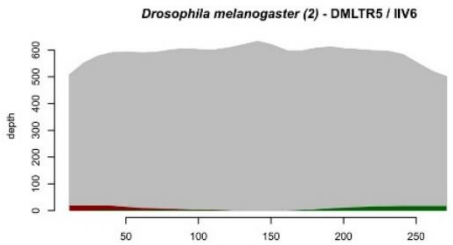

Drosophila melanogaster (2) - BLOOD_LTR / IIV6
depth
0
200
400
600
800
1000
0
100
200
300
400

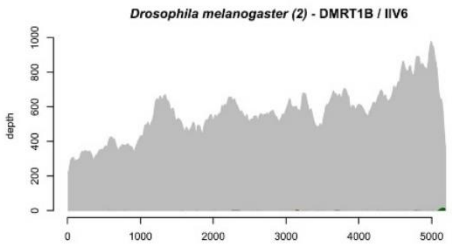

Drosophila melanogaster (2) - COPIA_DM_LTR / IIV6
depth
0
100
200
300
400
50
100
150
200
250

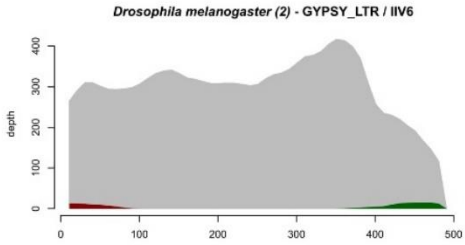

Drosophila melanogaster (2) - DM1731_LTR / IIV6
depth
0
50
100
200
300
0
50
100
150
200
250
300
350

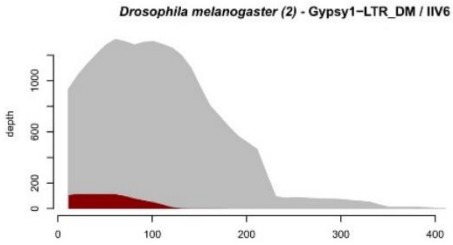

Drosophila melanogaster (2) - DM176 / IIV6
depth
0
500
1000
1500
0
2000
4000
6000

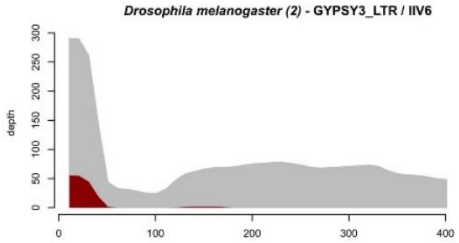

Drosophila melanogaster (2) - DM297_LTR / IIV6
depth
0
500
1000
2000
0
100
200
300
400

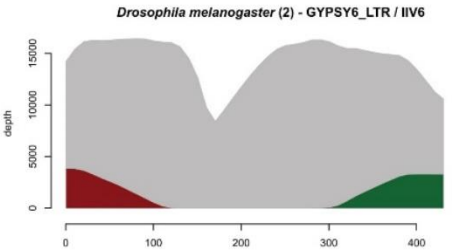

Drosophila melanogaster (2) - DMLTR5 / IIV6
depth
0 100 200 300 400 500 600
50 100 150 200 250

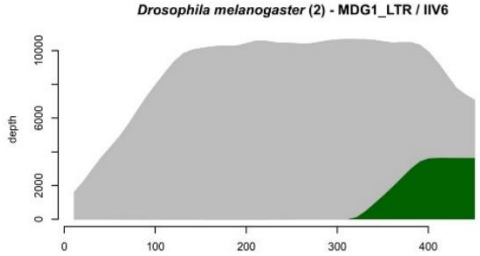

Drosophila melanogaster (2) - DMRT1B / IIV6
depth
0 200 400 600 800 1000
0 1000 2000 3000 4000 5000

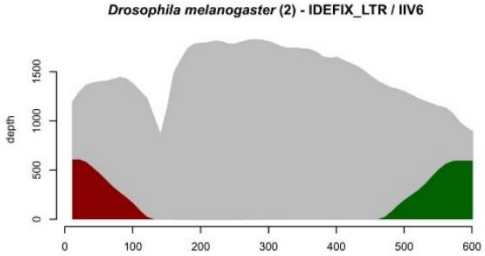

Drosophila melanogaster (2) - GYPSY_LTR / IIV6
depth
0 100 200 300 400
0 100 200 300 400 500

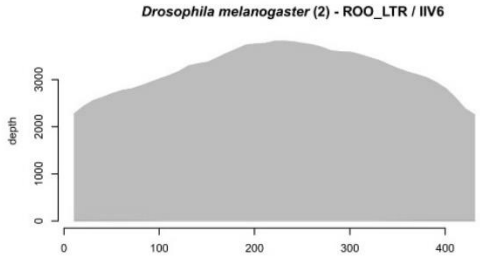

Drosophila melanogaster (2) - Gypsy1−LTR_DM / IIV6
depth
0 200 600 1000
0 100 200 300 400

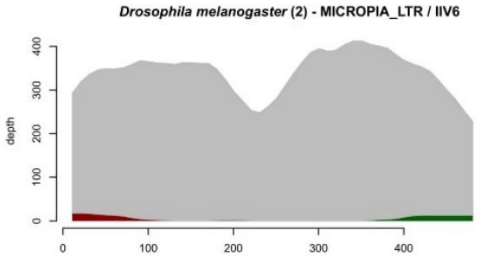

Drosophila melanogaster (2) - GYPSY3_LTR / IIV6
depth
0 50 100 150 200 250 300
0 100 200 300 400

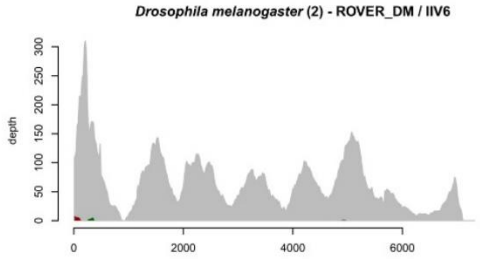

Drosophila melanogaster (2) - GYPSY6_LTR / IIV6
depth
0 5000 10000 15000
0 100 200 300 400

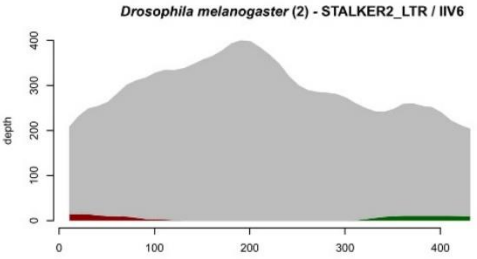

Drosophila melanogaster (2) - MDG1_LTR / IIV6
depth
0 2000 6000 10000
0 100 200 300 400


Drosophila melanogaster (2) - IDEFIX_LTR / IIV6
depth
0 500 1000 1500
0 100 200 300 400 500 600


Drosophila melanogaster (2) - ROO_LTR / IIV6
depth
0 1000 2000 3000
0 100 200 300 400


Drosophila melanogaster (2) - MICROPIA_LTR / IIV6
depth
0 100 200 300 400
0 100 200 300 400


Drosophila melanogaster (2) - ROVER_DM / IIV6
depth
0 50 100 150 200 250 300
0 2000 4000 6000


Drosophila melanogaster (2) - STALKER2_LTR / IIV6
depth
0 100 200 300 400
0 100 200 300 400

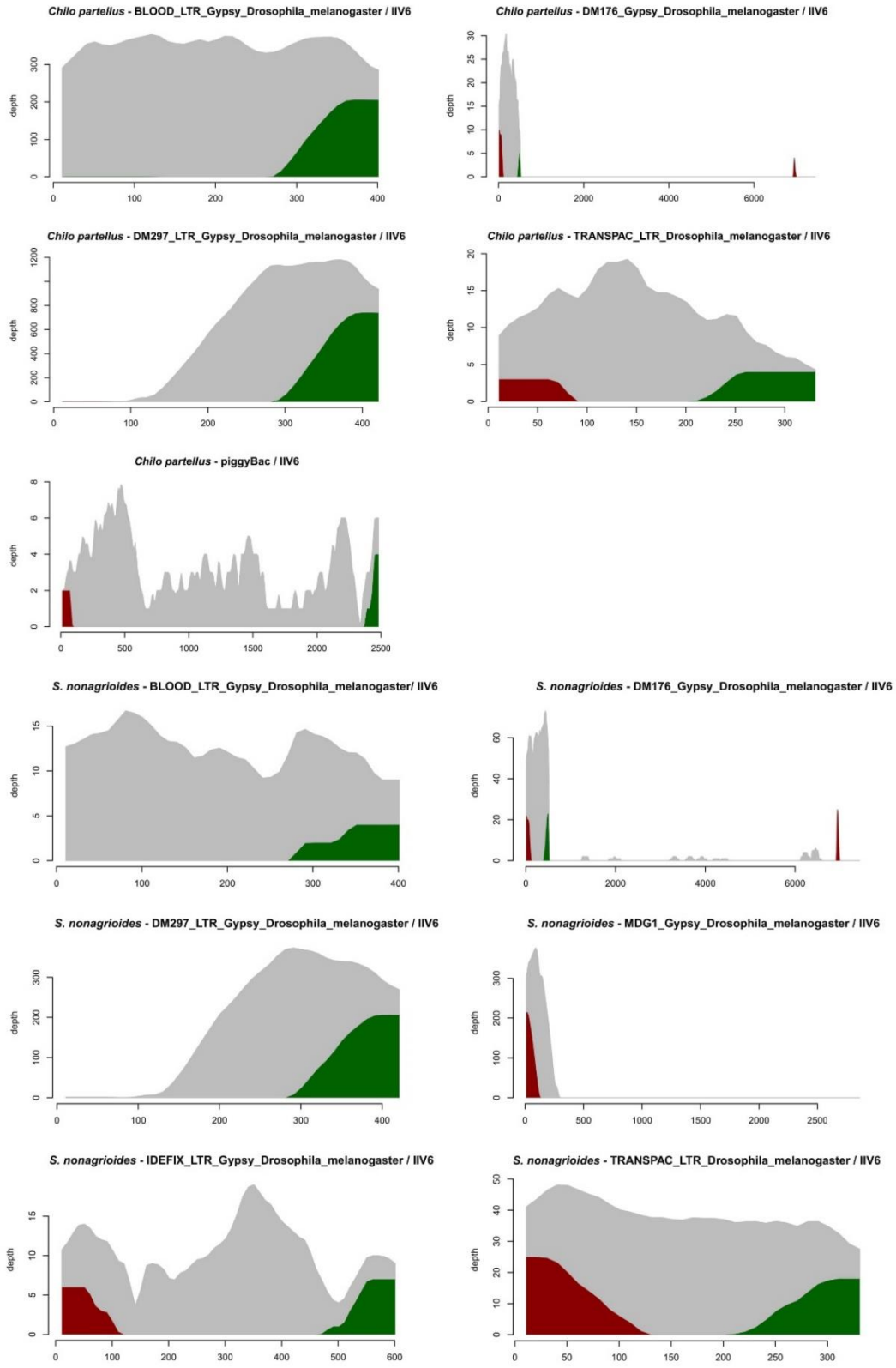



**Figure S2.2. Graphs illustrating read depth along all transposable elements (TEs) found integrated into viral genomes.** Depth of non-chimeric reads (reads mapping entirely on the TE) is shown in gray. Depth of chimeric reads (reads for which a portion maps on the viral genome only while the other portion maps on the TE only) is shown in red for reads mapping on the 5' extremity of TEs and in green for reads mapping on the 3' extremity of the TE. The figure shows that the vast majority of chimeric reads map at the extremities of TEs.

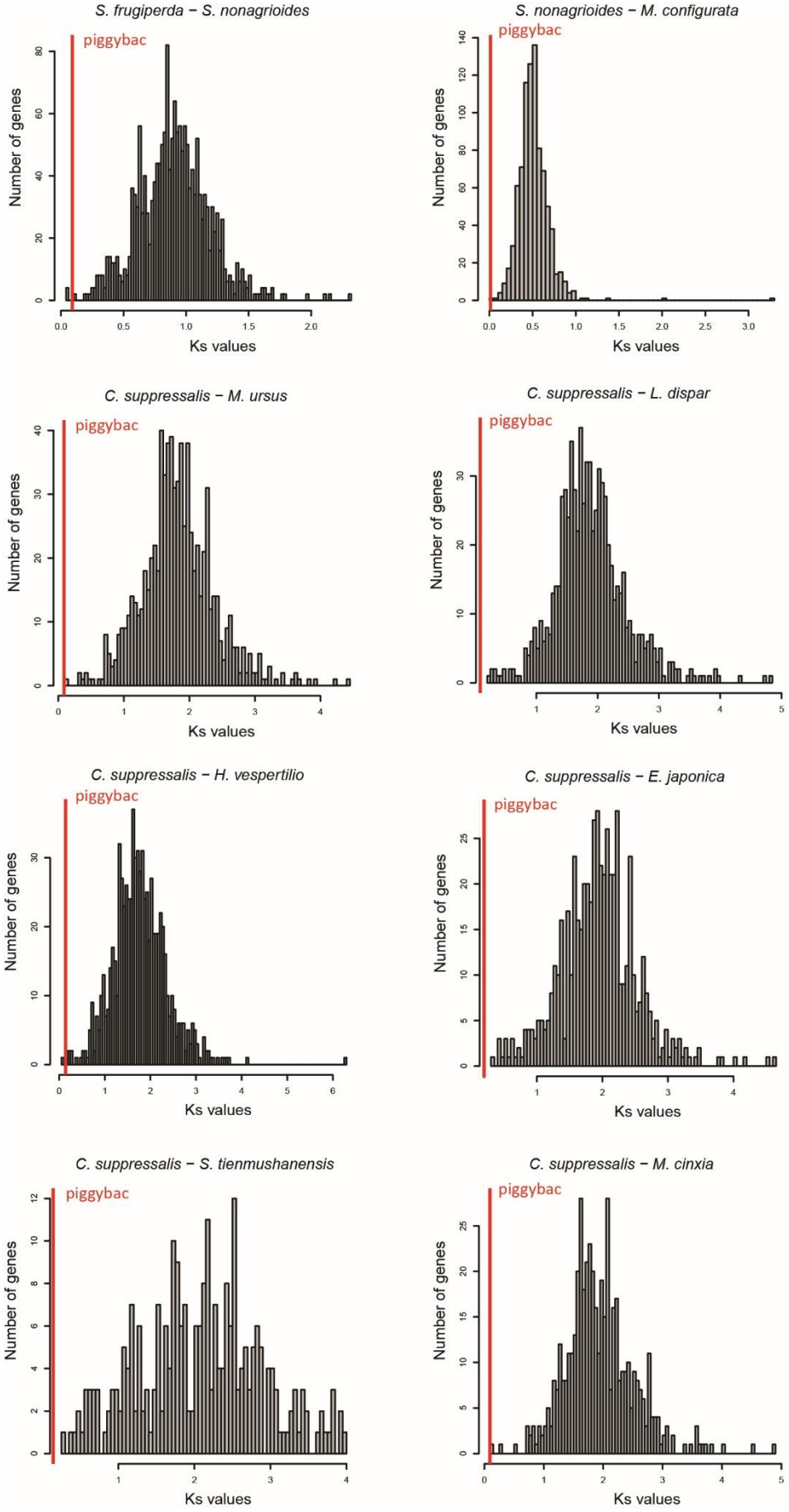

S. frugiperda − S. nonagrioides
piggybac
Number of genes
Ks values
S. nonagrioides − M. configurata
piggybac
Number of genes
Ks values
C. suppressalis − M. ursus
piggybac
Number of genes
Ks values
C. suppressalis − L. dispar
piggybac
Number of genes
Ks values
C. suppressalis − H. vespertilio
piggybac
Number of genes
Ks values
C. suppressalis − E. japonica
piggybac
Number of genes
Ks values
C. suppressalis − S. tienmushanensis
piggybac
Number of genes
Ks values
C. suppressalis − M. cinxia
piggybac
Number of genes
Ks values

**Figure S2.3. Graphs illustrating synonymous distances (Ks values) distribution for eight species pair comparisons.** The distances were computed on the complete unique copy of Insecta genes shared by the two species of a pair. Vertical red line represents the Ks value of the piggyBac gene.

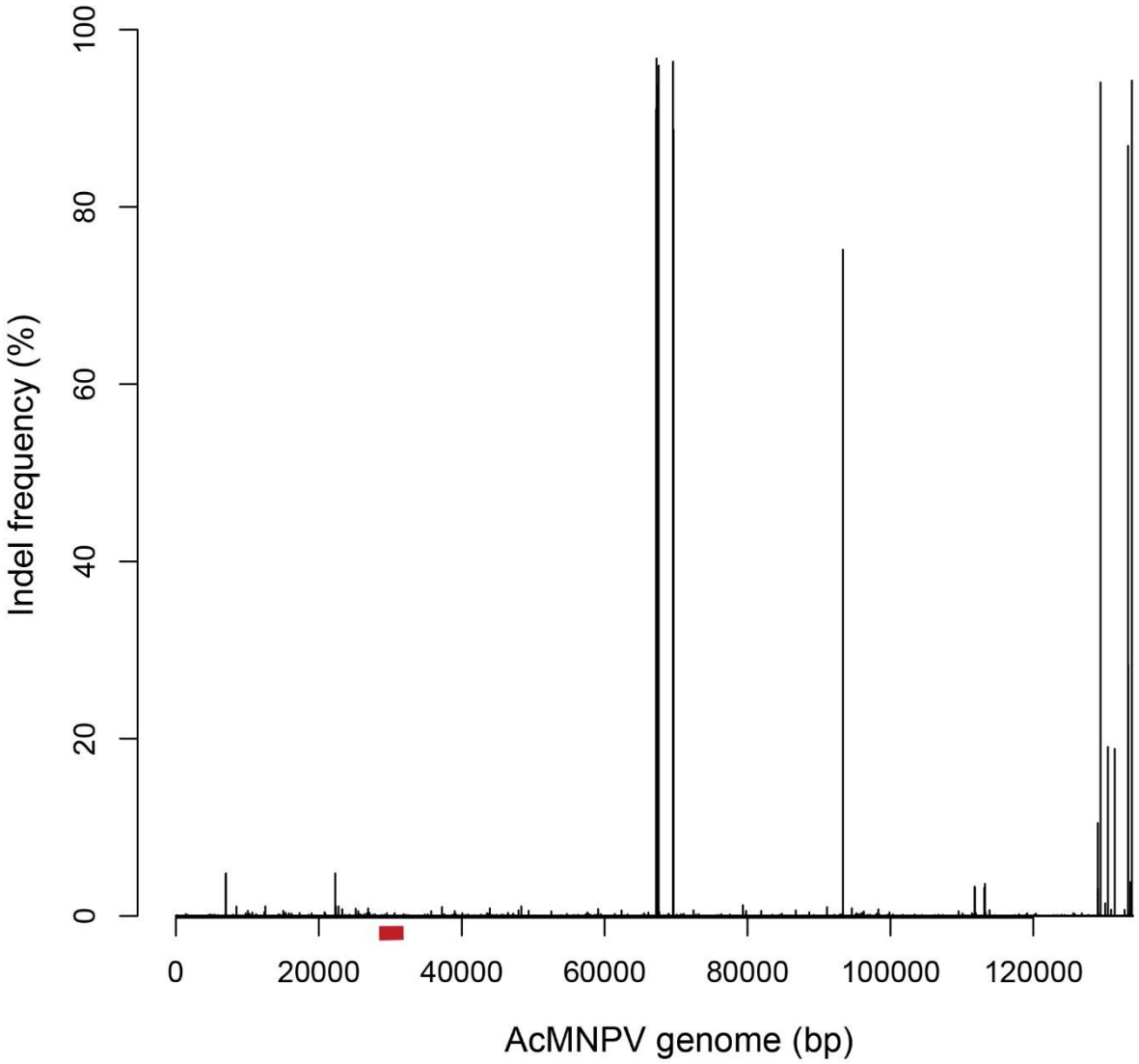



**Figure S2.4. Indel frequency along the AcMNPV genome (including position of the *Ac-GTA* gene).**

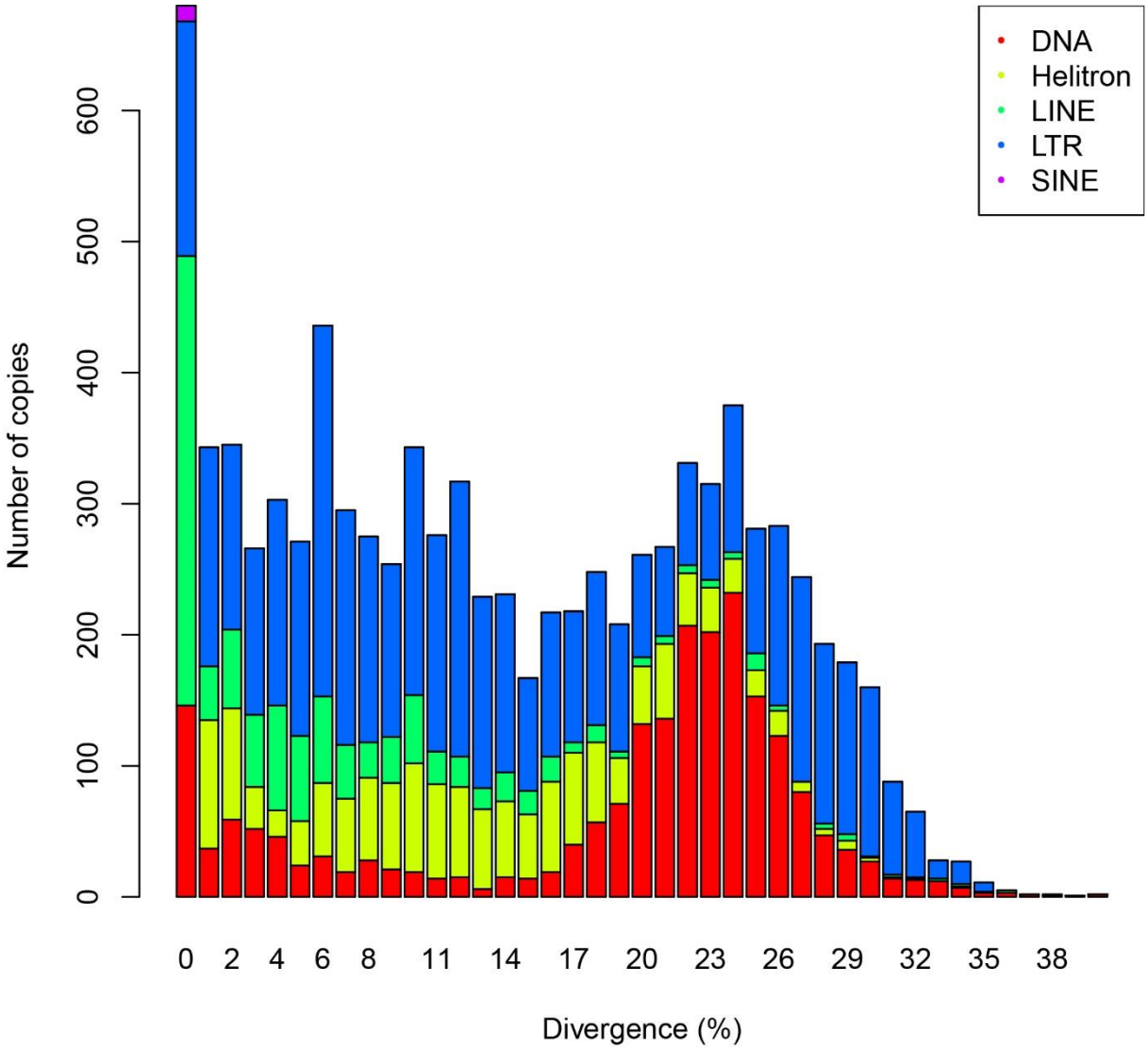



**Figure S2.5. TE landscape of *D. hydei*.** All TE copies >300 bp were considered. One bar corresponds to 1% of divergence. Mainly DNA and LTR TEs are represented, but LINEs have the greatest number of 0% divergent copies to the consensus sequence. The distribution mode reaches a peak at 0%, meaning some copies are functional. A second peak is reached at 23% of divergence corresponding to an ancient TE amplification, these copies being likely non-functional.

# Chapitre 3

# *Chapitre 3 : Analyse de longues lectures de séquençage pour l'identification d'insertions complètes d'ET et d'autres variants structuraux dans les génomes viraux*

Le premier chapitre a permis de révéler l'impact d'une infection virale sur l'activité des ET de l'hôte, certains étant surexprimés. De plus, certains ET insérés dans les génomes viraux sont transcrits, ce qui constitue une étape importante afin de pouvoir transposer à nouveau, potentiellement lors de l'infection d'un nouvel hôte. Le deuxième chapitre a étendu le spectre des systèmes hôte-virus connus pour lesquels des ET de l'hôte peuvent transposer dans les génomes viraux. Il a également apporté des précisions concernant la fréquence d'insertion des ET dans les génomes viraux et leur transposition au sein d'une population virale. Néanmoins, toutes ces analyses ont été effectuées à partir de courtes lectures de séquençage qui permettent de détecter les points d'intégration des ET mais ne permettent pas de caractériser l'intégralité de l'insertion. L'utilisation de longues lectures, malgré leur taux d'erreur nucléotidique élevé (environ 13%) et la présence d'insertions et de délétions artéfactuelles entravant la qualité de l'analyse, a permis de confirmer que les insertions d'ET pouvaient être complètes dans des génomes d'AcMNPV. Aussi il est possible que ces séquences complètes soient actives et puissent transposer, en accord avec la découverte de transposition virus-vers-virus dans le chapitre précédant. Au total, 524 séquences complètes d'ET provenant de la noctuelle exiguë *Spodoptera exigua* ont été détectées dans les génomes viraux. En accord avec ce qui avait déjà été montré (Gilbert et al., 2014, 2016), une majorité d'ET appartenant aux superfamilles piggybac, mariner et harbinger ont des séquences complètes insérées dans les génomes d'AcMNPV.

De plus, le séquençage à forte profondeur (>195000 X) de courtes et longues lectures, réalisé sur l'ADN viral offrait l'opportunité d'utiliser ces jeux de données différents afin de caractériser les variants structuraux (VS) génomiques autres que les ET présents dans les populations virales. Cela nous a conduits à la détection de 1141 VS, dont 464 délétions, 443 inversions, 160 duplications et 74 insertions. L'ensemble de ces VS affectait 39,9% des génomes d'AcMNPV,

ce qui a permis de mettre en lumière la présence importante des VS au sein des populations virales. Afin d'avoir une meilleure idée des VS affectant différentes populations virales, cette même analyse a été poursuivie sur trois autres jeux de données à forte profondeur de séquençage en lectures courtes correspondant à trois virus à ADN doubles-brin différents, le cytomégalovirus humain (HCMV) et les iridovirus IIV6 et IIV31.

Ces travaux ont fait l'objet d'une publication dont je suis premier auteur parue en janvier 2020 dans le journal *Virus Evolution*. L'article est présenté ci-dessous tel que disponible dans ce journal.

Virus Evolution, 2020, 6(1): vez060

doi: 10.1093/ve/vez060
Research article

# Wide spectrum and high frequency of genomic structural variation, including transposable elements, in large double-stranded DNA viruses

Vincent Loiseau,[1,*] Elisabeth A. Herniou,[2] Yannis Moreau,[2] Nicolas Lévêque,[3,4] Carine Meignin,[5] Laurent Daeffler,[5] Brian Federici,[6] Richard Cordaux,[7] and Clément Gilbert[1,*,†]

[1]Laboratoire Evolution, Génomes, Comportement, Ecologie, Unité Mixte de Recherche 9191 Centre National de la Recherche Scientifique et Unité Mixte de Recherche 247 Institut de Recherche pour le Développement, Université Paris-Saclay, Gif-sur-Yvette 91198, France, [2]Institut de Recherche sur la Biologie de l'Insecte, UMR 7261 CNRS - Université de Tours, 37200 Tours, France, [3]Laboratoire de Virologie et Mycobactériologie, CHU de Poitiers, 86000 Poitiers, France, [4]Laboratoire Inflammation, Tissus Epithéliaux et Cytokines, EA 4331, Université de Poitiers, 86000 Poitiers, France, [5]Modèles Insectes d'Immunité Innée (M3i), Université de Strasbourg, IBMC CNRS-UPR9022, Strasbourg F-67000, France, [6]Department of Entomology and Institute for Integrative Genome Biology, University of California, Riverside, CA 92521, USA and [7]Laboratoire Ecologie et Biologie des Interactions, Equipe Ecologie Evolution Symbiose, Unité Mixte de Recherche 7267 Centre National de la Recherche Scientifique, Université de Poitiers, 86000 Poitiers, France

*Corresponding author: E-mail: vincent.loiseau01@gmail.com; clement.gilbert@egce.cnrs-gif.fr

[†]https://orcid.org/0000-0002-2131-7467

## Abstract

Our knowledge of the diversity and frequency of genomic structural variation segregating in populations of large double-stranded (ds) DNA viruses is limited. Here, we sequenced the genome of a baculovirus (Autographa californica multiple nucle-opolyhedrovirus [AcMNPV]) purified from beet armyworm (Spodoptera exigua) larvae at depths >195,000 using both short-(Illumina) and long-read (PacBio) technologies. Using a pipeline relying on hierarchical clustering of structural variants (SVs) detected in individual short- and long-reads by six variant callers, we identified a total of 1,141 SVs in AcMNPV, including 464 deletions, 443 inversions, 160 duplications, and 74 insertions. These variants are considered robust and unlikely to result from technical artifacts because they were independently detected in at least three long reads as well as at least three short reads. SVs are distributed along the entire AcMNPV genome and may involve large genomic regions (30,496 bp on aver-age). We show that no less than 39.9 per cent of genomes carry at least one SV in AcMNPV populations, that the vast majority of SVs (75%) segregate at very low frequency (<0.01%) and that very few SVs persist after ten replication cycles, consistent with a negative impact of most SVs on AcMNPV fitness. Using short-read sequencing datasets, we then show that populations of two iridoviruses and one herpesvirus are also full of SVs, as they contain between 426 and 1,102 SVs carried by 52.4–80.1 per cent of genomes. Finally, AcMNPV long reads allowed us to identify 1,757 transposable elements (TEs) inser-tions, 895 of which are truncated and occur at one extremity of the reads. This further supports the role of baculoviruses as possible vectors of horizontal transfer of TEs. Altogether, we found that SVs, which evolve mostly under rapid dynamics of gain and loss in viral populations, represent an important feature in the biology of large dsDNA viruses.

Vc The Author(s) 2020. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Key words: large double-stranded DNA viruses; genomic structural variation; transposable elements; herpesvirus; iridovirus; baculovirus.

## 1. Introduction

Estimating the evolutionary potential of viral populations is key to our understanding of how and how fast viruses may evolve in response to new environmental constraints. Such potential is directly linked to the genetic diversity of viral populations, which has been well characterized in only a handful of viruses. RNA viruses display high mutation rates, large population sizes, and fast replication dynamics, which all together generate clouds of genetically linked single nucle-otide variants that functionally cooperate and collectively contribute to the fitness of the viral population (Lauring and Andino 2010; Acevedo, Brodsky, and Andino 2014). Such extremely high levels of polymorphism allow RNA vi-ruses to rapidly adapt to the various host and cellular environments they may be exposed to (Lauring, Frydman, and Andino 2013; Sanjuan and Domingo-Calap 2016). It also makes the outcome of infection difficult to predict, and thus poses major challenges to the prevention and treatment of viral diseases.

In contrast to most RNA viruses, which do not encode error-correcting polymerases, large double-stranded DNA (dsDNA) viruses use high-fidelity, proofreading polymerases (Duffy, Shackelton, and Holmes 2008; Sanjuan and Domingo-Calap 2016). As a consequence, the mutation rate of large dsDNA viruses is two to four orders of magnitude lower than that of RNA viruses. Despite these lower mutation rates, populations of large dsDNA viruses exhibit very high nucleotide diversity. For example, high-throughput sequencing approaches have revealed several thousands of single nucleotide polymorphisms (SNPs) segregating in populations of the human cytomegalovirus (HCMV) (Renzette et al. 2013, 2015, 2017), the Autographa californica multiple nucleopolyhedrovirus (AcMNPV) (Chateigner et al. 2015), and human herpesvirus 2 (Akhtar et al. 2019). Although the majority of these SNPs are at low frequency and likely neutral, a fraction was shown to be under positive selection and involved in rapid adaptation during intra-host evolution (Renzette et al. 2013).

In addition to SNPs, another source of genetic diversity found in viral populations is structural variation, which may be defined as deletions, insertions, inversions, duplications and translocations (Alkan, Coe, and Eichler 2011). Some forms of structural variants (SVs), leading to defective viral genomes (DVGs), have been the subject of extensive experimentation because of the negative impact they have on viral replication (Marriott and Dimmock 2010). DVGs were first discovered in populations of the Influenza A virus and have since been extensively studied in RNA viruses (Manzoni and Lopez 2018). Their presence in RNA viral populations is pivotal to the intra-host dynamics of viral infections, to the point that abnormal depletion in DVG can lead to severe disease outcomes (Vasilijevic et al. 2017). Such DVGs, provided they contain all the signals necessary for packaging, can outcompete complete viral genomes and rapidly cause important drops in overall virus titers (Li et al. 2011). RNA virus DVGs are also known to play a role in the induction of the interferon-mediated antiviral response (Lopez 2014) as well as in the Dicer-dependent viral DNA-mediated antiviral RNAi response in insects (Poirier et al. 2018).

Historically, DVGs have been less studied in large dsDNA viruses. With the development of protein expression vectors, most experiments focused on baculoviruses (De Gooijer et al. 1992). Experimental assays coupled to population genetics modeling characterized interactions between complete and DVGs (Bull, Godfray, and O'Reilly 2003; Zwart, Tromas, and Elena 2013) to assess what proportion of DVGs may be optimal to limit the persistence of complete viruses used as biopesticides (Kool et al. 1991; Godfray, Reilly, and Briggs 1997). These approaches mostly revealed a negative impact of DVGs on virus replication and production. Yet, in natural viral populations beneficial interactions may exist between defective and complete viral genomes, as mixtures are more pathogenic than clonal wild type populations (Simon et al. 2006). Besides baculoviruses, a high proportion of non-canonical viral genomes have also been detected in populations of human herpesviruses us-ing molecular combing or Sanger sequencing (Mahiet et al. 2012).

Despite the impact of SVs on viral population dynamics and infection outcome, our knowledge on their full spectrum and frequency remains limited. Next generation sequencing (NGS) offers potent tools to probe the extent of SV diversity in large viral populations (Acevedo and Andino 2014). However, most studies of viral SVs using NGS have so far focused on major variants through assembling and comparing consensus genomes from different viral strains (Szpara et al. 2014; Karamitros et al. 2016, 2018). Surveys of intra-host viral SVs, as detected in individual sequencing reads, remain scarce, and often limited to specific, targeted rearrangements (Elde et al. 2012; Sasani et al. 2018). One of the most comprehensive NGS-based analysis of SVs diversity has been conducted on the flock house virus (FHV; Alphanodaviridae) after replication in Drosophila melanogaster S2 cells (Routh et al. 2015; Jaworski and Routh 2017). These studies led to the characterization of hundreds of different recombination events along the FHV RNA1 and FHV RNA2 genome segments and unveiled the precise dynamics and mechanisms underlying the emergence of DVGs during serial passaging of the virus in cell culture over a 1-month period.

One limitation of NGS to study SVs is the well-known propensity of both long- and short-read sequencing technologies, to generate artificial chimeras, which are difficult to distinguish from biological recombination events, during library construction (Tsai et al. 2014; Griffith et al. 2018; Peccoud et al. 2018). In addition, the quantity of viral particles that are directly recovered from natural hosts is often relatively small, which makes it difficult to purify enough viral DNA to prepare sequencing libraries. All NGS studies of SVs in viral populations have thus so far been done using viruses passaged in cell lines. Here, we sought to estimate the diversity of SVs that segregate in large dsDNA virus populations following natural host infections. First, we sequenced a large population of AcMNPV genomes purified from *Spodoptera exigua* larvae using both short-read Illumina and long-read PacBio sequencing technologies in parallel. Using a novel pipeline involving hierarchical clustering of SVs detected by six variant callers, we counted SVs present in both sequencing datasets. As PacBio and Illumina technologies are subject to different biases, we reasoned that SVs retrieved

from both datasets are unlikely to derive from technical artifacts and can be considered robust. Based on the results obtained for AcMNPV, we then estimated SVs in populations of two other invertebrate large dsDNA viruses, the invertebrate iridescent virus 31 (IIV31) and 6 (IIV6) extracted from adults of the pillbug *Armadillidium vulgare* and from larvae of the moth *Sesamia nonagrioides*, respectively, and in a population of the HCMV purified from MRC5 cells.

## 2. Materials and methods

### 2.1 Infection of *S. exigua* larvae with AcMNPV

The AcMNPV-WP10 isolate (Chateigner et al. 2015) was used to infect 150 fourth instar larvae of the beet armyworm (S. exigua) using the diet plug method (Sparks, Li, and Bonning 2008). Each moth larva was fed 100,000 occlusion bodies (OBs) per 5 $mm^3$ diet plug. Upon host death, which occurred 2–5 days post-infection, OBs were first filtered through cheesecloth, purified twice by centrifugation (10 min at 7,000 rpm) with 0.1 per cent sodium dodecyl sulfate, then distilled water, and finally resuspended in water. Approximately 1.5 $10^{10}$ OBs were treated as described in Gilbert et al. (2014) to provide about 50 mg of high-quality dsDNA (about 5.82 $10^{11}$ genomes assuming 100 genomes per OB; Ackermann and Smirnoff, 1983; Slack and Arif, 2006; Rohrmann, 2014). Briefly, OBs were purified by a percoll gradient at pH 7.5, sucrose 0.25 M (9 V of percoll/sucrose solution were added to 1 V of virus solution) with a centrifugation step (30 min at 15,000 g, 4 C). OBs were dissolved using $Na_2CO_3$ to release nucleocapsids (O'Reilly, Miller, and Luckow 1992). Viral DNA was then extracted using the QIAamp DNA Mini kit (Qiagen).

### 2.2 Infection of *A. vulgare* with IIV31

A solution containing IIV31 viral particles was obtained through grinding a piece of cuticle from a naturally infected *A. vulgare* individual collected on the campus of the University of California Riverside. One *A. vulgare* individual was pricked with a thin needle soaked in the viral solution. Fourteen days after the infection, the pillbug became bluish and died about 4 weeks after infection, as described in Lupetti et al. (2013). Upon death, the pillbug was crushed with a pestle and put in a 1.5 ml Eppendorf tube in a Tris solution. An ultra-centrifugation step on sucrose cushion was then performed at 35,000 g for 90 min at 4 C. The pellet was resuspended in 100 ml of Tris solution. Viral DNA was then extracted using the QIAamp DNA Mini kit (Qiagen).

### 2.3 Infection of *S. nonagrioides* larvae with IIV6

Ten fourth instar larvae of the Mediterranean corn borer *S. nonagrioides* were infected with the IIV6 viral strain originally described in Fukaya and Nasu (1966 ). Larvae were pricked using a thin needle soaked in the viral solution. Fourteen days later, the larvae presented a purple iridescence and they finally died about four weeks after infection. Upon host death, viral particles were filtered through cheesecloth and two centrifugation steps were performed to eliminate most of host cells and tissues. Then, an ultracentrifugation step was performed as described above for IIV31. Viral DNA was then extracted using the QIAamp DNA Mini kit (Qiagen).

### 2.4 Infection of MRC5 cells with HCMV

MRC5 human fibroblasts were cultured in Dulbecco's modified Eagle medium (Invitrogen) supplemented with 10 per cent fetal bovine serum, 4.5 g/l glucose, and 1 per cent penicillin-streptomycin (Pen-Strep; Life Technologies) at 37°C in a 5 per cent (vol/vol) $CO_2$ atmosphere. Before HCMV infection, MRC5 cells were grown to confluence, resulting in 3.0 $10^4$ cells per $cm^2$. Once confluent, the medium was removed, and serum-free medium was added. Cells were maintained in serum-free medium for 24 h before infection at which point, they were infected at a multiplicity of infection of 10 pfu/cell with a clinical strain of HCMV isolated from a patient in 2015. After a 2 h adsorption period, the inoculum was aspirated, and fresh serum-free medium was added. Cells were harvested 8 days after infection through trypsinization followed by washing in Earle's balanced salt solution and centrifugation at 1,100 g. Pelleted cells were then transferred into a 15 ml Falcon tube and cell lysis was performed by several steps of freeze/thaw cycles in dry ice and water bath at 37°C. The solution was centrifuged at 5,000 g for 30 min at 4 C and the supernatant containing viral particles was collected. Purification of viral particles and viral DNA extraction was performed as described above for iridoviruses.

### 2.5 Sequencing

For each virus, an aliquot containing 2 mg of DNA was used to construct a paired-end library (insert size 260 bp), which was sequenced on a Illumina HiSeq™ 2500 machine (Illumina, San Diego, CA, USA), generating 298, 298, 582, and 308 million 151-bp paired reads for AcMNPV, HCMV, IIV6, and IIV31, respectively. For PacBio sequencing, about 15 mg of AcMNPV DNA was used to construct one library. This library was sequenced at the McGill University and Genome Quebec Innovation Center on eight SMRT cells using the PacBio Sequel instrument, which generated 3 million reads (31 Gb).

### 2.6 Assembly and annotation of the consensus viral genomes

A consensus viral genome was assembled for all four viruses sequenced in this study. For AcMNPV, the viral assembly was based on the long reads that altogether reached a 203,467 depth of the AcMNPV genome. All PacBio reads longer than 30 kb (68,173 out of 3,012,899 reads, corresponding to 21 coverage depth on the viral genome) were assembled with Canu v1.5 (Koren et al. 2017; main options: -d AsmCanu -auto -genomeSize = 134k -pacbio-raw). The raw assembly was then polished with Pilon v1.22 (Walker et al. 2014; default options) using the short reads that altogether reached a 196,093 depth of the AcMNPV genome. The polished assembly was then circularized with ToAmos v3.1.0 (Treangen et al. 2011) and minimus2 (Sommer et al. 2007) with default options. The assembly was then annotated based on the AcMNPV-E2 strain genome (accession number KM667940.1) with the General Annotation Transfer Utility program (Tcherepanov, Ehlers, and Upton 2006). The HCMV, IIV6, and IIV31 genomes were assembled as follows. For each virus, read subsamples corresponding to 500 and 1,500 depth coverage were assembled with tadpole (https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/tadpole-guide/, version of December 2018, options used: 'k = 17'; 'k = 31'; 'k = 60'; 'k = 90' with 'mincov = 100'). The different assemblies obtained with 17, 31, 60, and 90 mers with 500 and 1,500 depth coverage were then assembled with Geneious version 11.0.2 (https://www.geneious.com, options: de novo assembly, Geneious assembler, high sensitivity).

The final assemblies were annotated based on the available HCMV (NC_006273.2), IIV6 (NC_003038.1), or IIV31 (NC_024451.1) genomes with the General Annotation Transfer Utility program (Tcherepanov, Ehlers, and Upton 2006).

### 2.7 SV detection

Illumina reads were aligned on the viral genomes assembled in this study using BWA (Li and Durbin 2009, options: -R '@RG\tID: id\tSM: sample\tLB: lib') and blastn (options: -outfmt 6 -max_-target_seqs 2 -max_hsps 2). SVs were called with four different pipelines: Pindel (Ye et al. 2009), Lumpy (Layer et al. 2014), Fermikit (Li, 2015) and a custom Python script. Pindel and Lumpy were run on bam files produced by BWA (options: -R '@RG\tID: id\tSM: sample\tLB: lib'). Fermikit is an SV caller based on local read assembly, which uses raw reads as input. The custom python script is derived from that used in Chateigner et al. (2015) to find large deletions using short-read pairs. The script runs on tabular blastn output files and identifies deletions by comparing the observed distance separating both reads of a pair with the distance expected according to the mean library insert size. To account for experimental variation in insert size not due to structural variation, we only considered inter-read dis-tances longer than 700 bp as reflecting true SVs (Supplementary Fig. S3.1). Deletions of smaller genome fragments cannot be confidently identified with this script. All read pairs involved in a deletion event of approximately the same start position, end position (maximum length between start positions or end positions was 7 bp) and length were clustered in SV events each characterized by an average start and end position, as well as an average length and a number of read pairs supporting the deletion event. As we had no expectation regarding the final number of clusters, we followed previous studies (Mönchgesang et al. 2016; Parikh et al. 2016; Zhang et al. 2018) and used a hierarchical clustering method rather than the K-means method, which is based on a known number of clusters. Briefly, Euclidean distances of start and end positions were computed between all SVs to generate a distance matrix. Then the linkage step between SVs was performed according to the Ward method (Ward 1963). The threshold value was automatically defined using the inconsistency method (Jain and Dubes 1988). Clustering was performed with the 'scipy.cluster.hierarchy' Python package (https://docs.scipy.org/doc/scipy/reference/cluster.hierar chy.html).

PacBio reads were aligned on the AcMNPV genome using BWA and SVs were called on the bam file with sniffles (Sedlazeck et al. 2018). SVs were also called with PbHoney (English, Salerno, and Reid 2014), which takes raw long reads as input. Both SV callers were run using default parameters. All SV caller output files were treated as Variant Call Format files in downstream analyses.

### 2.8 SV analyses

To remove redundancy in SVs (a given SV may be detected by more than one SV caller), all SVs supported by three reads or more were clustered using a hierarchical clustering approach implemented in the 'fastcluster' R package (Müllner 2013), the R version of the Python package used for the custom SV Python SV caller. To take into account the relative imprecision in the coordinates of some clusters, we did not define clusters based on the inconsistency method, instead we performed multiple rounds of clustering (fourteen rounds, see below), each time using a different threshold value (Fig. 3.1A). The use of different thresholds allowed us to take into account the fact that high threshold values can induce erroneous clustering of different SV events which coordinates are very close to each other. Importantly however, clusters containing different types of SVs (e.g. a deletion and an inversion which have the same coordinates) are removed from the analyses. On the other hand, small threshold values can miss clustering of identical SVs detected by different programs due to slight differences in coordinate precision between programs, as more particularly noted in the case of the AcMNPV long-read dataset. The error-prone long reads can be mapped approximately due to artefactual SNPs and insertions/deletions (indels) present in the reads. Such approximations can lead to different start and end coordinates for a same SV between the different sequencing technologies and the different alignment programs. Due to these slight differences in coordinates for the same SV, a low threshold value will not cluster these different coordinates sets into one cluster but will give many clusters each with one pair of coordinates. With only one threshold value, downstream filters would often erroneously remove some clusters (i.e. some SVs) because all clusters supported by only one SV caller are not considered robust and discarded in our approach. For example, if a deletion was detected with a long-read SV caller at coordinates 5–50 and with a short-read SV caller at coordinates 6–54, a too low threshold value would not cluster both coordinate sets in one cluster (one SV) supported by the two SV callers but it would cluster them in two different clusters each supported by one SV caller. Then, a filter in downstream analyses would remove all clusters not supported by at least one long-read and one short-read SV callers. Thus, the two SVs detected would be erroneously removed whereas they in fact correspond to the same biological SV but with slightly different mapping coordinates. With a higher threshold, these two SVs detected would be clustered together and kept as one SV. That is why we used different threshold values.

To avoid redundancy in SV detection due to the use of many threshold values, duplicated SVs were removed. For each data-set, we performed a total of fourteen different clustering steps each with a different threshold value (5; 10; 30; 50; 100; 200; 300; 400; 500; 600; 700; 800; 900; and 1,000). To further improve the delineation between different SVs that may involve close genome breakpoint coordinates; we included read coverage in our analysis, reasoning that different SVs may often be supported by a different number of reads (Fig. 3.1B). Thus we repeated the above-described round of clustering (involving all different four-teen clustering threshold values) twenty-two times using twenty-two different thresholds for the minimal number of reads supporting each SV (4; 5; 6; 7; 8; 9; 10; 20; 30; 40; 50; 100; 200; 300; 400; 500; 600; 700; 800; 900; 1,000; and 1,500). The number of merged SVs differed depending on these different threshold values. Some SVs could be removed by downstream filters when too many discordant SVs were merged (more likely when the minimum read number threshold was low) whereas the same SVs had a lower chance to be removed by downstream filters when higher numbers of reads were used, inducing a less aggressive clustering. All SVs obtained through these clustering steps were retrieved and a final list of SVs was established after removing redundancy that is, all identical SVs found under different thresholds were counted only once.

For the AcMNPV virus, the clustering procedure was performed jointly on the Illumina and PacBio datasets. Also, to avoid

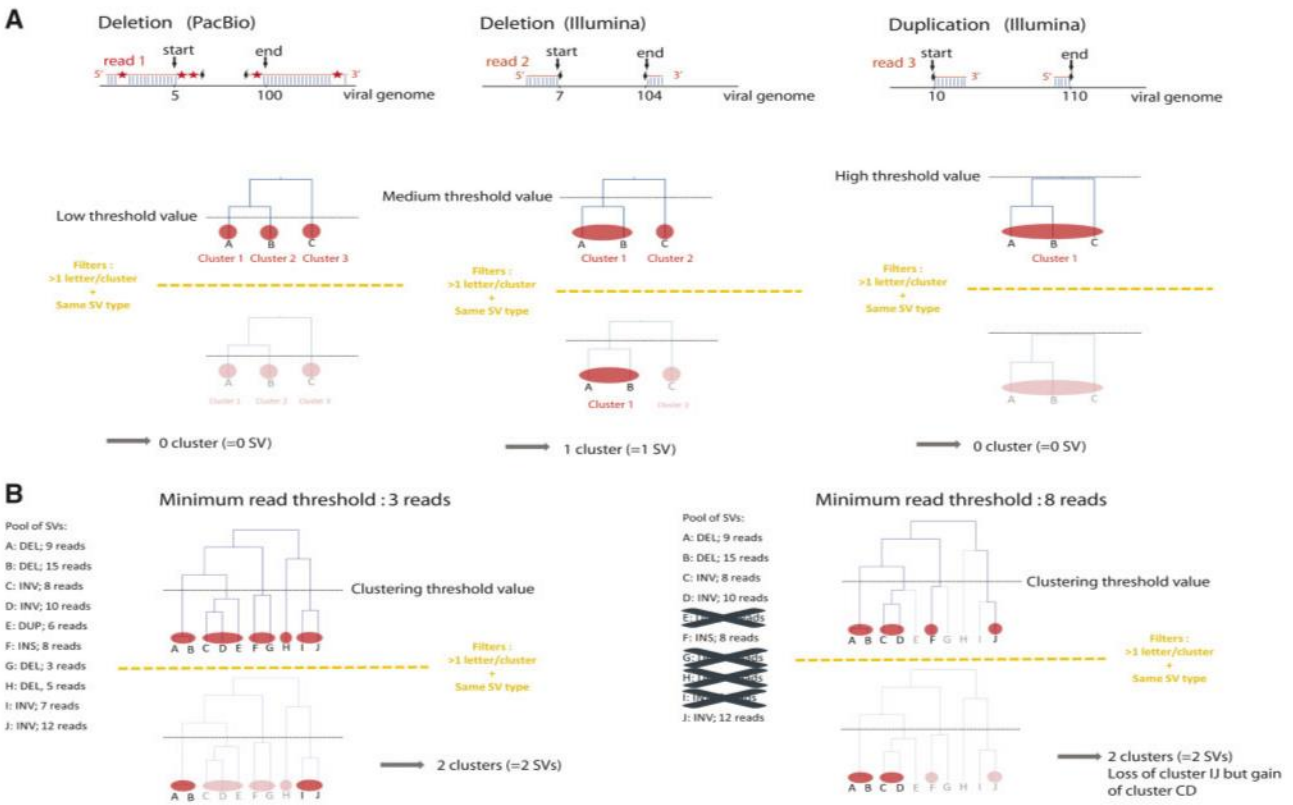



Figure 3.1. Illustration of two important steps of the hierarchical clustering of SVs. (A) Influence of the clustering threshold value. The top panel illustrates three reads (one long PacBio read and two short Illumina reads) mapped onto overlapping regions of the viral genome. Red asterisks correspond to sequencing errors that prevent accurate mapping of long reads. 'start' and 'end' correspond to start and end coordinates of the SV detected by SV callers (a deletion in the case of Reads 1 and 2 and a duplication for Read 3. The bottom panel shows how using multiple clustering thresholds prevents discarding well-supported SVs. With a low threshold, all clusters contain a single SV because none of the SVs have the exact same coordinates. Because a downstream filter of our pipeline requires that SVs must be detected either by both long and short reads (in the case of the AcMNPV population sequenced using both Illumina and PacBio technologies) or by two programs (in the case of the three other large dsDNA viruses sequenced only with Illumina) to be retained, none of the SVs are retained with this low clustering threshold. With a high clustering thresh-old, all SVs (two deletions and one duplication) end up in the same cluster because they are defined by coordinates that are close to each other. Because a downstream filter of our pipeline requires that all SVs within a cluster must be of the same nature for a cluster to be retained, the cluster is here not considered further. With a medium threshold, the deletions detected by Reads 1 and 2 are lumped into the same cluster because their coordinates are close enough and the duplication detected by Read 3 forms another cluster because its coordinates are too far from those of the deletion. After running the downstream filters of our pipeline, Cluster 2 is retained and one deletion is counted because it has been detected independently by long and short reads. The cluster containing the duplication is not considered further because it contains only one SV detected by short reads only. Note that although SVs supported by only one read are represented here for the sake of simplicity but our approach only retained SVs supported by a minimum of three reads. (B) Influence of the minimum number of reads supporting a SV. On the left panel, using three reads as the minimum number of reads required to retain SVs, ten SVs of different nature and/or supported by different numbers of reads have been detected by SV callers. Under a given clustering threshold value, these ten SVs form five clusters, only two of which are retained (A–B and I–J) by downstream filters because they contain several SVs which are all of the same nature. On the right panel, only six of the ten SVs detected on the left panel are detected by SV callers using eight reads as the minimum number of reads required to retain SVs. With the same given clustering threshold value as in the left panel, SVs form four clusters, two of which (A, B and C, D) are retained by downstream filters because they contain several SVs which are all of the same nature. Using multiple minimum numbers of reads supporting SVs ensure that well-supported SVs (here the inversion in C and D) are not eliminated by downstream filters.

false SV discovery due to a detection error caused by the circular nature of the AcMNPV genome, additional filters were added for this virus. Some long reads involved in the SVs were retrieved and aligned on the AcMNPV genome with Geneious version 11.0.2 (https://www.geneious.com). Some of them corresponded to the start and end coordinates of the AcMNPV consensus genome, thus did not capture an SV event and overestimated the number of reads supporting an SV. Empirically, we found that false SVs were mainly supported by a few number of reads (ten to twenty reads). SVs with a length >67 kb (half the size of the AcMNPV genome) and supported by <20 long reads or by <3 SV callers were discarded from the analysis. After obtaining a final list of SVs for each virus, average start position, end position and length were calculated for each SV. Finally, viral genes corresponding to the average start and end SV positions were identified based on the viral genome general feature format file.

### 2.9 SV frequencies in viral populations

Our calculation of the SV frequency was based on the approach commonly used to calculate SNP frequency that is, SNP cover-age/(SNP coverage þ reference coverage) at the SNP position. Thus, we calculated SVs frequency as follows: SV coverage/(SV coverage þ reference coverage) at the SV position, using a per-base coverage file of all alignments obtained with bedtools genomecov (Quinlan and Hall 2010; option: -d) for the four SV callers relying on the use of a bam mapping file.

### 2.10 Simulation of AcMNPV short reads

We simulated a mock dataset of short reads from the AcMNPV genome with 200,000 depth, equivalent to our real dataset. The mock reads were generated with the Grinder program (Angly et al. 2012) with point mutations and chimeras, to mimic

a real Illumina dataset (options: '-coverage_fold 200,000 - read_dist 150 uniform 0 -insert_dist 230 normal 50 -mate_orientation FR -chimera_perc 5 -chimera_dist 1 -chimera_kmer 0 -mutation_dist uniform 0.3 -mutation_ratio 99.7 0.3'). The simulation yielded 89,322,000 150-bp reads, on which we ran our SV detection pipeline.

### 2.11 Characterization of SVs in twenty-one AcMNPV datasets

A published experimental evolution dataset of AcMNPV, whereby a population of this virus purified after several rounds of infection on the cabbage looper moth (*Trichoplusia ni*) was Illumina-sequenced at 187,536 average depth and was independently passaged in ten lines of *T. ni* larvae and ten lines of *S. exigua* larvae, each line consisting of ten successive infection cycles (Gilbert et al. 2016). AcMNPV OB's recovered from the last infection cycle of each of twenty evolved AcMNPV populations were sequenced at between 9,211 and 33,783 average depth for the ten *T. ni* lines (total depth = 145,386 X) and between 3,497 and 35,434 average depth for the ten *S. exigua* lines (total depth =163,610 X). To detect SVs in each of the twenty-one AcMNPV Illumina datasets, we applied the method described above for the AcMNPV Illumina dataset, involving hierarchical clustering of the outputs of four SV callers. As long reads were not available for any of these twenty-one datasets, we restricted our analysis of SV frequency to the 4.98 per cent most robust SVs detected in each dataset by selecting SVs supported by two variants callers (in line with the 4.98 per cent SVs jointly detected in both short and long reads among all SVs in the first analysis above). When the number of SVs supported by two variant callers was <4.98 per cent of all SVs, we selected all SVs detected by two variant callers plus another set corresponding to the most frequent SVs to reach 4.98 per cent.

### 2.12 Transposable element insertions

Our search for host sequences integrated into viral genomes involved aligning viral reads on various databases of publicly available sequences from the very host species used in this study or from species related to these hosts. For *S. nonagrioides*, our database included all nuclear and mitochondrial genomic and transcriptomic data of all lepidopterans available in GenBank as of 20 January 2018 (Benson et al. 2005) and the databases of all beet armyworm and cabbage looper contigs used in Gilbert et al. (2016). To increase our chances to detect host transposable element (TE) insertions, we used all TE sequences available in Repbase as of 15 October 2017 (Bao, Kojima, and Kohany 2015) and those identified with RepeatModeler (http://www. repeatmasker.org) in 196 insect genomes in Peccoud et al. (2017). For human, we used the GRCh38.p12 version of the human genome (GenBank assembly accession: GCA_000001405.27). For the pillbug, we used the *A. vulgare* genome (Chebbi et al. 2019). For the different host/virus systems studied, we also retrieved non-viral reads and assembled them with the SPAdes assembler (Bankevich et al. 2012). Then we aligned the resulting contigs on the GenBank nr database. We also aligned these contigs against themselves to search for terminal inverted repeats or long terminal repeats that are specific sequences found at the end of full-length TE sequences (Craig, 2002). Contigs corresponding to full-length TE sequences were added to previous databases to refine the search for TEs integrated in viral genomes.

Junctions between viral and host sequences were searched in Illumina short reads following Gilbert et al. (2016). Briefly, the raw Illumina reads were trimmed to remove adapters. Then they were aligned separately to host genomic and transcriptomic databases and to the viral genome using blastn (option 'megablast'). Only reads aligning over at least 16 bp on the viral genome only and over at least 16 bp on a host sequence only were retained. Reads had to align on at least 130 bp (out of a total length of 151 bp) of their length. The overlap between alignment on the virus and on the host sequences was set to involve at most 20 bp and at least 5 bp (see Supplementary Fig. S6 in Gilbert et al. 2016).

Host sequences integrated into AcMNPV genomes were further searched by mapping long PacBio reads on the host databases with BLASR (Chaisson and Tesler 2012). BLASR tabular outputs obtained for each host database were merged, overlapping hits were identified and among them only the best-score hit was retained. Regions not mapping on host sequences were aligned to the viral genome with BLASR program to validate the host/virus chimeric nature of the reads.

The observed proportion of TE sequences at read ends was calculated by counting the number of TE sequences that were at a read end among all the TE sequences. The expected proportion of TE sequences at read ends was calculated by dividing the total TE sequences length by the total read length. A binomial test was performed to compare the observed and expected pro-portions of TE sequences at read ends. Statistical analyses were performed in R version 3.4.4 (R Core Team 2018).

## 3. Results

### 3.1 AcMNPV consensus genome

Our hybrid assembly of the AcMNPV genome yielded a 133,981-bp consensus genome which is 99.92 per cent identical to and 15 pb longer than the AcMNPV-E2 strain (Maghodia, Jarvis, and Geisler 2014). Both genomes diverge by sixty-three SNPs and sixteen short (<10 bp each) indels. All sixteen indels are supported by >80 per cent of Illumina reads covering these variants. These indels affect eight genes and disrupt the open reading frame in five of them (Ac-bro, Ac-odv-e18, Ac-gp64, AcOrf-91, and Ac-lef4). The sixty-three SNPs involve fourteen genes (AcOrf-34, AcOrf-18, Ac-IE-1, Ac-49K, Ac-IE-0, Ac-ME53, Ac-chitinase, AcOrf-114, Ac-helicase, AcOrf-74, Ac-lef3, A-lef8, Ac-pcna, and Ac-odv-e66). Only two of these sixty-three SNPs are fixed in the population, whereas the remaining sixty-one coexist at high frequencies (65.0–98.5%) with the alternative variant of the AcMNPV-E2 genome.

### 3.2 Nature, number and frequency of SVs in the AcMNPV population

Our search for SVs in the AcMNPV short-read Illumina data using our clustering pipeline applied to the results of four SV callers (Lumpy, fermikit, pindel, and custom Python script) yielded 22,892 variants, among which 1,141 (4.98%) were considered robust as they overlapped with the 9,421 SVs detected in the PacBio long-read data. The 1,141 SVs comprised 464 deletions, 443 inversions, 160 duplications, and 74 insertions (Fig. 3.2A, Table 3.1 and Supplementary Table S3.1). Examples of read alignments supporting twelve AcMNPV SVs are shown in Supplementary Figs S3.2–S3.14. SV size ranged from 50 bp (the minimum size cutoff that we used) for an insertion to 66,787 bp for an inversion (close to the maximum size cutoff), with an average of 30,496 bp (Fig. 3.2B). SVs were detected all along the

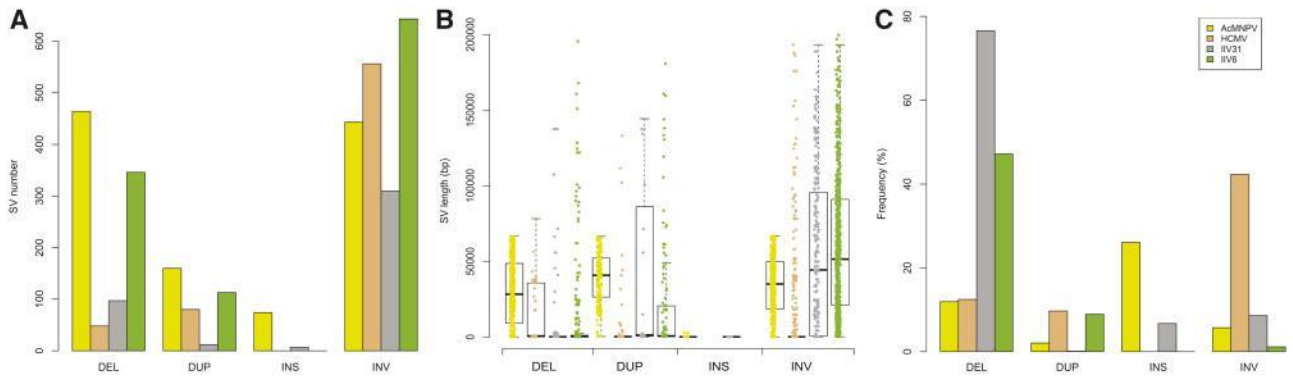



Figure 3.2. Number, size, and frequency of SVs in the four viral populations. (A) Number of detected SVs by SV type for the four viruses. No insertions were detected in the HCMV and IIV6 viral population. Insertions were only detected in long-read AcMNPV and in the IIV31 short reads. (B) Boxplots representing the size of detected SVs by SV type for the four viral populations. (C) Frequency of viral genomes carrying SVs shown by SV type for the four viral populations. The frequency was computed considering SV number per viral genome follows a Poisson distribution. DEL, deletion; DUP, duplication; INS, insertion; INV, inversion.

AcMNPV genome, with no apparent hotspot (Fig. 3.3 and Supplementary Fig. S3.15).

Most AcMNPV SV variants occurred at low to very low frequencies, with 92.4 and 75.4 per cent of SVs having a frequency <0.1 and 0.01 per cent, respectively (Fig. 3.4). Yet, taking all 1,141 SVs into account and assuming that the number of SVs per viral genome follows a Poisson distribution, we calculated that no less than 39.9 per cent of AcMNPV genomes are affected by a variant (Supplementary Table S3.1). It is noteworthy that in spite of being less numerous than other SVs in the AcMNPV population, insertions were generally segregating at higher frequency, with an overall estimate of 26.1 per cent of AcMNPV genomes being affected by an insertion (Fig. 3.2C and Supplementary Table S3.1). The most frequent SV in this viral population was an insertion of 70 bp that increased the length of the hr4b homologous region in 6.0 per cent of AcMNPV genomes. Interestingly, the five most frequent SVs in this population involve intergenic regions, non-essential or uncharacterized genes (Supplementary Table S3.2), which may reflect the lower effect of these SVs on viral fitness. We also found that the total frequency of all SVs involving genes, hr, or intergenic regions was fairly homogeneous and mostly comprised between 1.9 and 6.2 per cent (Supplementary Fig. S3.16). The only exception to this pattern is the hr4b region mentioned above, which is involved in 535 SV affecting 10.8 per cent of viral genomes. Next, we counted the number of SVs involving each of the 151 AcMNPV genes and classified SVs as either inactivating (SVs inducing gene truncations) or non-inactivating (i.e. the coding capacity of the gene remains intact). We found that 148 out of the 151 genes were more affected by non-inactivating than by inactivating SVs. For these 148 genes, there was on average 55 inactivating and 100 non-inactivating SVs. Notably, the three remaining genes are located at the extremities of the linear AcMNPV genome as we have used it for the analyses. Thus the higher number of inactivating SVs in these genes is due to a technical effect. We also looked at the cumulative frequency of inactivating and non-inactivating SVs affecting genes. The results were consistent with the raw numbers of SVs, with the vast majority of genes (N = 149) more frequently affected by non-inactivating SVs than by inactivating SVs.

We then used the long-read dataset to assess the extent to which a given viral genome can be affected by multiple SVs. This analysis was based on the set of SV-carrying long reads detected by Sniffles only, as only this program provides read names in the output SV list. All 15,044 reads supporting the 1,648 SVs detected by Sniffles were retrieved. The vast majority of these reads (N = 14,783) carried a single SV, and the remain-ing 261 reads (1.74%) carried more than one SVs. Among these, only 13 and 1 reads, respectively carried three and four SVs. For 161 of the 261 reads carrying more than 1 SV, SV coordinates overlapped, indicating nested SVs.

### 3.3 Comparison of simulated and observed SVs

To assess the extent to which technical chimeras produced during the construction of the Illumina sequencing library may have introduced biases in SV count and frequency calculation, we generated a mock short-read dataset in which a proportion of chimeric reads were introduced (see Section 2.10). To compare the numbers, nature and frequency of SVs detected with this simulated dataset to the 1,141 robust SVs detected with the real dataset, we selected 4.98 per cent of all detected SVs using the simulated datasets (i.e. the proportion of SVs detected by both sequencing technologies among all SVs detected using short reads only, see above). The 4.98 per cent most frequent SVs supported by at least 2 SV callers were selected, which yielded 802 SVs, corresponding to 737 deletions and 65 duplications (Supplementary Table S3.3). Taking all these SVs into account, we calculated that 1.47 per cent of AcMNPV genomes carry one SV detected with the simulated dataset (assuming the number of SVs by viral genome follows a Poisson distribution). These results show that technical chimeras can induce a substantial number of false positives using our SV detection pipeline. However, the SV profile and frequency of viral genomes calculated to carry these variants widely differ between the simulated (737 deletions and 65 duplications; 1.47%) and real dataset (464 deletions, 443 inversions, 160 duplications, 74 insertions; 39.9%), strongly suggesting that the vast majority of SVs detected by both sequencing technologies in the real data-set are indeed biological. Note that the number of SVs due to technical chimeras that we detect in the mock dataset is likely overestimated because we chose to simulate reads with 5 per cent of chimeras, which corresponds to some of the highest technical chimera rates observed in previous studies (Görzer et al. 2010; Peccoud et al. 2018).

The construction of PacBio sequencing libraries, which involves a blunt-end ligation step, can also induce the formation of a substantial number of artefactual chimeras (Tallon et al. 2014; Griffith et al. 2018). However, the conditions and rates at which such chimeras are generated have been less

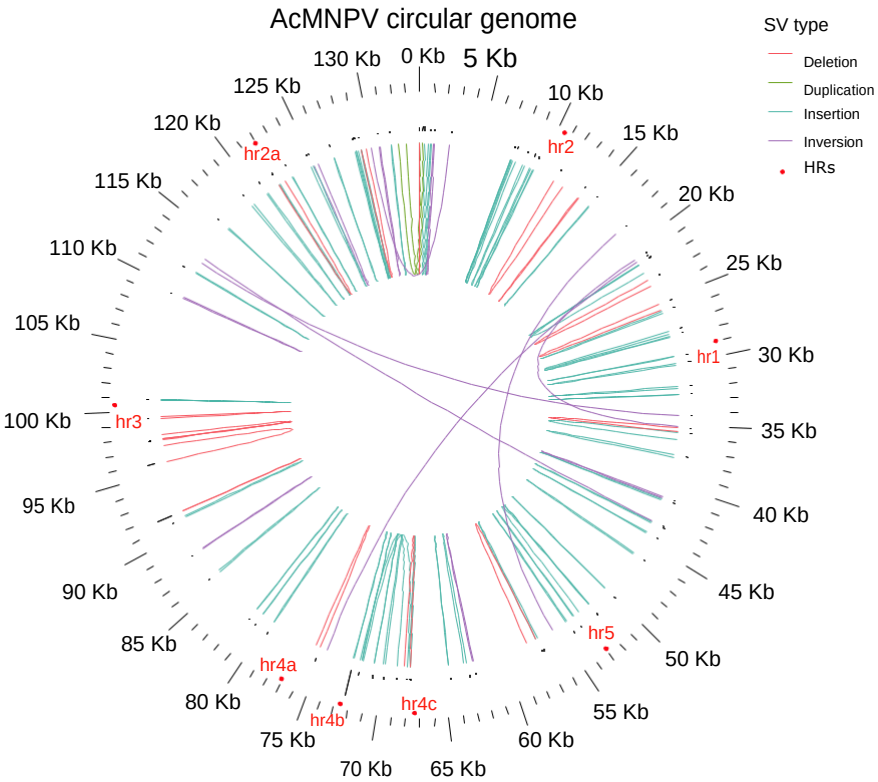



Figure 3.3. Map of the circular AcMNPV genome illustrating all SVs present in more than 0.1 per cent of the viral population sequenced with Illumina and PacBio technologies. Each SV is illustrated by a curve linking their start and end coordinates. Histograms on top of SVs correspond to the relative frequency of each SV, with the most frequent SV involving hr4b.

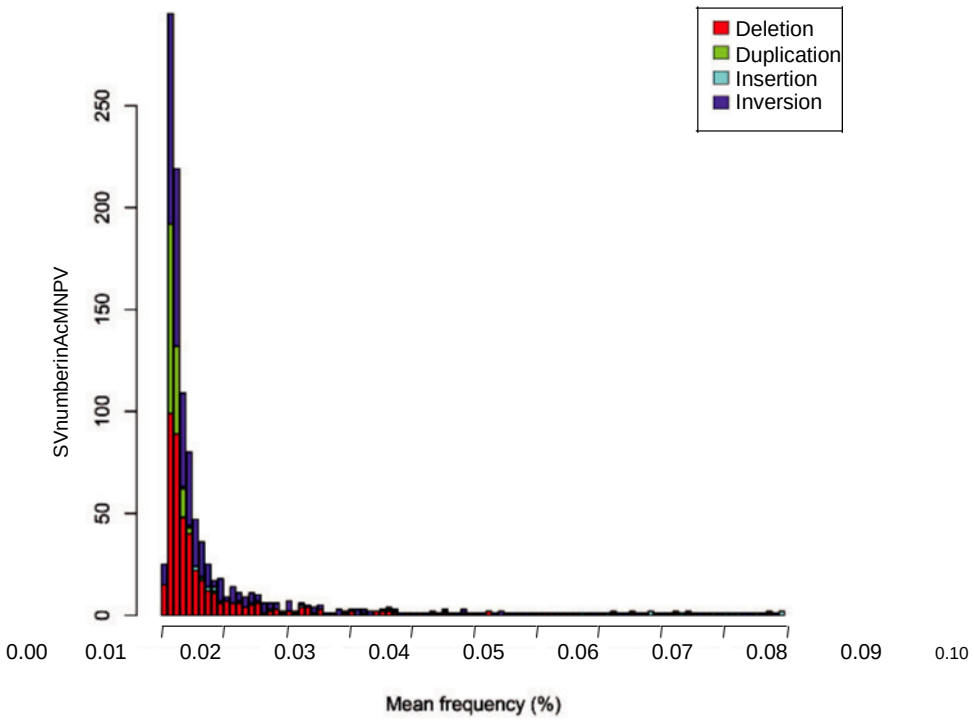



Figure 3.4. Number of SVs by 0.001 per cent frequency bin detected in the AcMNPV population sequenced using Illumina and PacBio technologies. Only the first 100 frequency bins are shown. The vast majority of SVs (92%) are present in viral genomes at a very low frequency (<0.01%).

Table 3.1. Numbers, frequencies and lengths of SVs detected in AcMNPV, HCMV, IIV6, and IIV31 populations.

| SV type | Number of SVs | Frequency (%) | Minimum length (bp) | Average length (bp) | Maximum length (bp) |
|---|---|---|---|---|---|
| AcMNPV | | | | | |
| DEL | 464 | 11.91 | 52 | 29,318 | 66,712 |
| DUP | 160 | 2.05 | 290 | 38,898 | 66,488 |
| INS | 74 | 26.11 | 50 | 159 | 2,723 |
| INV | 443 | 5.70 | 55 | 33,762 | 66,787 |
| Total | 1,141 | 39.87 | 50 | 30,496 | 66,787 |
| HCMV | | | | | |
| DEL | 48 | 12.43 | 54 | 16,211 | 78,246 |
| DUP | 80 | 9.69 | 156 | 7,386 | 132,989 |
| INS | 0 | 0 | 0 | 0 | 0 |
| INV | 556 | 42.30 | 55 | 10,825 | 205, 218 |
| Total | 684 | 54.37 | 54 | 10,800 | 205, 218 |
| IIV6 | | | | | |
| DEL | 346 | 47.16 | 52 | 10,060 | 207,988 |
| DUP | 113 | 8.90 | 171 | 21,386 | 180,808 |
| INS | 0 | 0 | 0 | 0 | 0 |
| INV | 643 | 1.14 | 105 | 61,168 | 206,392 |
| Total | 1,102 | 52.41 | 52 | 41,042 | 207,988 |
| IIV31 | | | | | |
| DEL | 97 | 76.58 | 55 | 15,004 | 209,948 |
| DUP | 12 | 0.11 | 212 | 36,295 | 139,680 |
| INS | 7 | 6.72 | 58 | 63 | 66 |
| INV | 310 | 8.60 | 100 | 58,550 | 203,305 |
| Total | 426 | 80.06 | 55 | 47,046 | 209,948 |

The frequency refers to the percentage of viral genomes affected by the SV type, assuming that it follows a Poisson distribution. DEL, deletions; DUP, duplications; INS, insertions; INV, inversions. The frequencies were computed assuming the number of SVs per viral genome follows a Poisson distribution. Details about each SV detected in the four viral genomes are provided in Supplementary Tables S3.1, S3.4–S3.6.

studied than those produced during the construction of Illumina libraries. Currently available simulators of long PacBio reads do not offer the possibility to generate chimeras (Ono, Asai, and Hamada 2013; Stöcker, Köster, and Rahmann 2016; Wei and Zhang 2018; Zhang, Jia, and Wei 2019). Thus, we did not estimate the number of SVs possibly due to artificial long-read chimeras that we can detect with our pipeline.

### 3.4 Characterization of SVs in twenty-one AcMNPV datasets

The finding of a large number of SVs in AcMNPV populations raised the question of their persistence over several rounds of infection. To investigate SV dynamics during viral evolution, we used a published experimental evolution dataset of AcMNPV, whereby a population of this virus purified after several rounds of infection on the cabbage looper moth (*T. ni*) was Illumina-sequenced at 187,536 average depth and was independently passaged in ten lines of *T. ni* larvae and ten lines of *S. exigua* larvae, each line consisting of ten successive infection cycles (see Gilbert et al. 2016 and Section 2). Overall, this analysis revealed that the number of SVs shared by the parental *T. ni* population and any of the twenty evolved populations was always low (from 1 [0.07%] to 46 [3.9%] out of the 1,158 SVs detected in the parental *T. ni* population; Fig. 3.5A and B), and that the vast majority of SVs were only present in one population (Fig. 3.5D). Of note, one SV present in the G0 population was found in eight *T. ni* datasets and in six *S. exigua* datasets. It is a duplication of 62,654 bp involving the hr2 and hr4b regions in very low frequency in the parental population (0.008%) that increased in frequency in some evolved populations (>4%, represented in red in Fig. 3.5C).

### 3.5 Analyses of SVs supported by short reads in populations of HCMV, IIV6, and IIV31

To assess the extent to which SVs may affect viruses other than AcMNPV, we generated short-read datasets for two invertebrate iridoviruses, IIV6, and IIV31, respectively passaged on caterpillars of the Mediterranean corn borer (*S. nonagrioides*) and the pillbug *A. vulgare*, and for human CMV passaged on MRC5 cells. The IIV6 genome we assembled was 210,812 bp in length, 99.51 per cent identical to and 1,670 bp shorter than the closest reference genome available in NCBI, that of the *Chilo iridescent* virus IIV6 (accession number AF303741.1). Over the 468 genes annotated in the reference IIV6 genome, 435 were recovered in our IIV6 assembly and used for downstream analyses. Among these four hundred and thirty-five genes, twenty-two have a weak protein similarity (<60%) with those in the reference genome. The differences between our assembly and the reference genome were due to seventy-one insertions and fifty-four deletions, including twenty-one insertions and fifteen deletions located within genes without changing the open reading frame. Our assembly of the IIV31 genome was 219,807 bp in length, 99.90 per cent identical to and 415 bp shorter than the *A. vulgare* iridescent virus reference genome (accession number HF920637.1). A total of 193 out of 203 genes from the reference genome were retrieved in the assembly. The difference in length with the reference genome was due to thirty-seven insertions and forty-six deletions, among which eight insertions and eleven deletions were localized in open reading frames (without disruption). Finally, the HCMV genome we assembled was 234,915 bp in length, 97.69 per cent identical to and 731 bp shorter than its closest reference genome available in NCBI, that of the Merlin strain (accession number KP745639.1). A total of 151 genes out of 154 annotated in the Merlin strain were recovered in our assembly. The difference in length was due to ninety-two insertions and eighty-three deletions among which

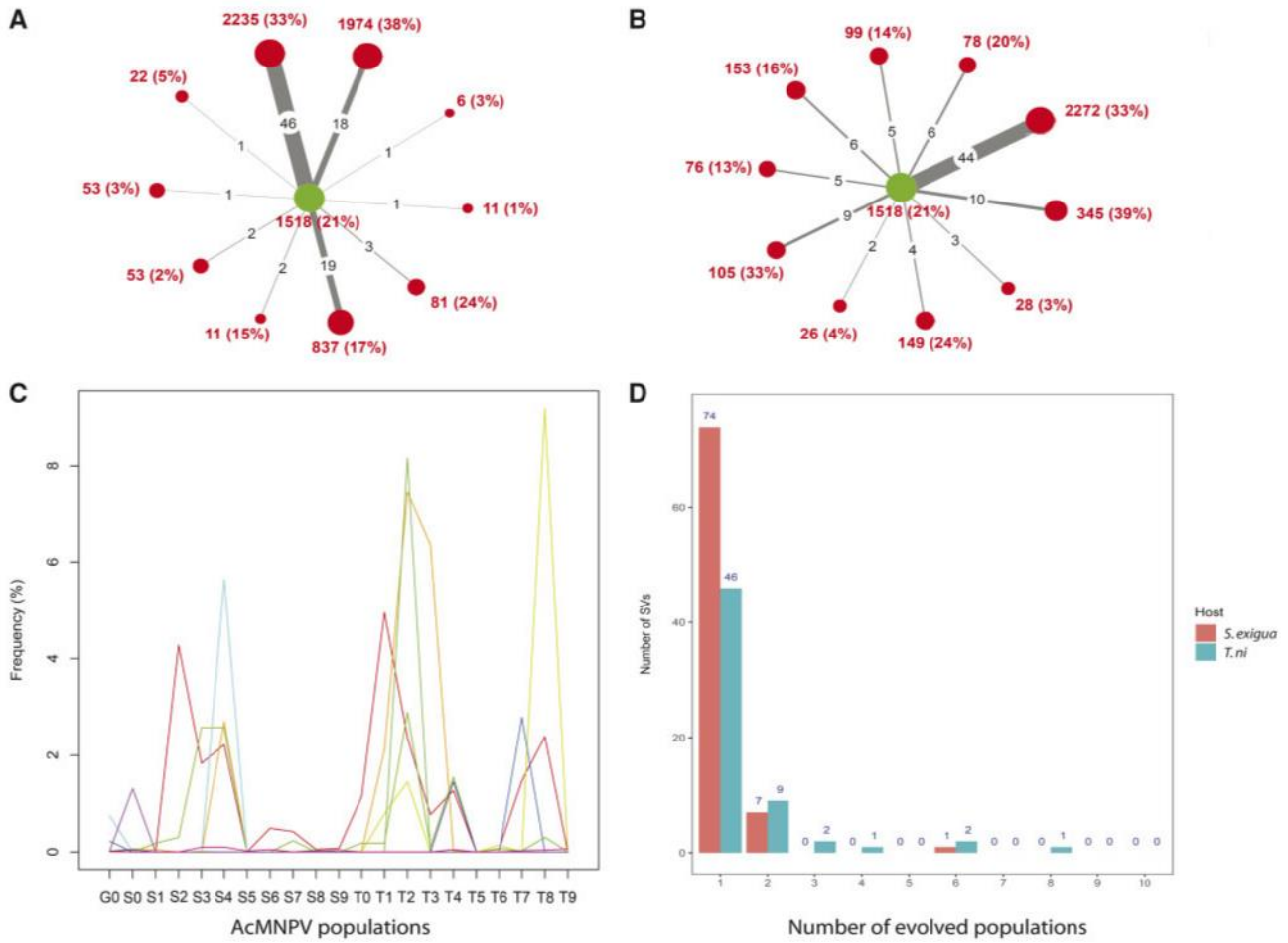



Figure 3.5. Dynamics of SVs in twenty evolved AcMNPV lines. (A) Red circles show the number of SVs detected in ten AcMNPV populations which were each purified after ten infection cycles on larvae of the beet armyworm (S. exigua). The green circle shows the number of SVs detected in the parental population of AcMNPV purified from larvae of the cabbage looper moth (T. ni). The size of the circle is proportional to the number of SVs and the frequency of viral genomes carrying a SV is given between brackets, assuming the number of SVs per viral genome follows a Poisson distribution. The thickness of the lines linking the parental AcMNPV population to each of the ten evolved populations is proportional to the number of shared SVs (numbers in black on the lines). (B) Same as in A except that the ten evolved AcMNPV popula-tions were purified after ten infection cycles on larvae of the same species (T. ni) as that used to generate the parental AcMNPV population. (C) Frequency of the ten most frequent SVs detected among the twenty-one viral populations and which were initially present in the parental AcMNPV population. One color corresponds to one SV. The 'G0' population refers to the parental population. The S0–S9 populations refer to the populations evolved on S. exigua larvae. The T0–T9 populations refer to populations evolved on T. ni. Note that no SVs reached >10 per cent in frequency in any AcMNPV population. (D) Number of SVs present in the parental AcMNPV population that were also detected in one to ten evolved viral populations. Most SVs were only detected in one evolved population (seventy-four in S. exigua and forty-six in T. ni).

among which thirty-four insertions and twenty-two deletions affected open reading frames (without disruption).

The four variant callers used to identify AcMNPV SVs in short reads were run on the three additional viruses. As previously, we conservatively estimated the number of robust SVs as 4.98 per cent of the total number of SVs identified with the four variant callers. Following this approach, we counted a total of 684,426 and 1,102 SVs in HCMV, IIV31, and IIV6 datasets, respectively (Table 3.1 and Supplementary Tables S3.4–S3.6). We estimated that overall 54.4, 80.1, and 52.4 per cent of the HCMV, IIV31, and IIV6 viral genomes, respectively, were affected by SVs, assuming the number of SVs in viral genomes follows a Poisson distribution (Fig. 3.2). The two most abundant SV types affecting IIV31 and IIV6 genomes were deletions and inversions, while duplications and inversions were the main events occurring in the HCMV genomes (Fig. 3.2A). It is noteworthy that about 76.6 and 47.2 per cent of IIV31 and IIV6 genomes were affected by deletions, respectively. Furthermore, >42 per cent of HCMV genomes carried an inversion (Fig. 3.2C). SV sizes were very heterogeneous, ranging from 52 bp for a deletion in IIV6 genomes to 209,948 bp for a deletion in IIV31 genomes. The average mean size of SVs was 10,800, 47,046, and 41,042 bp in HCMV, IIV31, and IIV6, respectively (Fig. 3.2B). For all three viruses, SVs were detected all along the genome, suggesting no region was devoid of SVs (Supplementary Fig. S3.15). The five most frequent SVs in IIV6, IIV31, and HCMV populations mainly involve intergenic regions, non-essential or uncharacterized genes (Supplementary Table S3.2). Strikingly, the five most frequent SVs in the IIV6 population involve the 444 gene (unknown function) and account for >25 per cent of the SV frequency in the viral population.

### 3.6 TE insertions in viral genomes

The seventy-four insertions detected in the AcMNPV population isolated from *S. exigua* and sequenced using both short- and long-read technologies all corresponded to insertions of AcMNPV sequences. These

Table 3.2. Characteristics of TE insertions detected in AcMNPV short-read data.

| TE superfamily | Number of unique chimeric reads | Number of chimeric reads including PCR duplicates | Insertion frequency (%) |
|---|---|---|---|
| Sola | 2,347 | 2,956 | 0.7 |
| Harbinger | 1,154 | 1,305 | 0.3 |
| Gypsy | 495 | 562 | 0.13 |
| Mariner | 361 | 1,078 | 0.24 |
| piggyBac | 349 | 398 | 0.09 |
| Copia | 180 | 319 | 0.07 |
| Transib | 51 | 57 | 0.01 |
| MuLE | 33 | 35 | 0.007 |
| Helitron | 22 | 22 | 0.005 |
| hAT | 1 | 10 | 0.002 |
| Total | 4,993 | 6,742 | 1.554 |

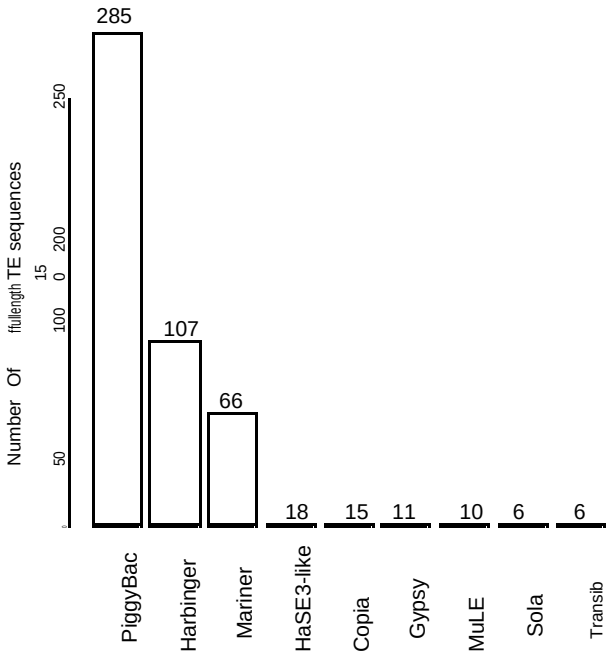



Figure 3.6. Number of TEs integrated as full-length copies for the nine TE superfamilies found in the AcMNPV genomes. Six and three TE superfamilies belong to Classes II and I TEs, respectively. The major part of full-length inserted TE sequences belong to the Class II TE superfamilies (480 complete TE sequences out of 524), mainly to the PiggyBac TE superfamily.

insertions could be considered duplications, but they were not classified as such by SV callers, presumably because duplicated sequences were not in tandem or located sufficiently close to each other along the AcMNPV genome. SV callers did not identify any insertion of non-AcMNPV DNA, which is somewhat surprising because our earlier works, based on twenty-one AcMNPV populations reanalyzed here, have shown that a large number of host TEs systematically integrate into AcMNPV genomes during infection (Gilbert et al. 2016). Using the same method as in Gilbert et al. (2016), we searched for host TEs in the short reads of the new AcMNPV population sequenced in this study. We identified 4,993 virus–host TE junctions involving one and nine superfamilies of Classes 1 and 2 TEs, respectively, and yielding an estimate of 1.5 per cent viral genomes harboring a host TE in this population (Table 3.2). Using the long-read dataset, we were further able to retrieve a total of 524 full-length TE copies from three Class 1 and six Class 2 TE superfamilies (Fig. 3.6). Another 1,233 TEs were identified in long reads as truncated copies. In contrast, no TE insertion was found using the Gilbert et al. (2016) pipeline in the HCMV, IIV6, and IIV31 genome populations.

## 4. Discussion

Although SVs have been implicated as an important source of viral evolution in several large dsDNA viruses (Lopez-Ferber et al. 2003; Elde et al. 2012; Mahiet et al. 2012; Chateigner et al. 2015; Filée, 2015; Karamitros et al. 2018; Sasani et al. 2018), the full spectrum and overall frequency of SVs carried by populations of these viruses has never been probed using high-throughput sequencing. Here, we begin tackling this issue by focusing on three invertebrate viruses for which obtaining large quantities of DNA from in vivo infections was possible (AcMNPV, IIV6, and IIV31) as well as on one human virus replicated in cell lines (HCMV). The rationale followed in this study to robustly estimate a minimum number of SVs segregating in populations of these large dsDNA viruses is that given that short- and long-read sequencing technologies are not affected by the same biases inducing the formation of technical chimeras (Tallon et al. 2014; Tsai et al., 2014; Griffith et al. 2018; Peccoud et al. 2018), SVs detected by both types of data can be considered biological. Although likely conservative, this approach revealed that populations of AcMNPV can carry more than one thousand SVs together affecting almost 40 per cent of genomes. Based on the proportion of AcMNPV SVs detected by both sequencing technologies among all SVs detected by Illumina sequencing only (4.98%) we have estimated that much like in AcMNPV, several hundreds to more than 1,000 SVs can be found in populations of three other large dsDNA viruses (HCMV, IIV6, and IIV31). The number of SVs found in these dsDNA viruses is thus similar to those circulating in RNA virus populations (Jaworski and Routh 2017). Our results are in agreement with earlier molecular biology studies (Mahiet et al. 2012) and further contribute to unveil SVs as an important facet of the biology of large dsDNA viruses.

The large number of programs that have been developed to detect SVs in short- and long-sequencing reads all rely on different approaches involving different algorithms and/or are based on different mapping strategies. Integrating the results of these various programs to detect as many robust SVs as possible was a challenge. While some meta-callers pooling the results of several SV callers exist (Wong et al. 2010; Mohiyuddin et al. 2015; Zarate et al. 2018), they were all geared to SVs detection in gigabase-sized genomes sequenced at <100 X, conditions that vastly differ from our study of kilobase-sized genomes sequenced at depth ranging from >3,000 to >200,000 X. We thus developed our own meta-calling approach based on hierarchical clustering of SVs detected by only some of the programs that are available. Our choice of programs was guided by limitations of some of these programs to effectively analyze our data. For example, Manta (Chen et al. 2016), Delly (Rausch et al. 2012), and Wham (Kronenberg et al. 2015) were unable to detect more than thirty SVs in the AcMNPV Illumina datasets, likely because they were benchmarked on data corresponding to <50 X sequencing depth or they automated the exclusion of deeply covered regions (Kronenberg et al. 2015; Chen et al. 2016). On a related methodological note, we further monitored the

influence sequencing depth has on the detection of SVs by subsampling our >200,000 X AcMNPV Illumina dataset at depths ranging from 50 X to 50,000 X and running our SV detection pipeline on these subsamples. We found that sequencing depth had a strong impact on the number of detected SVs, with <150 SVs detected at depths 50,000 X compared with the 1,141 SVs detected at >200,000 X (Supplementary Fig. S3.17). Thus, a large fraction of low frequency variants segregating in viral populations cannot be detected with our approach unless extremely high-sequencing depths are generated.

SVs have never been characterized in IIV6 and IIV31 so we cannot compare the nature of the SVs detected in this study to previous studies. In regard to HCMV, Karamitros et al. (2018) performed long-read sequencing of the TB40/E strain, which enabled precise characterization of a 1,348-bp deletion located be-tween genes UL144 and UL145. This SV was not identified in our HCMV short-read data; instead we found a deletion of a 350-bp intergenic region located between these same two genes with a frequency of 0.09 per cent. The two SVs are different, but they show that this genomic region is susceptible to deletions in both strains. Regarding AcMNPV, Chateigner et al. (2015) characterized short deletions in a population of this virus using short-read sequencing. The majority of deletions were found to involve hr1, hr2, hr3, and hr4b, which are homologous regions scattered around the AcMNPV genome thought to serve as origin of replication (Pearson et al. 1992; Kool et al. 1995). Here, we found 114 AcMNPV SVs involving hr regions in the population sequenced by both Illumina and PacBio technologies. We calculated that seven per cent of AcMNPV genomes harbor one SV involving an hr. These SVs correspond to 103 deletions, six inversions, three duplications, and two insertions. The hr regions most frequently involved in SVs are hr2, hr3, hr4b, and hr5, which is concordant with Chateigner et al. (2015) and further highlights the role of AcMNPV hr regions in producing SVs.

The fact that the majority of SVs are present at low to very low frequencies in all viral populations indicates they are likely deleterious and thus unlikely to persist over many rounds of replication. Accordingly, we found a higher frequency of non-inactivating versus inactivating SVs in most AcMNPV genes and very few SVs were shared between the parental *T. ni* AcMNPV population and the twenty populations of this virus that underwent ten infection cycles on *T. ni* or *S. exigua*. These findings echo the low number of TE insertions we found to be shared between the parental *T. ni* and the twenty evolved AcMNPV populations in our earlier study (Gilbert et al. 2016) and indicate that much like TE insertions, other SVs are continuously gained and lost at high rates during viral replication. This rapid SV turnover likely involve recombination (Crouch and Passarelli 2002; Kamita, Maeda, and Hammock 2003; Sijmons et al., 2015; Cudini et al. 2019) as well as errors in viral genome replication (Kilpatrick and Huang 1977; van Oers and Vlak 2007) and DNA repair (Kulkarni and Fortunato 2011; Xiaofei and Kowalik 2014; Renzette et al. 2015). It would be interesting to assess the relative importance of each of these mechanisms in generating structural diversity in the future.

It has been proposed that large SVs, including gene captures, gene duplications, and deletions play a crucial role in the adaptation of large dsDNA viruses to new hosts (Elde et al. 2012; Filée 2013; Thézé et al. 2015; Sasani et al. 2018). Our study was not designed to assess the adaptive role of SVs but it is noteworthy that the experiment producing the twenty-one AcMNPV short-read datasets involved a host switch from *T. ni* to *S. exigua* larvae. In this context, an independent increase in the frequency of a given SV in several AcMNPV lines replicated on *S. exigua* (coupled to no increase of the same SV in AcMNPV lines replicated on *T. ni*) would have provided an indication that this SV might have been involved in adaptation. Our analysis did not reveal any evidence of such a situation, nor did it reveal any case of polymorphic insertion involving a host gene. One reason for the absence of detectable adaptive AcMNPV SV after switching the virus from *T. ni* to *S. exigua* might be that in spite of diverging by more than 60 Myrs (Toussaint et al. 2012) the two noctuids used to generate those lines are too closely related to expect any major viral adaptation during a switch from one host to the other. Another possibility is that large SVs such as those on which we focused in this study are in fact rarely involved in viral adaptation because their effects on viral replication are too large. Interestingly, close inspection and comparison of the newly assembled consensus genome of the four viruses with the closest genomes available in GenBank revealed a number of differences involving small (<50 bp) variants. Since the GenBank viruses most closely related to AcMNPV, HCMV, and IIV6 were sequenced from different hosts (Sf9 cells for AcMNPV, $E_1$SM fibroblasts for HCMV, and CF-124 cells for IIV6) compared with this study, it is possible that the change in frequency of these small variants are due to their effect on the viral fitness in the different hosts. Interestingly, small variants are also found between our IIV31 genome and its closest relative in GenBank even though both viruses were isolated from *A. vulgare* (Piegu et al. 2014). These differences could be due to virus adaptation to different genetic backgrounds in pillbug populations or neutral genomic changes. Small variants may be more often involved in adaptation to host switches than larger ones because of their smaller effect on viral replication. Yet, it is also possible that several of these variants have no effect on viral fitness and became fixed through drift.

The short-read sequencing of a new AcMNPV population purified from *S. exigua* confirms our earlier observation of thousands of host TEs integrated in AcMNPV genomes (Gilbert et al. 2016). One limitation of short reads to analyze host TEs integrated into viral genomes is that it is impossible to assess whether reads mapping entirely on TEs originate from TE copies integrated into the virus genome or from copies integrated into contaminating fragments of the host genome. Thus, the completeness of TE copies integrated into AcMNPV genomes cannot be assessed using short reads. In agreement with previous low-throughput approaches (Fraser et al., 1995; Jehle et al. 1995, 1997), our long-read sequencing data shows that within an AcMNPV population, hundreds of TEs are integrated into AcMNPV genomes as full length copies. Although the high error rate of PacBio sequencing does not allow assessment of whether these TE copies are free from non-sense mutations, such high numbers of full length copies suggest that many of these TE are functional and potentially able to further jump from the viral genome into another genome, which may be that of another host infected by AcMNPV. Thus, our results further support the role of AcMNPV as potential vector of horizontal transfer of TEs between insects (Miller and Miller, 1982; Gilbert et al. 2014, 2016; Gilbert and Cordaux 2017).

The finding of many TEs integrated into AcMNPV genomes contrasts with the absence of TEs in all consensus baculovirus genomes sequenced so far, which suggests that TE insertions never reach high frequencies in viral populations (Gilbert et al. 2016). Thus, though the rate of TE transfer from host to virus is relatively high, the probability of TEs to have either a positive or no impacts on the virus fitness, and thus to increase in frequency in a viral population, is extremely low. The absence of polymorphic host gene insertions in AcMNPV populations

surprisingly contrasts with the relatively large number of host genes that have been captured by baculoviruses during their evolution (Hughes and Friedman 2003; Thézé et al. 2015). Thus, contrary to TEs, while host genes may rarely end up integrated into baculovirus genomes, their chances to improve viral fitness may be much higher than that of TEs.

The finding of many truncated TE copies in AcMNPV long reads is also interesting considering that the majority (895 out of 1,233) of these truncated copies begin or terminate the read in which they were found that is, they are flanked by viral sequences only at one of their ends. The high number of truncated TEs at the extremities of long reads does not correspond to what would be expected if truncated TEs were randomly dis-tributed in long reads (exact binomial test, P-value < $2.2*10^{-16}$, see Section 2). It is thus possible that at least a subset of truncated TE copies at read extremities correspond to the very extremity of linear AcMNPV genome molecules. In turn, such linear AcMNPV genomes could result from aborted transposition that led to the formation of truncated TE copies. Interestingly, linearized AcMNPV genomes are known to be 15- to 150-fold less infectious than circular forms (Kitts, Ayres, and Possee 1990). Thus, linearization of AcMNPV genomes mediated by aborted transposition could be viewed as beneficial by-product of transposition, which may impede or slow down AcMNPV replication. Here, the number of potentially linear AcMNPV genomes containing truncated TE copies is relatively low compared with TE-free genomes in the population we sequenced. Thus, the potential impact transposition-induced linearization may have on AcMNPV replication is unlikely to be significant in *S. exigua*. Yet, the possible antiviral protection conferred by aborted transposition of host TEs may be viewed as a form of cooperation between TEs and their hosts (Cosby, Chang, and Feschotte 2019) and worthy of further investigation in other host/baculovirus systems.

Finally, the absence of TE copies integrated in HCMV, IIV6, and IIV31 contrasts with their widespread occurrence in AcMNPV. It may be explained either by a low TE activity in human MRC5 cells, *S. nonagrioides* and *A. vulgare*, and/or by a weak capacity for the virus to carry supplementary genomic loads like TEs. This observation also contributes to make AcMNPV a better carrier of host TEs than other large dsDNA viruses, which, combined with its specificity for lepidopterans, may in part explain the higher number of horizontal transfer of TEs recently inferred in these compared with other arthropods (Reiss et al. 2019).

## Data availability

The various sequencing datasets produced during this study have been deposited in the SRA database of the NCBI under accession number PRJNA592818. All supplementary data, figures, tables and R scripts associated to this manuscript have been deposited in the DRYAD database (datadryad.org): DOI https://doi. org/10.5061/dryad.cfxpnvx25. This includes fasta files of the newly assembled AcMNPV, HCMV, IIV6 and IIV31 genomes, as well as their annotation files.

## Supplementary data

Supplementary data are available at Virus Evolution online.

Conflict of interest: None declared.

## Funding

This work was supported by Agence Nationale de la Recherche Grant ANR-15-CE32-0011-01 TransVir (to C.G.)

## References

Acevedo, A., and Andino, R. (2014) 'Library Preparation for Highly Accurate Population Sequencing of RNA Viruses', Nature Protocols, 9: 1760–9.

——, Brodsky, L., and Andino, R. (2014) 'Mutational and Fitness Landscapes of an RNA Virus Revealed through Population Sequencing', Nature, 505: 686–90.

Ackermann, H.-W., and Smirnoff, W. A. (1983) 'A Morphological Investigation of 23 Baculoviruses', Journal of Invertebrate Pathology, 41: 269–80.

Akhtar, L. N. et al. (2019) 'Genotypic and Phenotypic Diversity of Herpes Simplex Virus 2 within the Infected Neonatal Population', MSphere, 4: e00590–18

Alkan, C., Coe, B. P., and Eichler, E. E. (2011) 'Genome Structural Variation Discovery and Genotyping', Nature Reviews Genetics, 12: 363–76.

Angly, F. E. et al. (2012) 'Grinder: A Versatile Amplicon and Shotgun Sequence Simulator', Nucleic Acids Research, 40: e94.

Bankevich, A. et al. (2012) 'SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing', Journal of Computational Biology, 19: 455–77.

Bao, W., Kojima, K. K., and Kohany, O. (2015) 'Repbase Update, a Database of Repetitive Elements in Eukaryotic Genomes', Mobile DNA, 6:

Benson, D. A. et al. (2005) 'GenBank', Nucleic Acids Research, 33: D34–D38.

Bull, J. C., Godfray, H. C. J., and O'Reilly, D. R. (2003) 'A Few-Polyhedra Mutant and Wild-Type Nucleopolyhedrovirus Remain as a Stable Polymorphism during Serial Coinfection in Trichoplusia ni', Applied and Environmental Microbiology, 69: 2052–7.

Chaisson, M. J., and Tesler, G. (2012) 'Mapping Single Molecule Sequencing Reads Using Basic Local Alignment with Successive Refinement (BLASR): Application and Theory', BMC Bioinformatics, 13:238

Chateigner, A. et al. (2015) 'Ultra Deep Sequencing of a Baculovirus Population Reveals Widespread Genomic Variations', Viruses, 7: 3625–46.

Chebbi, M. A. et al. (2019) 'The Genome of Armadillidium vulgare (Crustacea, Isopoda) Provides Insights into Sex Chromosome Evolution in the Context of Cytoplasmic Sex Determination', Molecular Biology and Evolution, 36: 727–41.

Chen, X. et al. (2016) 'Manta: Rapid Detection of Structural Variants and Indels for Germline and Cancer Sequencing Applications', Bioinformatics, 32: 1220–2.

Cosby, R. L., Chang, N.-C., and Feschotte, C. (2019) 'Host–Transposon Interactions: Conflict, Cooperation, and Cooption', Genes & Development, 33: 1098–116.

Craig, N. L. (2002) 'Mobile DNA: An Introducton', in Craig, N. L., Craig, R., Gellert, M., and Lambowitz A. M. (eds.) Mobile DNA II, pp. 3–11. Washington DC: ASM Press.

Crouch, E. A., and Passarelli, A. L. (2002) 'Genetic Requirements for Homologous Recombination in Autographa californica Nucleopolyhedrovirus', Journal of Virology, 76: 9323–34.

Cudini, J. et al. (2019) 'Human Cytomegalovirus Haplotype Reconstruction Reveals High Diversity Due to Superinfection and Evidence of Within-Host Recombination', Proceedings of the

National Academy of Sciences of the United States of America, 116: 5693–8.

De Gooijer, C. D. et al. (1992) 'A Structured Dynamic Model for the Baculovirus Infection Process in Insect-Cell Reactor Configurations', Biotechnology and Bioengineering, 40: 537–48.

Duffy, S., Shackelton, L. A., and Holmes, E. C. (2008) 'Rates of Evolutionary Change in Viruses: Patterns and Determinants', Nature Reviews Genetics, 9: 267–76.

Elde, N. C. et al. (2012) 'Poxviruses Deploy Genomic Accordions to Adapt Rapidly against Host Antiviral Defenses', Cell, 150: 831–41.

English, A. C., Salerno, W. J., and Reid, J. G. (2014) 'PBHoney: Identifying Genomic Variants via Long-Read Discordance and Interrupted Mapping', BMC Bioinformatics, 15: 180.

File´ e, J. (2013) 'Route of NCLDV Evolution: The Genomic Accordion', Current Opinion in Virology, 3: 595–9.

——— (2015) 'Genomic Comparison of Closely Related Giant Viruses Supports an Accordion-like Model of Evolution', Frontiers in Microbiology, 6: 593.

Fraser, M. J. et al. (1995) 'Assay for Movement of Lepidopteran Transposon IFP2 in Insect Cells Using a Baculovirus Genome as a Target DNA', Virology, 211: 397–407.

Fukaya, M., and Nasu, S. (1966) 'A Chilo Iridescent Virus (CIV) from the Rice Stem Borer, Chilo Suppressalis WALKER (Lepidoptera : Pyralidae)', Applied Entomology and Zoology, 1: 69–72.

Gilbert, C., and Cordaux, R. (2017) 'Viruses as Vectors of Horizontal Transfer of Genetic Material in Eukaryotes', Current Opinion in Virology, 25: 16–22.

——— et al. (2014) 'Population Genomics Supports Baculoviruses as Vectors of Horizontal Transfer of Insect Transposons', Nature Communications, 5: 3348.

——— et al. (2016) 'Continuous Influx of Genetic Material from Host to Virus Populations', PLoS Genetics, 12: e1005838.

Godfray, H. C. J., Reilly, D. R. O., and Briggs, C. J. (1997) 'A Model of Nucleopolyhedrovirus (NPV) Population Genetics Applied to Co-occlusion and the Spread of the Few Polyhedra (FP) Phenotype', Proceedings of the Royal Society of London. Series B: Biological Sciences, 264: 315–22.

Go¨ rzer, I. et al. (2010) 'The Impact of PCR-Generated Recombination on Diversity Estimation of Mixed Viral Populations by Deep Sequencing', Journal of Virological Methods, 169: 248–52.

Griffith, P. et al. (2018). 'PacBio Library Preparation Using Blunt-End Adapter Ligation Produces Significant Artefactual Fusion DNA Sequences'. BioRxiv.

Hughes, A. L., and Friedman, R. (2003) 'Genome-Wide Survey for Genes Horizontally Transferred from Cellular Organisms to Baculoviruses', Molecular Biology and Evolution, 20: 979–87.

Jain, A. K. and Dubes R. C. (1988) Algorithms for Clustering Data, Prentice-Hall advanced reference series. Upper Saddle River, NJ: Prentice-Hall Inc.

Jaworski, E., and Routh, A. (2017) 'Parallel ClickSeq and Nanopore Sequencing Elucidates the Rapid Evolution of Defective-Interfering RNAs in Flock House Virus', PLoS Pathogens, 13: e1006365.

Jehle, J. A. et al. (1995) 'TCl4.7: A Novel Lepidopteran Transposon Found in Cydia Pomonella Granulosis Virus', Virology, 207: 369–79.

——— et al. (1997) 'Identification and Sequence Analysis of the Integration Site of Transposon TCp3.2 in the Gneome of Cydia Pomonella Granulovirus', Virus Research, 50: 151–7.

Kamita, S. G., Maeda, S., and Hammock, B. D. (2003) 'High-Frequency Homologous Recombination between Baculoviruses Involves DNA Replication', Journal of Virology, 77: 13053–61.

Karamitros, T. et al. (2016) 'De Novo Assembly of Human Herpes Virus Type 1 (HHV-1) Genome, Mining of Non-Canonical Structures and Detection of Novel Drug-Resistance Mutations Using Short- and Long-Read Next Generation Sequencing Technologies', PLoS One, 11: e0157600.

——— et al. (2018) 'Nanopore Sequencing and Full Genome de Novo Assembly of Human Cytomegalovirus TB40/E Reveals Clonal Diversity and Structural Variations', BMC Genomics, 19: 577.

Kilpatrick, B. A., and Huang, E.-S. (1977) 'Human Cytomegalovirus Genome: Partial Denaturation Map and Organization of Genome Sequences', Journal of Virology, 24: 16.

Kitts, P. A., Ayres, M. D., and Possee, R. D. (1990) 'Linearization of Baculovirus DNA Enhances the Recovery of Recombinant Virus Expression Vectors', Nucleic Acids Research, 18: 5667–72.

Kool, M. et al. (1991) 'Detection and Analysis of Autographa cali-fornica Nuclear Polyhedrosis Virus Mutants with Defective Interfering Properties', Virology, 183: 739–46.

——— et al. (1995) 'Replication of Baculovirus DNA', Journal of General Virology, 76: 2103–18.

Koren, S. et al. (2017) 'Canu: Scalable and Accurate Long-Read Assembly via Adaptive k-Mer Weighting and Repeat Separation', Genome Research, 27: 722–36.

Kronenberg, Z. N. et al. (2015) 'Wham: Identifying Structural Variants of Biological Consequence', PLoS Computational Biology, 11: e1004572.

Kulkarni, A. S., and Fortunato, E. A. (2011) 'Stimulation of Homology-Directed Repair at I-SceI-Induced DNA Breaks dur-ing the Permissive Life Cycle of Human Cytomegalovirus', Journal of Virology, 85: 6049–54.

Lauring, A. S., and Andino, R. (2010) 'Quasispecies Theory and the Behavior of RNA Viruses', PLoS Pathogens, 6: e1001005.

———, Frydman, J., and Andino, R. (2013) 'The Role of Mutational Robustness in RNA Virus Evolution', Nature Reviews Microbiology, 11: 327–36.

Layer, R. M. et al. (2014) 'LUMPY: A Probabilistic Framework for Structural Variant Discovery', Genome Biology, 15: R84.

Li, H. (2015) 'FermiKit: Assembly-Based Variant Calling for Illumina Resequencing Data', Bioinformatics, 31: 3694-3696.

———, and Durbin, R. (2009) 'Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform', Bioinformatics, 25: 1754–60.

Li, D. et al. (2011) 'Defective Interfering Viral Particles in Acute Dengue Infections', PLoS One, 6: e19447.

Lopez, C. B. (2014) 'Defective Viral Genomes: Critical Danger Signals of Viral Infections', Journal of Virology, 88: 8720–3.

Lo´ pez-Ferber, M. et al. (2003) 'Defective or Effective? Mutualistic Interactions between Virus Genotypes', Proceedings of the Royal Society of London. Series B: Biological Sciences, 270: 2249–55.

Lupetti, P. et al. (2013) 'Iridovirus Infection in Terrestrial Isopods from Sicily (Italy)', Tissue and Cell, 45: 321–7.

Maghodia, A. B., Jarvis, D. L., and Geisler, C. (2014) 'Complete Genome Sequence of the Autographa californica Multiple Nucleopolyhedrovirus Strain E2', Genome Announc, 2: e01202 14.

Mahiet, C. et al. (2012) 'Structural Variability of the Herpes Simplex Virus 1 Genome in Vitro and in Vivo', Journal of Virology, 86: 8592–601.

Manzoni, T. B., and Lo´ pez, C. B. (2018) 'Defective (Interfering) Viral Genomes Re-Explored: Impact on Antiviral Immunity and Virus Persistence', Future Virology, 13: 493–503.

Marriott, A. C., and Dimmock, N. J. (2010) 'Defective Interfering Viruses and Their Potential as Antiviral Agents', Reviews in Medical Virology, 20: 51–62.

Miller, D. W., and Miller, L. K. (1982) 'A Virus Mutant with an Insertion of a Copia-like Transposable Element', Nature, 299: 562–4.

Mohiyuddin, M. et al. (2015) 'MetaSV: An Accurate and Integrative Structural-Variant Caller for Next Generation Sequencing', Bioinformatics, 31: 2741–4.

Mönchgesang, S. et al. (2016) 'Natural Variation of Root Exudates in Arabidopsis thaliana-Linking Metabolomic and Genomic Data', Science Reports, 6: 1–11.

Mu¨ llner, D. (2013) 'Fastcluster: Fast Hierarchical, Agglomerative Clustering Routines for R and Python', Journal of Statistical Software, 53: 1–18.

O'Reilly, D. R., Miller, L. K., and Luckow, V. A. (1992). Baculovirus Expression Vectors, A Laboratory Manual. New York: W.H. Freeman and Co.

Ono, Y., Asai, K., and Hamada, M. (2013) 'PBSIM: PacBio Reads Simulator—Toward Accurate Genome Assembly', Bioinformatics, 29: 119–21.

Parikh, H. et al. (2016) 'Svclassify: A Method to Establish Benchmark Structural Variant Calls', BMC Genomics, 17:64.

Pearson, M. et al. (1992) 'The Autographa californica Baculovirus Genome: Evidence for Multiple Replication Origins', Science, 257: 1382–4.

Peccoud, J. et al. (2017) 'Massive Horizontal Transfer of Transposable Elements in Insects', Proceedings of the National Academy of Sciences of the United States of America, 114: 4721–6.

—— et al. (2018) 'A Survey of Virus Recombination Uncovers Canonical Features of Artificial Chimeras Generated during Deep Sequencing Library Preparation', G3 Genes Genomes Genetics, 8: 1129–38.

Piegu, B. et al. (2014) 'Genome Sequence of a Crustacean Iridovirus, IIV31, Isolated from the Pill Bug, Armadillidium vul-gare', Journal of General Virology, 95: 1585–90.

Poirier, E. Z. et al. (2018) 'Dicer-2-Dependent Generation of Viral DNA from Defective Genomes of RNA Viruses Modulates Antiviral Immunity in Insects', Cell Host & Microbe, 23: 353–65.e8.

Quinlan, A. R., and Hall, I. M. (2010) 'BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features', Bioinformatics, 26: 841–2.

Rausch, T. et al. (2012) 'DELLY: Structural Variant Discovery by Integrated Paired-End and Split-Read Analysis', Bioinformatics, 28: i333–i339.

R Core Team. (2018) R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing.

Reiss, D. et al. (2019) 'Global Survey of Mobile DNA Horizontal Transfer in Arthropods Reveals Lepidoptera as a Prime Hotspot', PLoS Genetics, 15: e1007965.

Renzette, N. et al. (2013) 'Rapid Intrahost Evolution of Human Cytomegalovirus is Shaped by Demography and Positive Selection', PLoS Genetics, 9: e1003735.

—— et al. (2015) 'Limits and Patterns of Cytomegalovirus Genomic Diversity in Humans', Proceedings of the National Academy of Sciences of the United States of America, 112: E4120–8.

—— et al. (2017) 'On the Analysis of Intrahost and Interhost Viral Populations: Human Cytomegalovirus as a Case Study of Pitfalls and Expectations', Journal of Virology, 91: e01976–16.

Rohrmann, G. F. (2014) 'Baculovirus Nucleocapsid Aggregation (MNPV vs SNPV): an Evolutionary Strategy, or a Product of Replication Conditions? ', Virus Genes, 49: 351–7.

Routh, A. et al. (2015) 'ClickSeq: Fragmentation-Free Next-Generation Sequencing via Click Ligation of Adaptors to Stochastically Terminated $3^0$-Azido cDNAs', Journal of Molecular Biology, 427: 2610–6.

Sanjuan, R., and Domingo-Calap, P. (2016) 'Mechanisms of Viral Mutation', Cellular and Molecular Life Sciences, 73: 4433–48.

Sasani, T. A. et al. (2018) 'Long Read Sequencing Reveals Poxvirus Evolution through Rapid Homogenization of Gene Arrays', Elife, 7: e35453.

Sedlazeck, F. J. et al. (2018) 'Accurate Detection of Complex Structural Variations Using Single-Molecule Sequencing', Nature Methods , 15: 461–8.

Sijmons, S. et al. (2015) 'High-Throughput Analysis of Human Cytomegalovirus Genome Diversity Highlights the Widespread Occurrence of Gene-Disrupting Mutations and Pervasive Recombination', Journal of Virology, 89: 7673–95.

Simon, O. et al. (2006) 'Dynamics of Deletion Genotypes in an Experimental Insect Virus Population', Proceedings of the Royal Society B: Biological Sciences, 273: 783–90.

Slack, J., and Arif, B. M. (2006). 'The Baculoviruses Occlusion-Derived Virus: Virion Structure and Function', Advances in Virus Research, 69: 99–165.

Sommer, D. D. et al. (2007) 'Minimus: A Fast, Lightweight Genome Assembler', BMC Bioinformatics, 8: 64.

Sparks, W., Li, H., and Bonning, B. (2008). 'Protocols for Oral Infection of Lepidopteran Larvae with Baculovirus', Journal of Visualized Experiments. 19: 888.

Stocker, B. K., Ko¨ ster, J., and Rahmann, S. (2016) 'SimLoRD: Simulation of Long Read Data', Bioinformatics, 32: 2704–6.

Szpara, M. L. et al. (2014) 'Evolution and Diversity in Human Herpes Simplex Virus Genomes', Journal of Virology, 88: 1209–27.

Tallon, L. J. et al. (2014) 'Single Molecule Sequencing and Genome Assembly of a Clinical Specimen of Loa loa, the Causative Agent of Loaiasis', BMC Genomics, 15: 788.

Tcherepanov, V., Ehlers, A., and Upton, C. (2006) 'Genome Annotation Transfer Utility (GATU): Rapid Annotation of Viral Genomes Using a Closely Related Reference Genome', BMC Genomics, 7:150.

Thézé, J. et al. (2015) 'Gene Acquisition Convergence between Entomopoxviruses and Baculoviruses', Viruses, 7: 1960–74.

Toussaint, E. F. A. et al. (2012) 'Palaeoenvironmental Shifts Drove the Adaptive Radiation of a Noctuid Stemborer Tribe (Lepidoptera, Noctuidae, Apameini) in the Miocene', PLoS One, 7: e41377.

Treangen, T. J. et al. (2011) 'Next Generation Sequence Assembly with AMOS', Current Protocols in Bioinformatics, 33: 11.8.

Tsai, I. J. et al. (2014) 'Summarizing Specific Profiles in Illumina Sequencing from Whole-Genome Amplified DNA', DNA Research, 21: 243–54.

van Oers, M., and Vlak, J. (2007) 'Baculovirus Genomics', Current Drug Targets, 8: 1051–68.

Vasilijevic, J. et al. (2017) 'Reduced Accumulation of Defective Viral Genomes Contributes to Severe Outcome in Influenza Virus Infected Patients', PLoS Pathogens, 13: e1006650.

Walker, B. J. et al. (2014) 'Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement', PLoS One, 9: e112963.

Ward, J. H. (1963) 'Hierarchical Grouping to Optimize an Objective Function', Journal of the American Statistical Association, 58: 236–44.

Wei, Z.-G., and Zhang, S.-W. (2018) 'NPBSS: A New PacBio Sequencing Simulator for Generating the Continuous Long Reads with an Empirical Model', BMC Bioinformatics, 19:177.

Wong, K. et al. (2010) 'Enhanced Structural Variant and Breakpoint Detection Using SVMerge by Integration of

Multiple Detection Methods and Local Assembly', Genome Biology, 11: R128.

Xiaofei, E., and Kowalik, T. (2014) 'The DNA Damage Response Induced by Infection with Human Cytomegalovirus and Other Viruses', Viruses, 6: 2155–85.

Ye, K. et al. (2009) 'Pindel: A Pattern Growth Approach to Detect Break Points of Large Deletions and Medium Sized Insertions from Paired-End Short Reads', Bioinformatics, 25: 2865–71.

Zarate, S. et al. (2018). 'Parliament2: Fast Structural Variant Calling Using Optimized Combinations of Callers', BioRxiv.

Zhang, W., Jia, B., and Wei, C. (2019) 'PaSS: A Sequencing Simulator for PacBio Sequencing', BMC Bioinformatics, 20:352.

Zhang, Y. et al. (2018) 'Large-Scale Comparative Epigenomics Reveals Hierarchical Regulation of non-CG Methylation in Arabidopsis', Proceedings of the National Academy of Sciences of the United States of America, 115: E1069–E1074.

Zwart, M. P., Tromas, N., and Elena, S. F. (2013) 'Model-Selection-Based Approach for Calculating Cellular Multiplicity of Infection during Virus Colonization of Multi-cellular Hosts', PLoS One, 8: e64657.

## Supplementary data

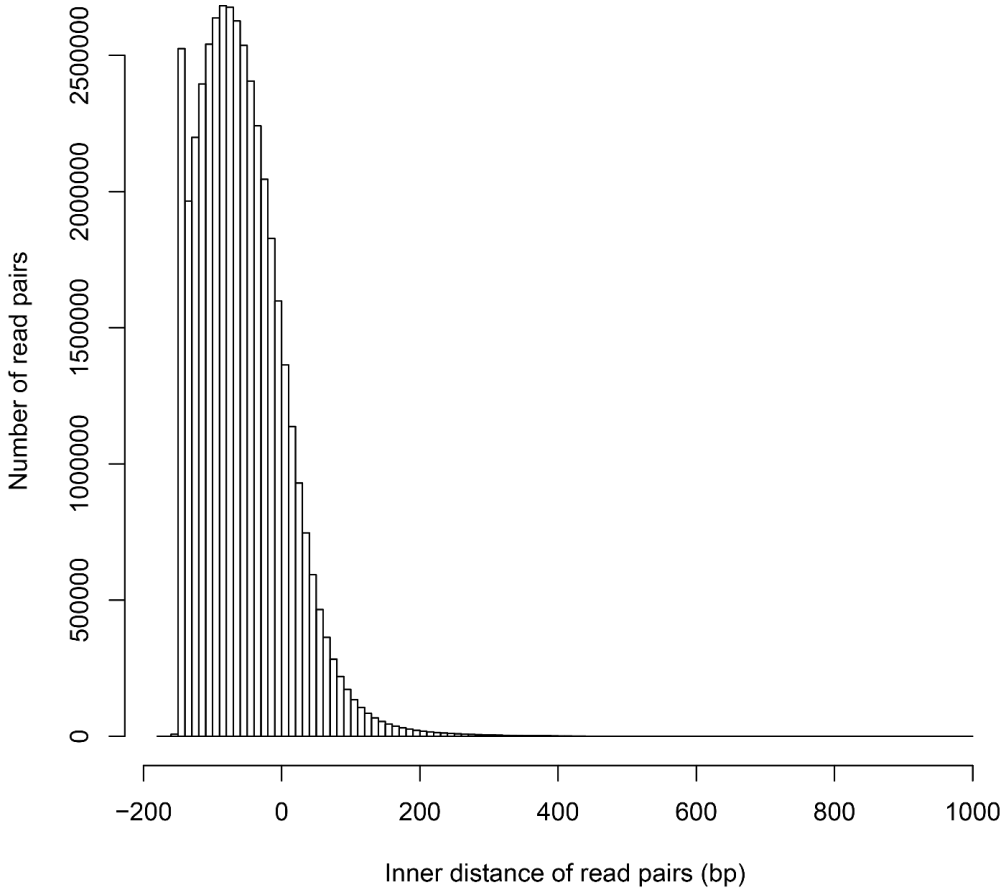



**Figure S3.1: Distribution of inner distance of all Illumina read pairs aligning on the AcMNPV genome.** Most of the pairs have an inner distance <700bp, the threshold used to consider the distance as the result of a large deletion. The inner distance can be negative when both reads of a pair overlap.

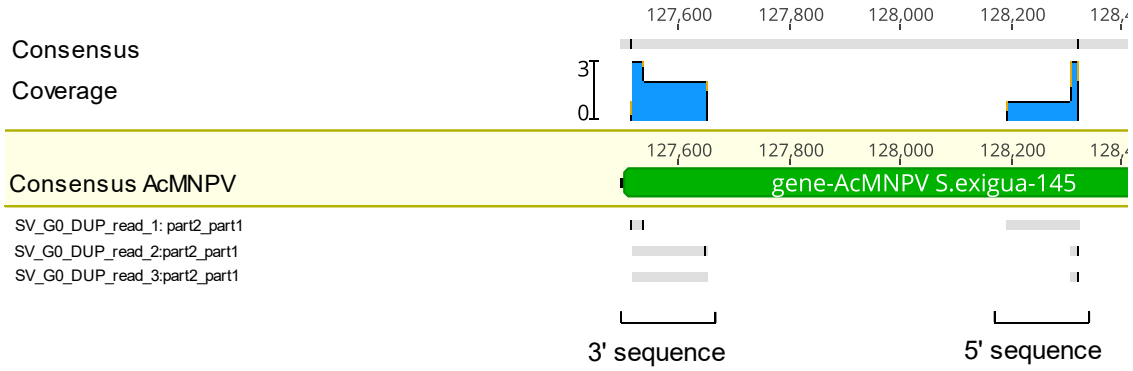



**Figure S3.2: Duplication detected in three reads in the AcMNPV G0 data.** The reads are split in a way supporting a duplication event. This duplication was detected in only three reads, supporting SV detection by three reads as true events and emphasizing the precision of our frequency calculation.


**Figure S3.3: The most frequent 'NA_NA' deletion detected in short reads from the AcMNPV data.** The short deletion is visible into each read with a precise alignment. Only a sample of all reads supporting the deletion is visible here.

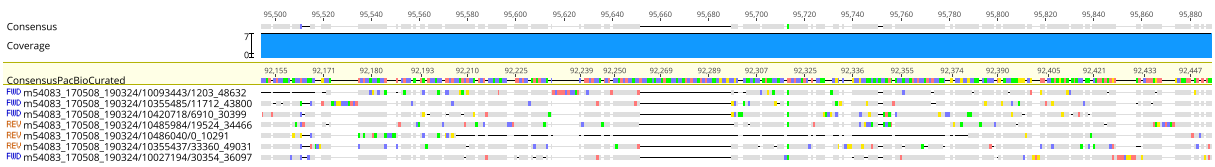



**Figure S3.4: The most frequent 'NA_NA' deletion detected in long reads from the AcMNPV data.** Only some reads supporting the deletion were retrieved here. As expected, the high error rate of long-read sequencing hinders the detection of SV breakpoint at a nucleotide

resolution. Regarding the fifth read with a larger deletion in the alignment, our clustering step can gather reads with a deletion of different sizes, likely two different deletion events. That is why our approach is conservative in the number of detected SVs, likely greater than the ones we detected.

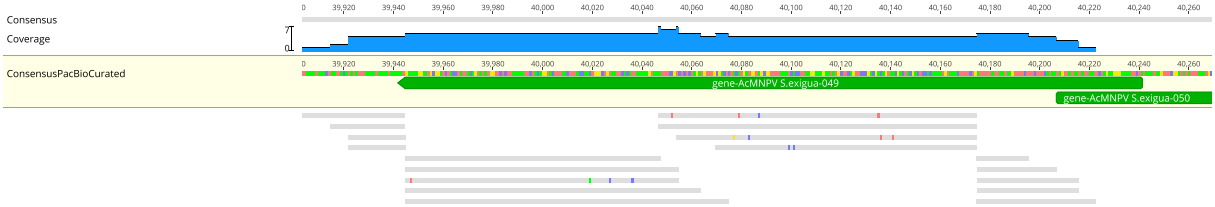



**Figure S3.5: Inversion detected in short reads from the AcMNPV data.** Short reads can also detect inversion with a high precision. This inversion is supported by nine reads and occurred into a viral gene.

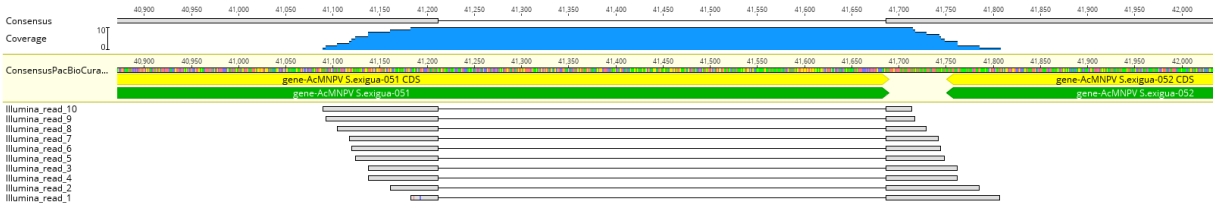



**Figure S3.6: Deletion detected in short reads from the AcMNPV data.** This deletion is supported by ten reads and occurred into a viral gene.

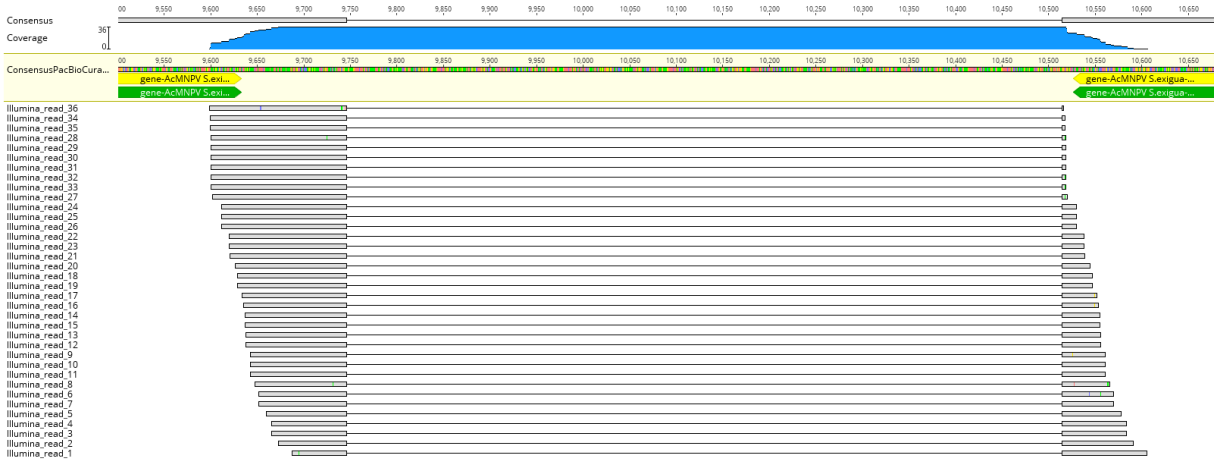



**Figure S3.7: Deletion detected in short reads from the AcMNPV data.** This deletion is supported by 36 reads and occurred into an intergenic region.

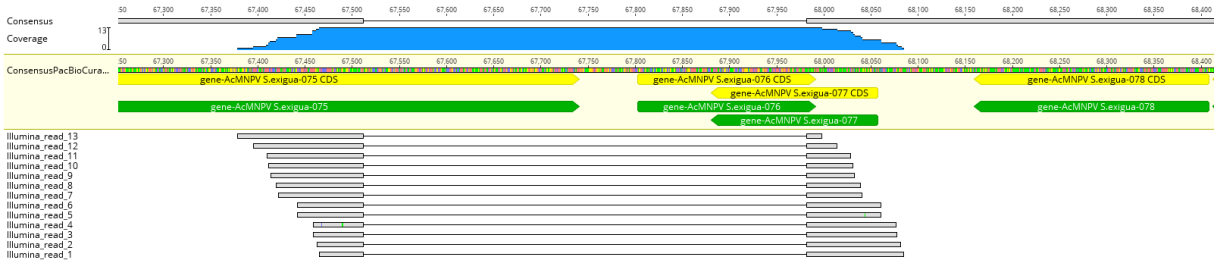



**Figure S3.8: Deletion detected in short reads from the AcMNPV data.** This deletion is supported by 13 reads and involved two genes.

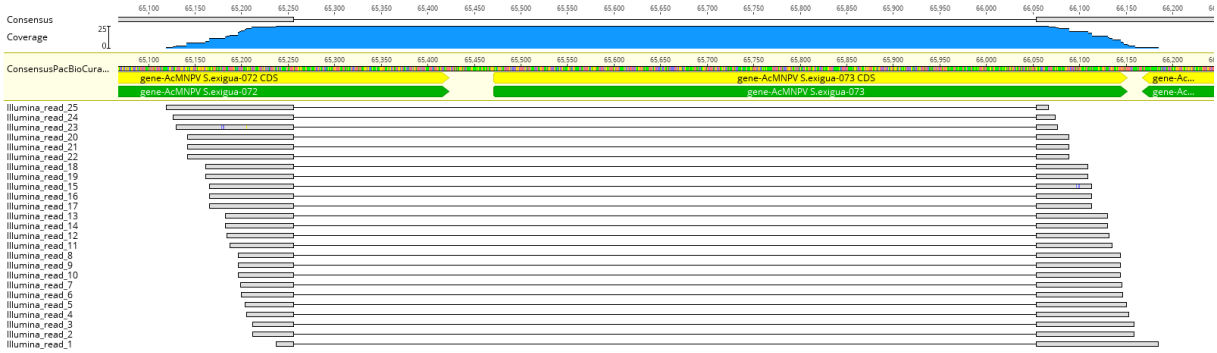



**Figure S3.9: Deletion detected in short reads from the AcMNPV data.** This deletion is supported by 25 reads and involved two genes.

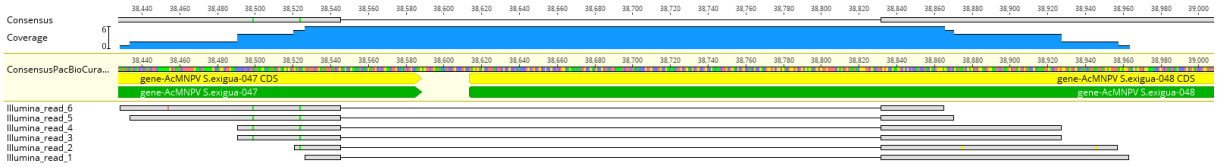



**Figure S3.10: Deletion detected in short reads from the AcMNPV data.** This deletion is supported by six reads and involved two genes.

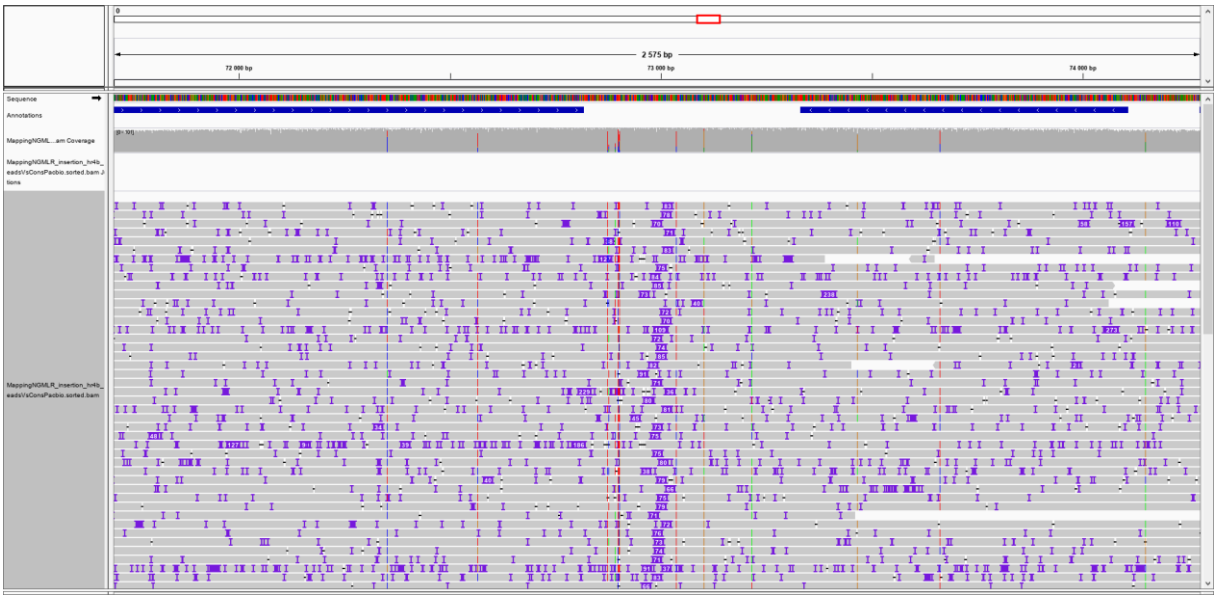


**Figure S3.11: Insertion detected in long reads from the AcMNPV data.** Long reads can also detect insertion. This insertion is supported by many reads and occurred into the hr4b region. The read alignment is visualized with IGV. Insertions are represented as purple rectangles indicating the number of inserted base pairs. The inserted length corresponds to about 70 bp.

**Figure S3.12: Deletion detected in long reads from the AcMNPV data.** This deletion occurs between an intergenic region and the PNK|PNL gene. The read alignment is visualized with IGV. The numbers indicate the deletion length, about 1,347 bp.

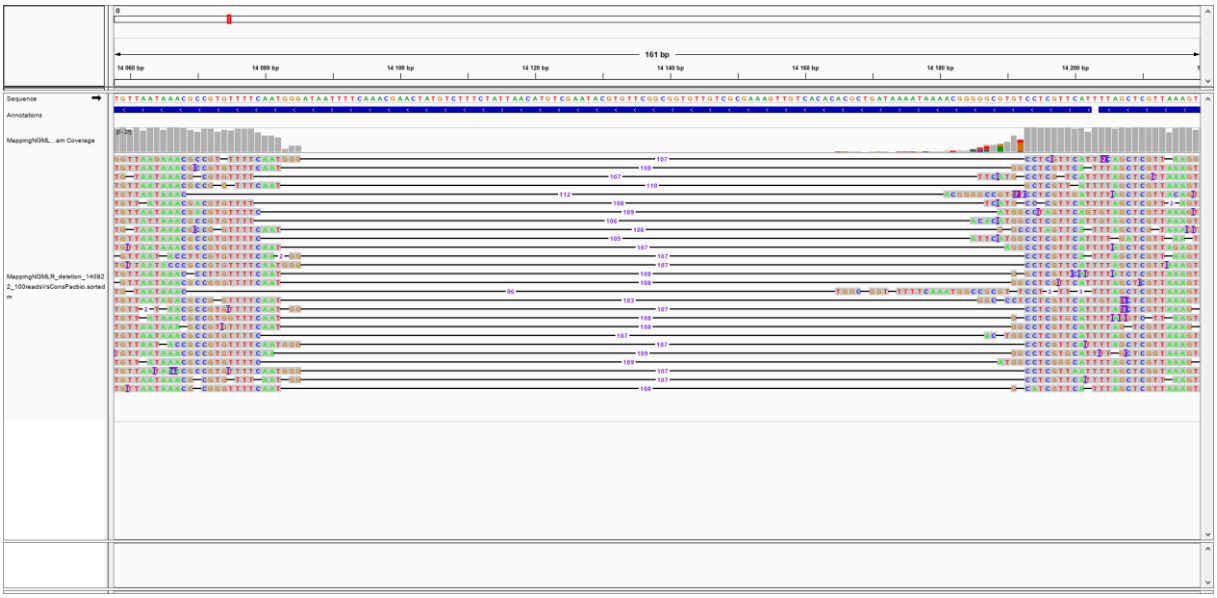


**Figure S3.13: Deletion detected in long reads from the AcMNPV data.** This deletion occurs in the Ac-IAP1 gene. The read alignment is visualized with IGV. The numbers indicate the deletion length, about 108 bp.

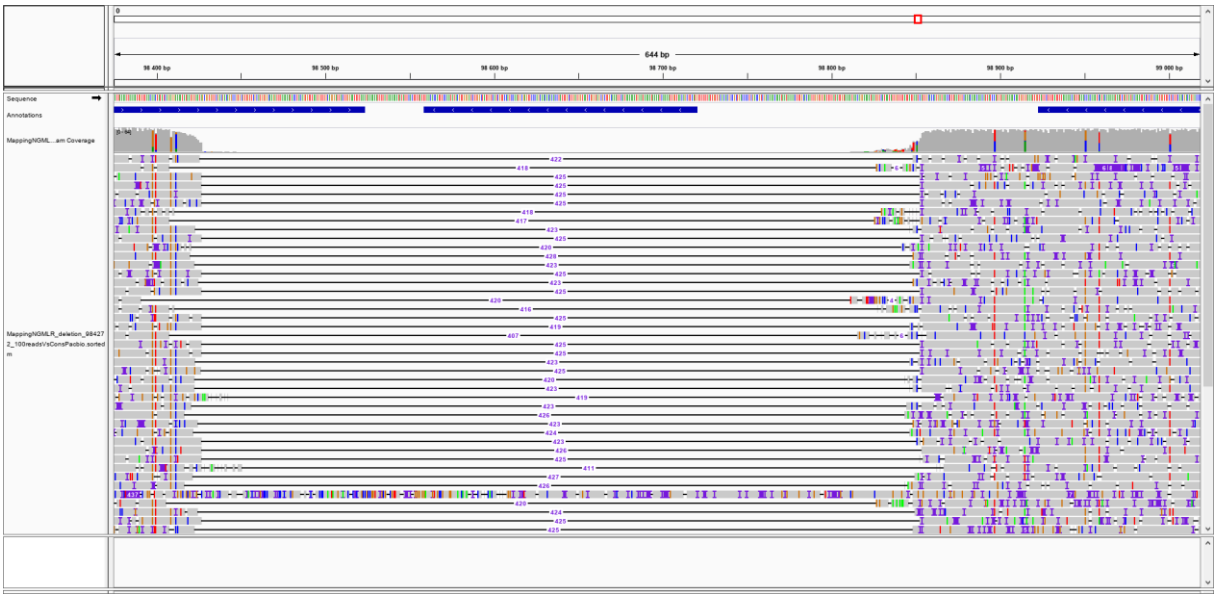


**Figure S3.14: Deletion detected in long reads from the AcMNPV data.** This deletion begins into the Ac-PNK|PNL gene and encompasses the AcOrf-85 gene.The read alignment is visualized with IGV. The numbers indicate the deletion length, about 425 bp.

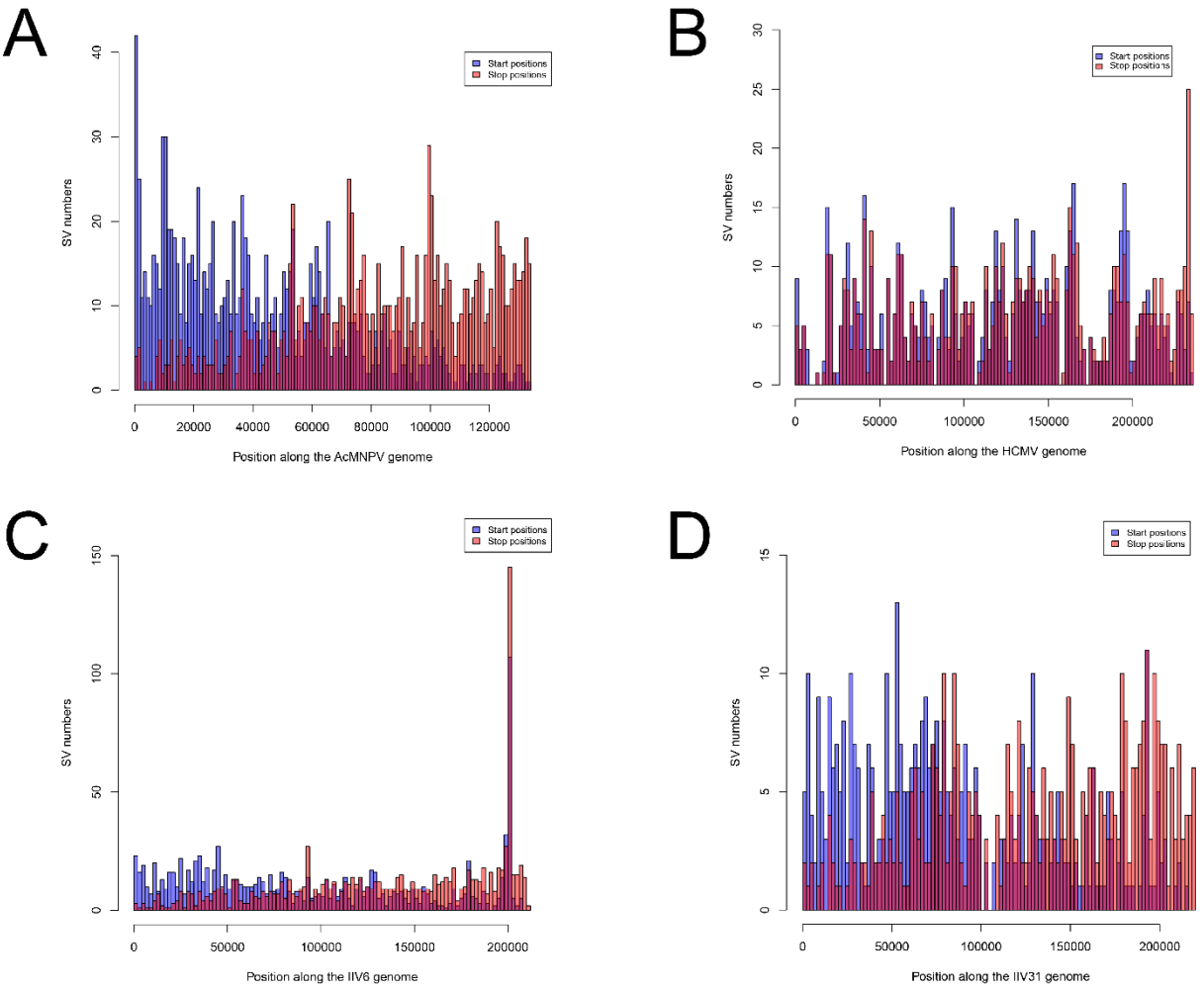


**Figure S3.15: SV breakpoint positions along the viral genomes.** Breakpoints are found all along the viral genomes. A: Breakpoint positions of detected SVs along the AcMNPV genome. B: Breakpoint positions of detected SVs along the HCMV genome. C: Breakpoint positions of detected SVs along the IIV6 genome. D: Breakpoint positions of detected SVs along the IIV31 genome.

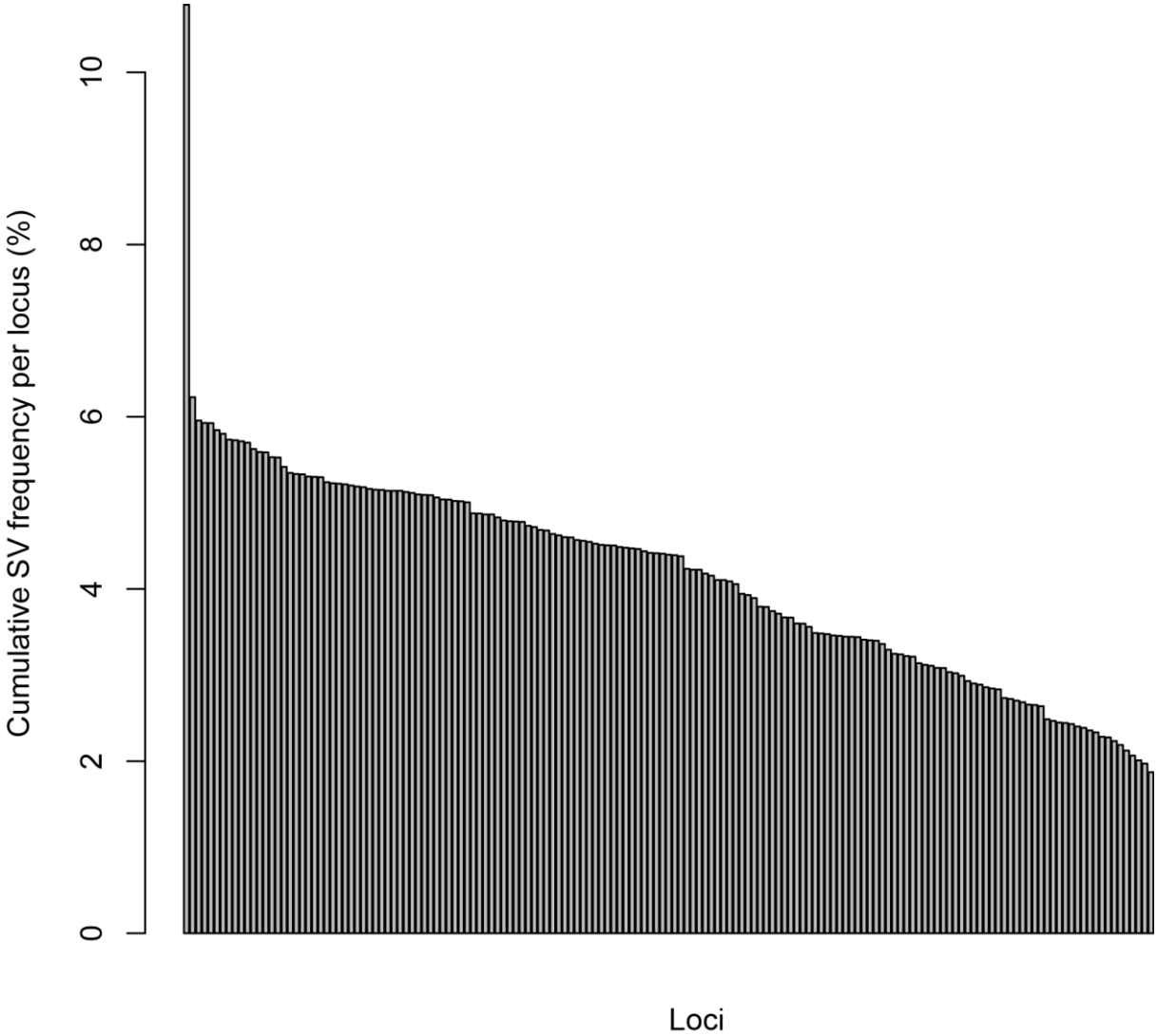



**Figure S3.16: Cumulative frequency of SVs affecting each locus of the AcMNPV genome.** The frequency was computed assuming the number of SVs per viral genome follow a Poisson distribution. Most of loci have an SV frequency ranging from 1.9% to 6.2%.

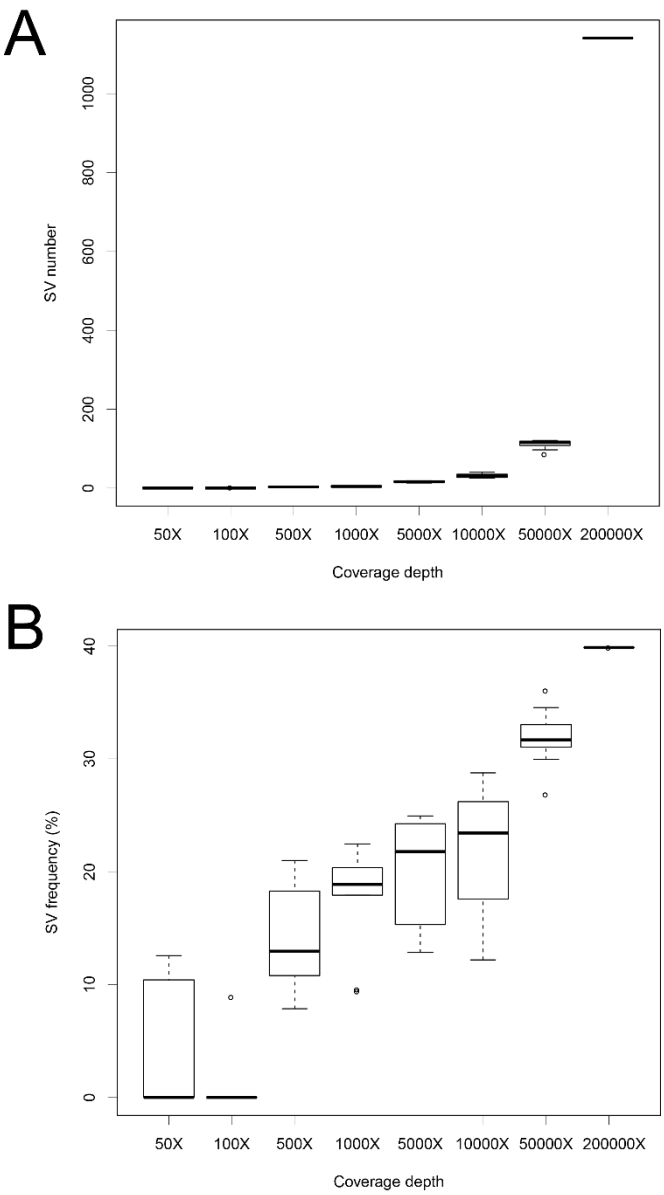



**Figure S3.17: Number and frequency of detected SVs according to the coverage depth. A:** SV number detected for each coverage depth. **B:** Cumulative mean frequency computed for each coverage depth. Frequencies were computed assuming the number of SVs per viral genome follows a Poisson distribution. On the two graphs, the SV number and frequency computed with the full data (200,000X) were added for comparison. The subsampling of AcMNPV long-read and short-read datasets was performed to get a number of reads corresponding to 50X, 100X, 500X, 1,000X, 5,000X, 10,000X and 50,000X coverage on the AcMNPV genome with ten replicates for each coverage depth. The SV detection was performed just as for the entire data (200,000X). Paired Wilcoxon tests showed SV numbers were always statistically different (excepted between 50X and 100X) between 100X and 500X; 500X and 1,000X; 1,000X and 5,000X; 5,000X and 10,000X and 10,000X and 50,000X (Wilcoxon tests, W=63, p-value=0.2113; W=1.5, p-value=0.0001218; W=23, p-value=0.03849; W=0, p-value=0.0001697; W=0, p-value=0.0001786; W=100, p-value=0.00001083; Kruskall-Wallis $X^2$=66.207, df=6, p-value=$2.444*10^{-12}$). Wilcoxon tests with SV frequency gave the same general

pattern , excepted SV frequency between 50X and 100X; 500X and 1,000X, 1,000X and 5,000X and 5,000X and 10,000X were not significantly different (Wilcoxon tests, W=67, p-value=0.1012; W=1, p-value=0.0001212; W=33, p-value=0.2176; W=34, p-value=0.2475; W=34, p-value=0.2475; W=2, p-value=0.0000433; Kruskall-Wallis $X^2$=55.53, df=6, p-value=$3.622*10^{-10}$). These results suggest the global number and frequency of SVs depend on coverage depth, due to a high number of SVs supported by only few reads, concordant with the very low SV frequency in the viral population. A coverage depth below 5,000X does not allow an accurate detection in terms of number and frequency of SVs, due to a too small set of viral genomic data. We confirmed there was a high homogeneity in SV number and frequency among replicates (SV number: Kruskall-Wallis $X^2$=0.39001, df=9, p-value=1; SV frequency: Kruskall-Wallis $X^2$=1.675, df=9, p-value=0.9956). It is worthy to note the detection of an insertion of ~70bp located in the hr2 region that was present in ~10% of viral genomes, depending the dataset. This was the most frequent SV detected in most of the datasets and in all the 10,000X and 50,000X datasET. This insertion was detected at all the coverage depth tested here. Strangely, it was not detected in our SV dataset with all the reads. To be sure the problem did not come from our clustering step, we searched for the presence of this insertion in the VCF output files from the SV callers that detected it, Pindel and Sniffles programs. We found the insertion was present in the Pindel output from the subsampled data and from the entire data (200,000X). However, the insertion was present in the Sniffles output from the subsampled data but not from the entire data. The large amount of data with all long or short reads could hamper the optimal use of SV callers and lead in some case to a lack of detection of some SVs.

Supplementary Tables S3.1, S3.4, S3.5 and S3.6 represent large tables of each AcMNPV, HCMV, IIV6 and IIV31 SV characteristics and are thus not shown here.

**Table S3.2: Viral regions affected by the five most frequent SVs in the AcMNPV, HCMV, IIV6 and IIV31 populations.** As expected, the majority of most frequent SV breakpoints involve intergenic regions, non-essential or uncharacterized genes, these regions encoding no essential proteins for the viral life cycles. The two genes encompassing most frequent SV breakpoints in IIV6 have not known functions, in part because this virus is less studied than the two others. The frequencies are computed considering the SV number per viral genome follows a Poisson distribution.

| SVs | regions | Frequency (%) | Location of the SV start coordinate | Location of the SV end coordinate |
|---|---|---|---|---|
| **AcMNPV** | **/** | **39.9** | **/** | |
| Insertion | hr4b | 6.03 | Homologous repeated region | / |
| Deletion | NA-NA | 3.65 | intergenic region | intergenic region |
| Deletion | NA-PNK\|PNL | 1.65 | intergenic region | Uncharacterized |
| Insertion | AcOrf-145 | 0.95 | Chitin binding | / |
| Duplication | AcOrf-145-AcOrf-145 | 0.94 | Chitin binding | Chitin binding |
| **HCMV** | **/** | **54.4** | **/** | |
| Inversion | NA-NA | 5.62 | intergenic region | intergenic region |
| Deletion | RL1-RL1 | 3.26 | Non-essential gene for viral growth | Non-essential gene for viral growth |
| Deletion | NA-NA | 3.11 | intergenic region | intergenic region |
| Deletion | NA-NA | 2.56 | intergenic region | intergenic region |
| Duplication | NA-NA | 2.51 | intergenic region | intergenic region |
| **IIV6** | **/** | **52.4** | **/** | |
| Deletion | 444-444 | 13.9 | Uncharacterized | Uncharacterized |
| Deletion | 444-444 | 12.4 | Uncharacterized | Uncharacterized |
| Deletion | 444-444 | 11.9 | Uncharacterized | Uncharacterized |
| Deletion | 444-444 | 3.9 | Uncharacterized | Uncharacterized |
| Deletion | 444-444 | 3.1 | Uncharacterized | Uncharacterized |
| **IIV31** | **/** | **80.1** | **/** | |
| Deletion | 34R-NA | 37.3 | Hypothetical protein | intergenic region |
| Deletion | NA-122R | 32.8 | intergenic region | Hypothetical protein |
| Deletion | 120R-120R | 24.8 | DNA-directed RNA polymerase subunit | DNA-directed RNA polymerase subunit |
| Deletion | NA-077R | 23.8 | intergenic region | Bro-like protein, GIY-YIG domain |
| Insertion | NA-NA | 6.6 | intergenic region | intergenic region |

**Table S3.3: Numbers and frequencies of AcMNPV SVs detected in real and simulated short-reads.** The frequencies were computed considering the number of SVs per viral genome follows a Poisson distribution.

| | Real data | Simulated data |
|---|---|---|
| | AcMNPV | AcMNPV |
| SV number | 1263 | 802 |
| Total frequency (%) | 20.61 | 1.47 |
| Deletion number | 308 | 737 |
| Duplication number | 149 | 65 |
| Insertion number | 0 | 0 |
| Inversion number | 806 | 0 |

# Chapitre 4

# *Chapitre 4 : transfert horizontal d'un rétrovirus murin dans la lignée cellulaire humaine Hep2-clone 2B 85011412-1VL*

Les études présentées dans les chapitres précédents se sont attachées à décortiquer les premières étapes du transfert horizontal (TH) hôte-virus, apportant des évidences d'intégration d'ET hôtes dans les génomes viraux. La présente étude se démarque quelque peu des précédentes et se focalise sur le TH virus-hôte en caractérisant l'intégration d'un virus xénotrope de la leucémie murine (ou XMLV) proche du XMLV Bxv1, dans la lignée cellulaire humaine Hep2-clone 2B. Cette intégration a été identifiée par l'analyse de données génomiques et transcriptomiques de cette lignée cellulaire. Le génome entier du virus est intégré dans le gène codant la *pseudouridylate synthase 1* (*PUS1*) et est exprimé. Un événement d'épissage favorisant l'expression du gène viral d'enveloppe (*env*) a été identifié. Cette étude s'inscrit dans la thématique des éléments viraux endogènes qui sont de plus en plus décrits au fur et à mesure de l'étude des génomes.

Cette étude a été initiée en collaboration avec le laboratoire de virologie du centre hospitalier universitaire de Poitiers. Si au début le projet était destiné à détecter de potentiels nouveaux virus impliqués dans un phénotype de lyse cellulaire, il a rapidement été question de contamination cellulaire. L'analyse des données génomiques laissait penser à des cellules HeLa et non pas Hep2-clone 2B. Or, cette dernière est connue pour être contaminée par des cellules HeLa. Cela paraissait censé puisque la présence d'un fragment de génome du papillomavirus humain 18 inséré dans le génome de la lignée cellulaire a été détectée et ce fragment est connu pour être intégré au génome d'HeLa. Ensuite, l'intégration du génome complet de Bxv1 a permis d'étoffer l'analyse présentée ci-dessous.

Cette étude fait l'objet d'un article dont je suis premier auteur qui a été accepté pour publication dans la revue *Scientific Reports*.

# Characterization of a new case of XMLV (Bxv1) contamination in the human cell line Hep2 (clone 2B)

Vincent Loiseau[1], Richard Cordaux[2], Isabelle Giraud[2], Agnès Beby-Defaux[3], Nicolas Lévêque[4], Clément Gilbert[1*]

[1] Laboratoire Evolution, Génomes, Comportement, Écologie, Unité Mixte de Recherche 9191 Centre National de la Recherche Scientifique et Unité Mixte de Recherche 247 Institut de Recherche pour le Développement, Université Paris-Sud, 91198, Gif-sur-Yvette, France.
[2] Université de Poitiers, CNRS UMR 7267 Laboratoire Ecologie et Biologie des Interactions, Equipe Ecologie Evolution Symbiose, 5 Rue Albert Turpain, TSA 51106, 86073, Poitiers Cedex 9, France.
[3] Laboratoire de Virologie et de Mycobactériologie, CHU de Poitiers, Poitiers, France; Unité de Microbiologie Moléculaire et Séquençage, CHU de Poitiers, Poitiers, France.
[4] Laboratoire de Virologie et de Mycobactériologie, CHU de Poitiers, Poitiers, France; EA4331-LITEC, Université de Poitiers, Poitiers, France.

* Author for correspondence: clement.gilbert@egce.cnrs-gif.fr

## Abstract

The use of misidentified cell lines contaminated by other cell lines and/or microorganisms has generated much confusion in the scientific literature. Detailed characterization of such contaminations is therefore crucial to avoid misinterpretation and ensure robustness and reproducibility of research in molecular and cell biology. Here we use DNA-seq data produced in our lab to first confirm that the Hep2 (clone 2B) cell line (Sigma-Aldrich catalog number: 85011412-1VL) is indistinguishable from the HeLa cell line by mapping integrations of the human papillomavirus 18 (HPV18) at their expected loci on chromosome 8 of the HeLa genome. We then show that the cell line is also contaminated by a xenotropic murine leukemia virus (XMLV) that is nearly identical to the mouse Bxv1 provirus and we characterize one Bxv1 provirus, located in the second intron of the pseudouridylate synthase 1 (*PUS1*) gene. Using an RNA-seq dataset, we confirm the high expression of the E6 and E7 HPV18 oncogenes, show that the entire Bxv1 genome is moderately expressed, and retrieve a Bxv1 splicing event favouring expression of the *env* gene, as previously found in the Bxv1-contaminated human JY B-cell line. Hep2 (clone 2B) is the fourth human cell line so far known to be contaminated by the Bxv1 XMLV. We believe that contamination by this virus happened through contact with a contaminated cell line or reagent as Hep2 (clone 2B) was not generated through passage in immunodeficient mice. This contamination has to be taken into account when using the cell line in future experiments.

## Introduction

Continuous cell lines are a cornerstone of cellular and molecular biology as well as of biomedical research. A long-known problem faced by researchers using such cell lines is contamination, whereby foreign cells or microorganisms are inadvertently introduced and remain unnoticed over multiple passages [1,2]. Characterizing cell culture contamination is of prime importance as contaminants may generate flawed experimental results and considerable confusion in the scientific literature [3,4]. The catalog of cell lines misidentified or cross-contaminated with another cell line contains 529 entries as of October 2019 [5]. Among the 143 different contaminants listed, the most common is HeLa (118 entries), the first continuous human cell line ever established. HeLa derives from a cervical adenocarcinoma biopsy sampled in 1951 from Henrietta Lacks. Sequencing of the HeLa genome resolved a highly rearranged region on chromosome 8 containing multiple integrated partial genomes of the human papillomavirus 18 (HPV18), thought to have induced tumorigenesis [6]. In addition to standard short tandem repeat (STR) typing, known virus-HeLa cell junctions of HPV18 integrations have been used to unveil cases of cancer cell lines contamination by HeLa [7].

Multiple microorganisms such as bacteria, fungi, viruses or yeasts have also been identified as persistent contaminants of widely-used cell lines [8–10]. Viral contaminations may be particularly problematic because viruses may present a risk for the persons handling cell cultures and unlike other organisms, viruses cannot be easily detected under light microscopy [9]. Screenings of hundreds of human cell lines have so far revealed only a relatively small proportion of them as contaminated with human viruses [11–13]. Most cases involve contamination by the Epstein Barr herpesvirus (EBV), either because primary cell cultures were established from infected patients or because EBV was intentionally used to generate transformed lymphoblastoid cell lines.

Non-human viruses have also been found in human cell lines, with murine leukemia viruses (MLV) being responsible for most cases of contaminations [11,13–19]. MLV are gammaretroviruses found under both exogenous and endogenous (proviral) forms in wild and laboratory mouse (*Mus musculus*) strains, which are associated with neoplasias of hematopoietic origin [20]. They are classified in several groups according to their host tropism, which is determined by interactions between their envelope-encoded receptor binding domain and host receptors [21]. While ecotropic MLV interact with the mCAT-1 receptor and can only infect mouse, polytropic MLV interact with the XPR1 receptor and can infect both mouse and other mammalian species. The xenotropic MLV (XMLV) also interact with the XPR1 receptor but they can only infect non-mouse species. Mouse strains are immunized against XMLV thanks to the presence of

several *Xpr1* variants that restrict interaction with these viruses [22]. Most cases of human cell contaminations by MLV involve XMLV and it has been shown that contamination by XMLV often leads to important changes in cellular behavior, which can severely affect the conclusions of studies using such infected cell lines [15].

Contamination by XMLV is known to occur either when a human tumor is xenografted in immunodeficient mice to establish or replicate human tumor cell lines, or through unintentional contact between contaminated and non-contaminated cell lines [23–25]. One of the best characterized cases of XMLV contamination is that of the so-called xenotropic murine leukemia virus-related virus (XMRV), which was originally thought to be a possible cause of prostate cancer and chronic fatigue syndrome [26]. XMRV was later shown to result from recombination between two distinct XMLV sequences integrated in the mouse genome (proviruses) and its presence in human cell lines was traced back to a contamination that occurred around 1993, when a prostate cancer cell line (CWR22Rv1) was xenotransplanted into immunocompromised mice [27]. Other well-characterized cases of XMLV contamination involve a virus originating from expression of a mouse chromosome 1 proviral locus called the B10 xenotropic virus 1 (Bxv1) [28]. This provirus is present in many severe combined immunodeficient (SCID) and nude mice strains [20] and Bxv1 infection of human cell lines passaged onto these mice strains has been experimentally demonstrated [24]. So far, Bxv1 contamination has been detected in two prostate cancer cell lines, one B-cell line and two pancreatic β cell lines, in which it was shown to produce infectious viral particles [16,17,29].

Here, we report a new case of a human cell line (Hep2[clone2B], Sigma-Aldrich catalog number: 85011412-1VL) contaminated by Bxv1 using high-throughput RNA- and DNA-seq. As part of a study aiming to characterize the potential presence of infectious agents in the cell lines of our laboratory, we set out to sequence total RNA and DNA extracted from Hep2 (clone 2B) cells. We characterized sequences from two viruses, the HPV18 papillomavirus and the Bxv1 XMLV. The presence of the former is consistent with the known contamination status of Hep2 (clone 2B) cell line by HeLa and the presence of the latter most likely results from mixing with a cell line or reagents contaminated with this virus.

## Materials and methods

### Cell cultures and DNA- and RNA-seq

Hep2 (clone 2B) cells were purchased from Sigma-Aldrich (catalog number: 85011412-1VL) on June 19, 2013 and cultured in Dulbecco's modified Eagle medium (DMEM; Invitrogen) supplemented with 10% fetal bovine serum and 1% penicillin-streptomycin (Pen-Strep; Life Technologies) at 37°C in a 5% (vol/vol) CO2 atmosphere. Upon confluence, cells contained in a T75 flask were harvested in a 15ml Falcon tube after trypsinization, washed in Earle's balanced salt solution (EBSS) and centrifuged at 1,100 g for 5 min. Total RNA and DNA were then extracted using the AllPrep DNA/RNA Mini Kit (Qiagen). A paired-end DNA library (mean insert size 191 bp) was constructed with the DNA sample and a stranded RNA library was constructed with the RNA sample after rRNA depletion. Both libraries were tagged and sequenced on an Illumina HiSeq2500 platform in a 2 x 125 bp configuration in High Output mode (V4 chemistry). Raw DNA- and RNA-seq fastq reads generated for this study have been deposited in dbGaP under accession number phs001944.v1.p1. All reads mapping onto the Bxv1 (JF908815) genome are provided in Supplementary Data 1 and 2.

### Bioinformatics analyses

Demultiplexing was performed by the sequencing company (Genewiz), yielding 140,313,660 reads with a mean quality score of 35.15 for the DNA sample and 57,593,594 reads with a mean quality score of 35.82 for the RNA sample. Remaining adapter sequences were removed and reads were trimmed with Trimmomatic using default parameters ("-phred33, Illuminaclip:adapter_file.fasta:2:30:10, minlen:126", Bolger et al. 2014). To identify infectious agents potentially present in the cells, we mapped paired reads onto the human genome (UCSC genome data, human genome version hg19) using Bowtie2 in "sensitive-local" mode (Langmead and Salzberg 2012). Reads that did not map onto the human genome were assembled for each dataset. DNA-seq and RNA-seq reads were assembled with Masurca [30] and Trinity [31], respectively, using default settings. The resulting contigs were used as queries to perform blastn and blastx searches on the non-redundant (NR) database of GenBank. To assess sequencing depth of the two viruses identified in this study (Bxv1 XMLV and HPV18), the reads were mapped onto the genome of these viruses (accessions number: JF908815 for Bxv1 and GQ180792 for HPV18) with Bowtie2 in "sensitive-local" mode. To identify recombination breakpoints within each of the two genomes and between the two genomes and the Hep2 (clone

2B) genome we used all reads as queries to perform separate blastn searches (megablast option) on the human genome and on the Bxv1 and HPV18 genomes. Virus-virus and cell-virus junctions were then searched within reads using the pipeline described in [32]. Briefly, only reads aligning over at least sixteen bp on a genome region *only* and over at least sixteen bp on another genome region *only* were retained. Reads had to align on at least 100 bp of their length. The overlap between alignment on the virus and on the host sequences was set to involve at most 20 bp and at least -5 bp (see figure S6 in Gilbert et al. 2016).

**Checking for contamination by rodent DNA**

To assess the presence of rodent DNA or RNA in our samples, reads were mapped using Bowtie2 in "end-to-end" mode on the latest version of *Mus musculus* (GRCm38/mm10) and *Rattus norvegicus* (GRSC 6.0/rn6). Using the SAMtools depth program, the mean rodent genome coverage as well as the percentage of the genome covered by the reads were determined. The nature of regions covered by more than 8000 reads was checked in the UCSC genome browser.

**PCR verifications**

To rule out the possibility that the Bxv1 contamination occurred in our laboratory, we purchased a second batch of Hep2 (clone 2B) cells from Sigma-Aldrich in May 2018 and searched by PCR for the presence of Bxv1 in this new batch and in the Hep2 (clone 2B) cells we used for the sequencing (batch ordered at Sigma-Aldrich in 2014). Amplification reactions were performed from 5 ng of DNA extracted from Hep2 (clone 2B) cells by using 10 µmol.$L^{-1}$ of each primer (Bxv1_1-F: AAGAGAAAGAGAGGGACCGC; Bxv1_1-R: TTTCCTCCAGTAGCCCCTTG), 3 mM MgCl2, 0.2 mM dNTP and 0.75 unit of DreamTaq DNA polymerase (Thermo Fischer Scientific) under a 35-cycle PCR program (95°C for 4 min; 35 cycles of 95°C for 30 sec, 56°C for 30 sec, 72°C for 15 sec, and 72°C for 10 min). Then a migration of the PCR products was performed on a 1.5% agarose gel at 100V during 25 minutes and bands were visualized with a Bio-Rad Transilluminator Universal Hood II under UV light. To check whether regions of the Bxv1 genome not covered by our DNA-seq dataset were in fact present in Hep2 (clone 2B) cells, we PCR-amplified and Sanger-sequenced three such regions as well as another region covered by our DNA-seq dataset using the following primer pairs: Bxv1_2-F: CCCCAGAAGAGAGAGAAGAAC; Bxv1_2-R: CATTGGTCCTTATCGAGTTGG; Bxv1_3-F: TGCCTTTGAGTGGAGAGATC; Bxv1_3-R: CTAGGGTTTGTAGAAGGGCC; Bxv1_4-F: CCTTCTCAACAACCTGGGAC; Bxv1_4-R:

ACAGGGTCAGCTTGTGTTG; Bxv1_5-F: CAGGCAAGCTAACTATGGGA; Bxv1_5-R: CCCAGATTACCTCGGTTTCA.

## Results and discussion

### Confirmation of Hep2 (clone 2B) contamination by HeLa based on HeLa-HPV18 characteristics

As mentioned on the Sigma Aldrich website, the Hep2 (clone 2B) cell line (catalog number: 85011412-1VL) was originally derived "from tumours produced in irradiated-cortisonised weanling rats after injecting with epidermoid carcinoma tissue from the larynx of a 56-year-old male, but it was later found to be indistinguishable from HeLa by STR PCR DNA profiling." It is a typical case of non-existent cell line that may cause important problems in cancer research [33]. Consistent with HeLa contamination, assembling of RNA- and DNA-seq reads obtained from sequencing this cell line and not mapping on the hg19 human genome yielded various contigs that were almost identical to the HPV18 (GenBank accession number: NC_001357.1). As previously reported [7] for HPV18 sequences integrated into the HeLa genome, mapping of both RNA- and DNA-seq reads onto the HPV18 genome shows that a portion of the E2 and L1 genes as well as the entire E4, E5 and L2 genes are missing (Figure 4.1). Moreover, the mapped regions displayed the same 23 SNPs between the reads and the HPV18 genome than those identified between HeLa-HPV18 and HPV18 (Cantalupo et al. 2015) (Supplementary Table 4.1). Mean DNA-seq depth varies from 1.3X for the partial E2 gene up to 26.4X for the L1 gene (Figure 4.1A), reflecting the complex and partially duplicated structure of the integrants [6]. Mean RNA-seq depth varies from 137X for the partial E2 gene up to 4078 and 8093X for the E6 and E7 genes, respectively. This is in agreement with the known high expression of the latter two genes, which are involved in oncogenesis through neutralization of tumour suppressors [6,34,35]. Finally, our search for integration loci in DNA-seq reads revealed four virus-cell junctions supported by one or more reads and also supported by RNA-seq reads (Table 4.1; Supplementary Table 4.2). All four junctions fall within the chromosome 8 region where HPV18 is known to be integrated in HeLa cells (8q24.21; between positions 128,228,000 and 128,243,000) [6,7]. Much like in Cantalupo et al. (2015), we also identified a number of other junctions in RNA-seq reads, among which all those supported by more than one reads fall within the 8q24.21 region (Table 4.1; Supplementary Table 4.2). As noticed by Cantalupo et al. (2015), many of the virus-cell RNA junctions involve the 929 5' splice site in E1 of HPV18, likely indicating that after read-through transcription, splicing events fused the 929 5' donor to

a downstream acceptor site located in the human genome [7,36]. Altogether, these results confirm that Hep2 (clone 2B) was contaminated by HeLa at some point during its propagation and they show that HPV18 sequences integrated in this subclone of HeLa have the same characteristics as in other subclones in terms of position in the genome and expression pattern [6,7,34].

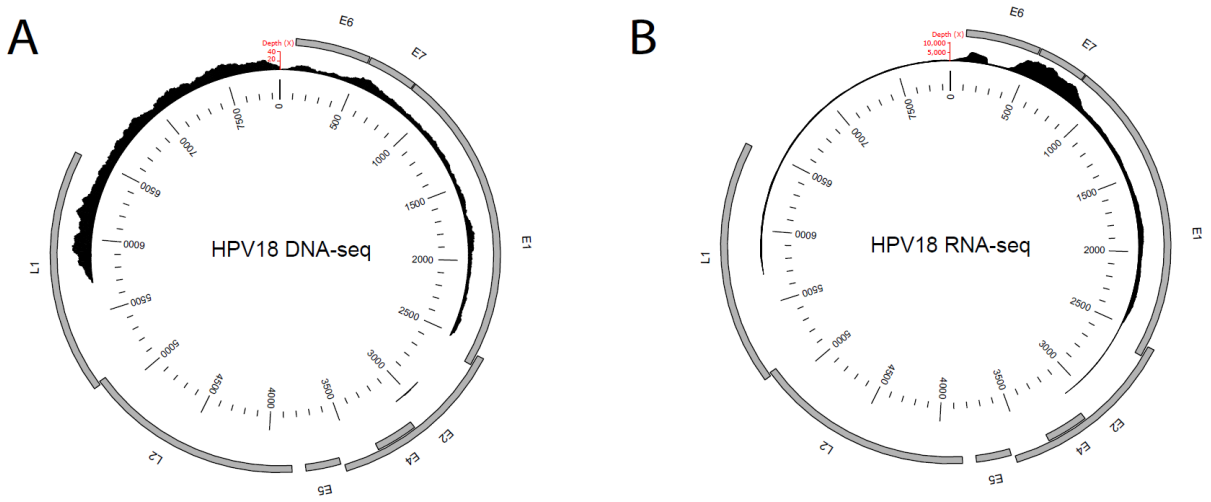



**Figure 4.1: HPV18 sequencing depth by DNA-seq (A) and RNA-seq (B) reads.** The coverage patterns indicate that a portion of the E2 and L1 genes as well as the entire E4, E5 and L2 genes are missing and that the E6 and E7 oncogenes are highly expressed.

**Hep2 (clone 2B) (HeLa) cells are contaminated by Bxv1 XMLV sequences**

In addition to HPV18 integrants, our assembly of non-human reads also yielded contigs 100% identical to Bxv1, an XMLV proviral locus known to generate infectious particles able to infect human cells and so far identified as a contaminant in four human cell lines [17,20,29]. As several studies have shown that the presence of XMLV in human cell lines could be due to contamination by mouse DNA [19,37], we checked for such contamination by mapping both DNA-seq and RNA-seq reads onto the mouse genome. Given that the original Hep2 cells were obtained by passaging larynx carcinoma cells onto immune-compromised laboratory rats [38,39], we also monitored the possible presence of rat DNA by mapping all reads onto the rat genome. Only ≃4% of the reads mapped onto the rodent genomes, with only ≃1% of the genome covered and a mean sequencing depth of 0.26X for the two species. Importantly, mapped regions corresponded exclusively to RNA genes (7SK, 7SL, U1, U2) that are known to be highly similar between mammalian species and no read was found to map onto intracisternal-A particle elements, which are rodent-specific transposable elements typically used as markers of contamination [19,37]. Thus, the presence of Bxv1 is not due to contamination by rodent DNA but

rather results from mixing with a contaminated cell line or reagent. To further verify that contamination did not occur in our laboratory subsequent to reception of Hep2 (clone 2B) cells on June 19 2013 we purchased a second batch from Sigma-Aldrich on May 11, 2018 and validated the presence of Bxv1 sequences by PCR (Supplementary Figure 4.1). We conclude that Hep2 (clone 2B) cells from Sigma-Aldrich (catalog number: 85011412-1VL) are not only undistinguishable from HeLa but they are also contaminated by Bxv1 XMLV sequences.

**Characterization and expression of a Bxv1 provirus**

To further validate the presence of Bxv1 in the HeLa-contaminated Hep2 (clone 2B) cell line, we searched for evidence of integration of the retrovirus. We identified two virus-cell junctions covered by more than one DNA-seq read and/or also supported by an RNA-seq read (Supplementary Data 3). Interestingly, the two junctions involve the very first and very last position of the long terminal repeat of Bxv1 and the same position in the human genome, suggesting that they correspond to the 5' and 3' extremities of one same proviral locus. Alignment of the junctions with the corresponding region of the human genome revealed that Bxv1 generated a 5-bp target site duplication (AAACC) upon integration (Supplementary Data 3). Much like Bxv1 proviruses from two prostate cancer cell lines (LAPC4 and VCaP) which all integrated into introns [17], the Bxv1 provirus characterized here lies within the second intron of the pseudouridylate synthase 1 (*PUS1*) gene. Furthermore, in agreement with the known propensity of MLV to preferentially target transcription start sites (TSS) and CpG islands [40,41], the Hep2 (clone 2B) Bxv1 lies only 946 bp downstream of the nearest *PUS1* TSS, well within a 1463-bp long CpG island. Altogether, these results indicate that the Bxv1 integration identified in Hep2 (clone 2B) cells is a *bona fide* proviral locus, confirming contamination by this virus of the Hep2 (clone 2B) cell line.

To assess how many proviral loci may segregate in the cell line, we mapped all reads on the Bxv1 genome (accession number: JF908815). While several short segments amounting to 15% of the Bxv1 genome are not covered by DNA-seq reads (Figure 4.2A), we believe this is due to stochastic under-representation rather than true absence of some regions because Bxv1 is fully covered by RNA-seq reads (Figure 4.2B). In agreement with this, we were able to PCR-amplify and Sanger-sequence three regions not covered by DNA-seq reads (Figure 4.2A, Supplementary Data 4). Thus we believe that Hep2 (clone 2B) cells contain at least one full-length Bxv1 provirus. Mean DNA-seq depth of Bxv1 is 2.2X, which is lower than that calculated over the entire human genome (6.4X). It is thus possible that the copy that we were

able to map is the only one segregating in the cell line. However, we cannot exclude that multiple proviruses are present, which altogether amount to one or less than one provirus per cell (i.e. some cells may be free of integrated Bxv1). The low number of Bxv1 proviruses may be due to the low capacity of the virus to replicate in these cells. Yet, the entire Bxv1 genome is expressed in Hep2 (clone 2B) cells, including the long terminal repeats which are necessary for the virus to replicate and integrate into the host genome, with mean RNA-seq depths varying from 365X for the *gag-pro-pol* open reading frame (ORF) to 1955X for the *env* ORF (Figure 4.2B). However, the level of Bxv1 expression is relatively low, with only 26,985 (or 0.1% of) RNA-seq reads mapping to the virus, which only ranks 21,906th when listing all human transcripts by decreasing number of mapped RNA-seq reads (Supplementary Table 4.3). This is much lower than the level of expression of Bxv1 measured in the JY B cell line, in which proviral transcripts ranked first in terms of number of aligned reads, being mapped by almost 1,000,000 reads representing 2.6% of all RNA-seq reads [16].

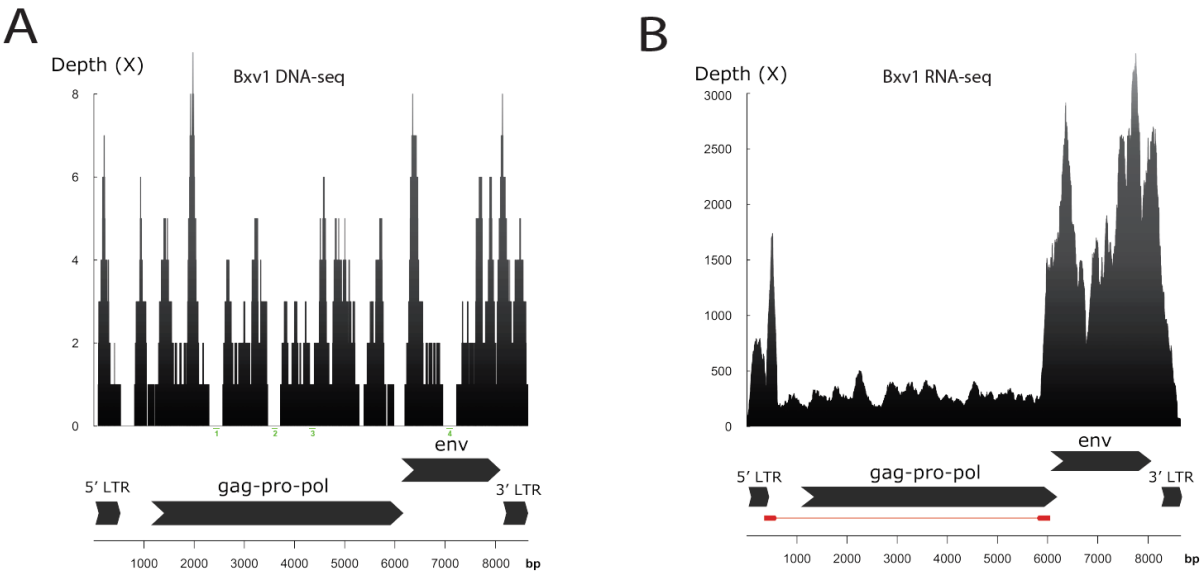



**Figure 4.2: Bxv1 sequencing depth by DNA-seq (A) and RNA-seq (B) reads.** The coverage patterns indicate that the entire Bxv1 genome is present in Hep2 (clone 2B) cells and that the *env* ORF is more expressed than the gag-pro-pol ORF. The green lines in Figure 4.2 A show the position of viral regions that were Sanger-sequenced. Numbers below the green lines correspond to those used to name Sanger-sequencing reads in Supplementary Data 4. The red line in Figure 4.2 B represents the spliced region between positions 587 and 5887.

In the Lin et al. (2012) study, 768 RNA-seq reads were found to contain a junction between positions 587 and 5887 of Bxv1 (accession number: JF908815), which corresponds to a splicing

event leading to the expression of the *env* ORF. Our search for virus-virus junctions unveiled a nearly identical splicing event, with 757 reads (145 reads when potential PCR duplicates are removed), supporting a junction between positions 587 and 5884 (Figure 4.2). The position of the splice donor site matches exactly between our study, that of Lin et al. (2012) and the annotation of the N417 MLV (accession number: HQ246218) which is 99.98% identical to Bxv1 JF908815. However, the position of the splice acceptor site differs (5884 in Lin et al.; 5887 here; 5219 in HQ246218 corresponding to 5601 in JF908815), which may be due to context-dependent variations in splicing. To assess whether the presence of Bxv1 in Hep2 (clone 2B) cells could result from mixing with the JY B cell line, we checked for the presence of RNA-seq and/or DNA-seq reads mapping onto the EBV genome, which is known to be present in JY B cells [16]. We did not find any read mapping on the EBV genome. While this suggests that the Bxv1 contamination is not due to a mix with JY B cells, we cannot exclude that a transient mix occurred between Hep2 (clone 2B) and JY B cells at some point but that JY B cells are no longer detectable. Interestingly, the RNA-seq coverage of the region corresponding to the alternative *env* transcript is three to five times higher than that of the rest of the Bxv1 sequence (Figure 4.2B) showing that for some reason, the splicing generating this transcript is markedly favoured in Hep2 (clone 2B) cells. Rather than alternative splicing, the junction we found in RNA-seq reads between positions 587 and 5887 of Bxv1 could be due to transcription of a deleted Bxv1 copy, that would segregate in some cells in addition to a full length copy. To check for the presence of a deleted Bxv1 copy we designed PCR primers on both sides of the deletion. All PCRs performed using those primers were negative, suggesting that the truncated transcripts unlikely result from transcription of a deleted Bxv1 proviral copy. We have not tested whether the higher expression of the *env* ORF translates into an accumulation of the ENV protein, which has been linked to the generation of cytopathic effects in some MLV [42]. That said, we have not observed any cytopathic effect, in agreement with the fact that most MLV do not generate such effects [43].

In their study of the JY B cell line contaminated by Bxv1, Lin et al. (2012) also found evidence of G-to-A editing likely resulting from the activity of the APOBEC3G restriction factor, which induces deamination of cytidine to uridine in single stranded DNA viral intermediates [44]. Specifically, Lin et al. (2012) found that 44 out of the 45 SNPs identified in their RNA-seq reads were G-to-A changes, indicating that they likely resulted from transcription of APOBEC3G-edited viral genomes. In agreement with the known absence of APOBEC3G in HeLa cells [45], the APOBEC3G transcript ranks only 37,052th in the list of human transcripts

ordered by the number of mapped reads (Supplementary Table 4.3). The number of SNPs supported by more than 10 reads and having a frequency >2% in our RNA-seq data is anyway too low to draw any conclusion on the possible activity of APOBEC3G in Hep2 (clone 2B) cells, but it is worth noting that four of the five SNPs we identified using those criteria in our RNA-seq data are G-to-A changes.

## Conclusion

In this study, we have confirmed that the Hep2 (clone 2B) cell line (Sigma-Aldrich catalog number: 85011412-1VL) is contaminated by HeLa cells based on the characterization of HPV18 integrants and expression patterns. We have further demonstrated that the cell line is also contaminated by a virus nearly identical to the Bxv1 XMLV provirus, the presence of which is likely due to direct contact between the cell line and another contaminated cell line or reagent. Based on sequencing depth, we show that the cell line might contain only one Bxv1 provirus, which we were able to map to the second intron of the *PUS1* gene. While our study does not demonstrate that Bxv1 is able to replicate in Hep2 (clone 2B) cells, it shows that it is expressed at a moderately high level, which may impact various cellular pathways. Thus, the presence of this virus in otherwise HeLa-contaminated Hep2 (clone 2B) cells will have to be taken into consideration in future studies using this cell line to avoid erroneous interpretations of experimental results. Furthermore, this study should also encourage others using this cell line and related ones, such as the (code A) 86030501-1VL, or 85020207-1VL Hep-2C (HeLa derivative) Human Negroid cervix carcinoma (code C), to check for the presence of Bxv1 using the PCR primers we provide in the materials and methods section.

**Table 4.1: Characteristics of HeLa – HPV18 junctions supported by both DNA- and RNA-seq reads. *** Several steps involved in the construction of an Illumina library (including cDNA library synthesis and illumina PCR) may generate artificial chimeras[46,47]. Thus, relying only on one read to identify a breakpoint is not good practice. However, this Table only reports HPV18-HeLa breakpoints that are supported by reads generated in two independent sequencing experiments (RNA-seq and DNA-seq), including one in which they are supported by multiple reads (here more than 15). For example, the first lane of the Table describes a breakpoint between position 929 of the HPV18 genome and position 128 241 377 of human chromosome 8 that is retrieved in 1 DNA-seq read and in 272 RNA-seq reads independently. The position of all breakpoints reported here is consistent with those reported in earlier studies (see text for details).

| **Number of DNA-seq reads supporting the junction** | **Number of RNA-seq reads supporting the junction** | **Viral breakpoint position** | **Viral gene** | **Position of breakpoint in human genome** | **Human chr. Band** |
|---|---|---|---|---|---|
| 1* | 272 | 929 | E1 | 128241377 | 8q24.21 |
| 25 | 97 | 5735 | L1 | 128230628 | 8q24.21 |
| 1 | 57 | 2497 | E1 | 128241551 | 8q24.21 |
| 1 | 15 | 930 | E1 | 128231213 | 8q24.21 |

## Acknowledgements

This study was supported by Agence Nationale de la Recherche (ANR-15-CE32-0011-01 TransVir).

## Author Contribution Statement

C.G, N. L., A. B.-D., R.C. conceived the study. I.G. performed research. V.L. and C.G. performed research and wrote the first draft of the manuscript. All authors reviewed the manuscript.

## Competing interests

The author(s) declare no competing interests.

## Data availability

In addition to the Supplementary Figure 4.1 and Supplementary Data 4.1-4.4 included in this published article, the raw fastq reads generated during the current study are available in the dbGaP repository under accession number phs001944.v1.p1.

## References

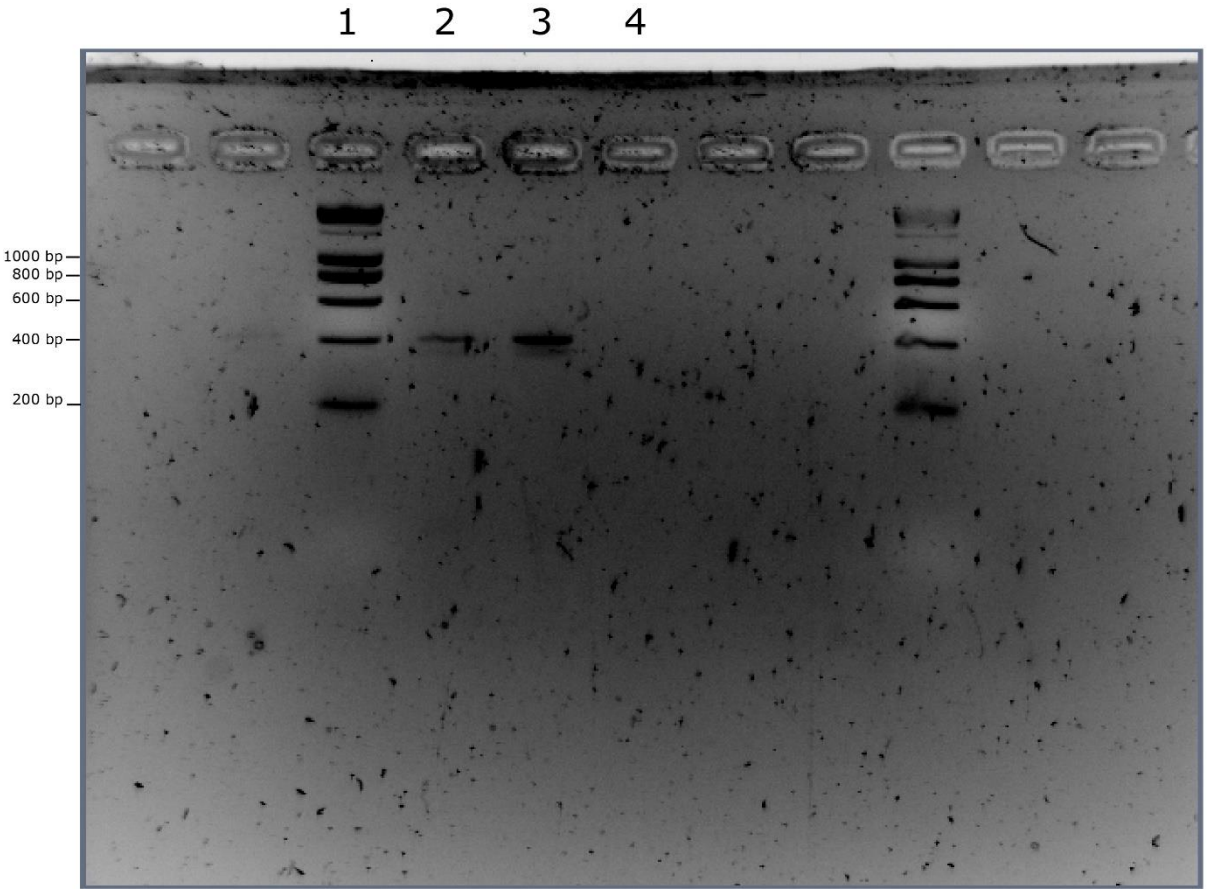

1. Gartler, S. M. Apparent HeLa Cell Contamination of Human Heteroploid Cell Lines. *Nature* **217**, 750–751 (1968).
2. Nelson-Rees, W., Daniels, D. & Flandermeyer, R. Cross-contamination of cells in culture. *Science* **212**, 446–452 (1981).
3. Horbach, S. P. J. M. & Halffman, W. The ghosts of HeLa: How cell line misidentification contaminates the scientific literature. *PLOS ONE* **12**, e0186281 (2017).
4. Nardone, R. M. Eradication of cross-contaminated cell lines: A call for action. *Cell Biol. Toxicol.* **23**, 367–372 (2007).
5. Capes-Davis, A. *et al.* Check your cultures! A list of cross-contaminated or misidentified cell lines. *Int. J. Cancer* **127**, 1–8 (2010).
6. Adey, A. *et al.* The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature* **500**, 207–211 (2013).
7. Cantalupo, P. G., Katz, J. P. & Pipas, J. M. HeLa Nucleic Acid Contamination in The Cancer Genome Atlas Leads to the Misidentification of Human Papillomavirus 18. *J. Virol.* **89**, 4051–4057 (2015).
8. Drexler, H. G. & Uphoff, C. C. Mycoplasma contamination of cell cultures: Incidence, sources, effects, detection, elimination, prevention. *Cytotechnology* **39**, 75–90 (2002).
9. Merten, O.-W. Virus contaminations of cell cultures - A biotechnological view. *Cytotechnology* **39**, 91–116 (2002).
10. Mirjalili, A., Parmoor, E., Moradi Bidhendi, S. & Sarkari, B. Microbial contamination of cell cultures: A 2 years study. *Biologicals* **33**, 81–85 (2005).
11. Cao, S. *et al.* High-Throughput RNA Sequencing-Based Virome Analysis of 50 Lymphoma Cell Lines from the Cancer Cell Line Encyclopedia Project. *J. Virol.* **89**, 713–729 (2015).
12. Shioda, S. *et al.* Screening for 15 pathogenic viruses in human cell lines registered at the JCRB Cell Bank: characterization of *in vitro* human cells by viral infection. *R. Soc. Open Sci.* **5**, 172472 (2018).
13. Uphoff, C. C., Denkmann, S. A., Steube, K. G. & Drexler, H. G. Detection of EBV, HBV, HCV, HIV-1, HTLV-I and -II, and SMRV in Human and Other Primate Cell Lines. *J. Biomed. Biotechnol.* **2010**, 1–23 (2010).
14. Cmarik, J. L., Troxler, J. A., Hanson, C. A., Zhang, X. & Ruscetti, S. K. The Human Lung Adenocarcinoma Cell Line EKVX Produces an Infectious Xenotropic Murine Leukemia Virus. *Viruses* **3**, 2442–2461 (2011).
15. Hempel, H. A., Burns, K. H., De Marzo, A. M. & Sfanos, K. S. Infection of Xenotransplanted Human Cell Lines by Murine Retroviruses: A Lesson Brought Back to Light by XMRV. *Front. Oncol.* **3**, (2013).
16. Lin, Z. *et al.* Detection of Murine Leukemia Virus in the Epstein-Barr Virus-Positive Human B-Cell Line JY, Using a Computational RNA-Seq-Based Exogenous Agent Detection Pipeline, PARSES. *J. Virol.* **86**, 2970–2977 (2012).

17. Sfanos, K. S. *et al.* Identification of Replication Competent Murine Gammaretroviruses in Commonly Used Prostate Cancer Cell Lines. *PLoS ONE* **6**, e20874 (2011).
18. Takeuchi, Y., McClure, M. O. & Pizzato, M. Identification of Gammaretroviruses Constitutively Released from Cell Lines Used for Human Immunodeficiency Virus Research. *J. Virol.* **82**, 12585–12588 (2008).
19. Uphoff, C. C., Lange, S., Denkmann, S. A., Garritsen, H. S. P. & Drexler, H. G. Prevalence and Characterization of Murine Leukemia Virus Contamination in Human Cell Lines. *PLOS ONE* **10**, e0125622 (2015).
20. Kozak, C. A. The mouse" xenotropic" gammaretroviruses and their XPR1 receptor. *Retrovirology* **7**, 101 (2010).
21. Kozak, C. Origins of the Endogenous and Infectious Laboratory Mouse Gammaretroviruses. *Viruses* **7**, 1–26 (2014).
22. Yan, Y. *et al.* Evolution of Functional and Sequence Variants of the Mammalian XPR1 Receptor for Mouse Xenotropic Gammaretroviruses and the Human-Derived Retrovirus XMRV. *J. Virol.* **84**, 11970–11980 (2010).
23. McALLISTER, R. M. *et al.* C-Type Virus Released from Cultured Human Rhabdomyosarcoma Cells. *Nature. New Biol.* **235**, 3–6 (1972).
24. Naseer, A. *et al.* Frequent Infection of Human Cancer Xenografts with Murine Endogenous Retroviruses in Vivo. *Viruses* **7**, 2014–2029 (2015).
25. Zhang, Y.-A. *et al.* Frequent detection of infectious xenotropic murine leukemia virus (XMLV) in human cultures established from mouse xenografts. *Cancer Biol. Ther.* **12**, 617–628 (2011).
26. Johnson, A. D. & Cohn, C. S. Xenotropic Murine Leukemia Virus-Related Virus (XMRV) and the Safety of the Blood Supply. *Clin. Microbiol. Rev.* **29**, 749–757 (2016).
27. Paprotka, T. *et al.* Recombinant Origin of the Retrovirus XMRV. *Science* **333**, 97–101 (2011).
28. Kozak, C. & Rowe, W. Genetic mapping of xenotropic leukemia virus-inducing loci in two mouse strains. *Science* **199**, 1448–1449 (1978).
29. Kirkegaard, J. S. *et al.* Xenotropic retrovirus Bxv1 in human pancreatic β cell lines. *J. Clin. Invest.* **126**, 1109–1113 (2016).
30. Zimin, A. V. *et al.* The MaSuRCA genome assembler. *Bioinformatics* **29**, 2669–2677 (2013).
31. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
32. Gilbert, C. *et al.* Continuous Influx of Genetic Material from Host to Virus Populations. *PLoS Genet.* **12**, e1005838 (2016).
33. Gorphe, P. A comprehensive review of Hep-2 cell line in translational research for laryngeal cancer. *Am. J. Cancer Res.* **9**, 644–649 (2019).
34. Akagi, K. *et al.* Genome-wide analysis of HPV integration in human cancers reveals recurrent, focal genomic instability. *Genome Res.* **24**, 185–199 (2014).

35. Bouallaga, I., Massicard, S., Yaniv, M. & Thierry, F. An enhanceosome containing the Jun B/Fra-2 heterodimer and the HMG-I(Y) architectural protein controls HPV18 transcription. *EMBO Rep.* **1**, 422–427 (2000).
36. Wang, X., Meyers, C., Wang, H.-K., Chow, L. T. & Zheng, Z.-M. Construction of a Full Transcription Map of Human Papillomavirus Type 18 during Productive Viral Infection. *J. Virol.* **85**, 8080–8092 (2011).
37. Oakes, B. *et al.* Contamination of human DNA samples with mouse DNA can lead to false detection of XMRV-like sequences. *Retrovirology* **7**, 109 (2010).
38. Moore, A. E., Sabachewsky, L. & Toolan, H. W. Culture characteristics of four permanent lines of human cancer cells. *Cancer Res.* **15**, 598–602 (1955).
39. Toolan, H. W. Transplantable human neoplasms maintained in cortisone-treated laboratory animals: H.S. No. 1; H.Ep. No. 1; H.Ep. No. 2; H.Ep. No. 3; and H.Emb.Rh. No. 1. *Cancer Res.* **14**, 660–666 (1954).
40. De Rijck, J. *et al.* The BET Family of Proteins Targets Moloney Murine Leukemia Virus Integration near Transcription Start Sites. *Cell Rep.* **5**, 886–894 (2013).
41. Sultana, T., Zamborlini, A., Cristofari, G. & Lesage, P. Integration site selection by retroviruses and transposable elements in eukaryotes. *Nat. Rev. Genet.* **18**, 292–308 (2017).
42. Zhao, X. & Yoshimura, F. K. Expression of Murine Leukemia Virus Envelope Protein Is Sufficient for the Induction of Apoptosis. *J. Virol.* **82**, 2586–2589 (2008).
43. Sliva, K., Erlwein, O., Bittner, A. & Schnierle, B. S. Murine leukemia virus (MLV) replication monitored with fluorescent proteins. *Virol. J.* **1**, 14 (2004).
44. Sheehy, A. M., Gaddis, N. C., Choi, J. D. & Malim, M. H. Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. *Nature* **418**, 646–650 (2002).
45. Kremer, M. *et al.* Vaccinia virus replication is not affected by APOBEC3 family members. *Virol. J.* **3**, 86 (2006).
46. Peccoud, J. *et al.* A Survey of Virus Recombination Uncovers Canonical Features of Artificial Chimeras Generated During Deep Sequencing Library Preparation. *G3amp58 GenesGenomesGenetics* **8**, 1129–1138 (2018).
47. Tsai, I. J. *et al.* Summarizing Specific Profiles in Illumina Sequencing from Whole-Genome Amplified DNA. *DNA Res.* **21**, 243–254 (2014).

## Supplementary data

Only the supplementary Figure 4.1 and the supplementary Tables 4.1 and 4.2 are provided.




**Supplementary Figure 4.1: Bxv1 PCR products visualized on an agarose gel.** Lane 1: ladder; Lane 2: a band of the expected size (400 bp) obtained on the 2018 Hep2 (clone 2B) batch; Lane 3: a band of the expected size (400 bp) obtained on the 2013 Hep2 (clone 2B) batch 400-bp band A band is visualised on a 1.5% agarose gel after PCR on Hep2 DNA; Lane 4: H2O.

**Supplementary Table 4.1: 23 SNPs between HeLa-specific HPV18 and the HPV18 genome.** The SNVs exactly match to the HeLa-specific HPV18 genome identified by Cantalupo and colleagues (2015).

| Gene | Position | SNV |
|---|---|---|
| E6 | 104 | T → C |
| | 287 | C → G |
| | 485 | T → C |
| | 549 | C → A |
| | | |
| E7 | 751 | C → T |
| | 806 | G → A |

| | | |
|---|---|---|
| E1 | 1012 | A → T |
| | 1194 | C → A |
| | 1353 | T → A |
| | 1807 | T → C |
| | 1843 | T → G |
| | 2269 | C → T |
| L1 | 5875 | C → A |
| | 6401 | A → G |
| | 6460 | C → G |
| | 6625 | C → G |
| | 6842 | C → G |
| | 7258 | T → A |
| | 7486 | C → T |
| LCR | 7529 | C → A |
| | 7567 | A → C |
| | 7592 | T → C |
| | 7670 | A → T |

**Supplementary Table 4.2: Junctions between both viruses and human genomes.** Junctions were mainly detected in RNA data. Most of junctions involve the E1 HPV18 gene and 8q24.21 human chromosome band.

| Samples with the detected breakpoint | Chimeric reads covering the junction | Viral breakpoint pos. | Involved viral genes | Human breakpoint pos. | Human chr. Band | Involved human genes | Involved human TEs |
|---|---|---|---|---|---|---|---|
| **HPV18-human junctions** | | | | | | | |
| RNA & DNA | 273 (272+1, respectively) | 929 | E1 | 128241377 | 8q24.21 | BC106081-exon | MIR3 |
| RNA & DNA | 122 (97+25, respectively) | 5735 | L1 | 128230628 | 8q24.21 | CCAT1-intron | / |
| RNA | 68 | 22 | / | 128231054 | 8q24.21 | CCAT1-intron | MIR |
| RNA & DNA | 58 (57+1, respectively) | 2497 | E1 | 128241551 | 8q24.21 | / | / |
| RNA | 52 | 929 | E1 | 128241370 | 8q24.21 | BC106081-exon | MIR3 |
| RNA & DNA | 16 (15+1, respectively) | 930 | E1 | 128231213 | 8q24.21 | CCAT1-exon | / |
| RNA | 10 | 929 | E1 | 128240876 | 8q24.21 | BC106081-exon | / |
| RNA | 8 | 929 | E1 | 128239788 | 8q24.21 | / | / |
| RNA | 5 | 929 | E1 | 128221964 | 8q24.21 | CCAT1-intron | / |
| RNA | 5 | 942 | E1 | 128241377 | 8q24.21 | BC106081-exon | MIR3 |
| RNA | 4 | 1357 | E1 | 128241379 | 8q24.21 | BC106081-exon | MIR3 |
| RNA | 4 | 21 | / | 128231059 | 8q24.21 | CCAT1-intron | MIR |
| RNA | 4 | 929 | E1 | 128235913 | 8q24.21 | / | / |
| RNA | 3 | 414 | E6 | 128231052 | 8q24.21 | CCAT1-intron | MIR |
| RNA | 3 | 929 | E1 | 128241374 | 8q24.21 | BC106081-exon | MIR3 |
| RNA | 3 | 931 | E1 | 128241507 | 8q24.21 | / | / |
| RNA | 2 | 1890 | E1 | 128239463 | 8q24.21 | / | / |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| RNA | 2 | 2289 | E1 | 128241082 | 8q24.21 | BC106081-exon | / |
| RNA | 2 | 7454 | / | 128233698 | 8q24.21 | / | MIRb |
| RNA | 2 | 776 | E7 | 128241375 | 8q24.21 | BC106081-exon | MIR3 |
| RNA | 2 | 903 | E1 | 128241375 | 8q24.21 | BC106081-exon | MIR3 |
| RNA | 2 | 908 | E1 | 128241341 | 8q24.21 | BC106081-exon | MIR3 |
| RNA | 2 | 929 | E1 | 128200362 | 8q24.21 | JX003871-exon, CASC19-intron | / |
| RNA | 2 | 930 | E1 | 128232653 | 8q24.21 | / | / |
| RNA | 1 | 120 | E6 | 128241337 | 8q24.21 | BC106081-exon | MIR3 |
| RNA | 1 | 1500 | E1 | 128232770 | 8q24.21 | / | / |
| RNA | 1 | 1539 | E1 | 98115605 | 12q23.1 | LOC643711-intron | L1HS |
| RNA | 1 | 1539 | E1 | 68132275 | 15q23 | / | / |
| RNA | 1 | 1539 | E1 | 180949634 | 3q26.33 | SOX2-OT-intron | / |
| RNA | 1 | 1552 | E1 | 128241418 | 8q24.21 | / | MIR3 |
| RNA | 1 | 1556 | E1 | 133216516 | 4q28.3 | / | / |
| RNA | 1 | 1890 | E1 | 128233294 | 8q24.21 | / | / |
| RNA | 1 | 1987 | E1 | 128241150 | 8q24.21 | BC106081-exon | / |
| RNA | 1 | 21 | / | 128235782 | 8q24.21 | / | / |
| RNA | 1 | 2105 | E1 | 128237342 | 8q24.21 | / | / |
| RNA | 1 | 2254 | E1 | 9934741 | 3p25.3 | JAGN1-exon | / |
| RNA | 1 | 23 | / | 128231055 | 8q24.21 | CCAT1-exon | MIR |
| RNA | 1 | 233 | E6 | 128174331 | 8q24.21 | / | / |
| RNA | 1 | 557 | E6 | 128239674 | 8q24.21 | / | MIR |
| RNA | 1 | 5811 | L1 | 128230765 | 8q24.21 | CCAT1-intron | MIRb |
| RNA | 1 | 6365 | L1 | 169742576 | 6q27 | / | MER94 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| RNA | 1 | 680 | E7 | 128241019 | 8q24.21 | BC106081-exon | / |
| RNA | 1 | 7702 | / | 893593 | 19p13.3 | / | AluSg |
| RNA | 1 | 860 | E7 | 70824505 | 12q15 | KCNMB4-exon | / |
| RNA | 1 | 884 | E7 | 128241206 | 8q24.21 | BC106081-exon | / |
| RNA | 1 | 926 | E1 | 128241377 | 8q24.21 | BC106081-exon | MIR3 |
| RNA | 1 | 928 | E1 | 128241374 | 8q24.21 | BC106081-exon | MIR3 |
| RNA | 1 | 929 | E1 | 128091380 | 8q24.21 | PCAT2-intron | L1M2 |
| RNA | 1 | 929 | E1 | 128231391 | 8q24.21 | CCAT1-exon | / |
| RNA | 1 | 930 | E1 | 128181364 | 8q24.21 | / | L1MC1 |
| RNA | 1 | 930 | E1 | 128200133 | 8q24.21 | JX003871-intron, CASC19-intron | / |
| RNA | 1 | 930 | E1 | 128215467 | 8q24.21 | / | L1ME3 |
| RNA | 1 | 931 | E1 | 128236722 | 8q24.21 | / | L2C |
| RNA | 1 | 931 | E1 | 128241374 | 8q24.21 | BC106081-exon | MIR3 |
| RNA | 1 | 936 | E1 | 128241377 | 8q24.21 | BC106081-exon | MIR3 |
| RNA | 1 | 978 | E1 | 128241348 | 8q24.21 | BC106081-exon | MIR3 |
| RNA | 1 | 979 | E1 | 128233301 | 8q24.21 | / | / |
| DNA | 1 | 1395 | E1 | 15446806 | 5p15.1 | / | / |
| DNA | 1 | 6171 | L1 | 92915224 | 6q15 | / | THE1D, HERVL-MaLR |
| DNA | 1 | 6371 | L1 | 129014698 | 9q33.3 | / | THE1B, HERVL-MaLR |
| DNA | 1 | 692 | E7 | 64100489 | 1p31.3 | PGM1-intron | / |
| DNA | 1 | 7326 | / | 203156088 | 1q32.1 | / | / |
| DNA | 1 | 741 | E7 | 48589338 | 12q13.11 | / | L1ME5 |
| DNA | 1 | 7460 | / | 56118967 | 5q11.2 | MAP3K1-intron | L2b |

| Bxv1-human junctions | | | | | | | |
|---|---|---|---|---|---|---|---|
| DNA | 3 | 0 | / | 132414759 | 12q24.33 | PUS1-intron | / |
| RNA & DNA | 2 (1+1, respectiveley) | 525 | / | 132414764 | 12q24.33 | PUS1-intron | / |
| RNA | 1 | 153 | / | 36894260 | 6p21.2 | C6orf89-exon | AluSx1 |
| RNA | 1 | 182 | / | 53383324 | 8q11.23 | / | / |
| RNA | 1 | 3208 | Gag-pro-pol | 50721247 | 22q13.33 | PLXNB2-exon | / |
| RNA | 1 | 5885 | Gag-pro-pol | 286632 | 9p24.3 | DOCK8-exon | / |
| RNA | 1 | 5905 | Gag-pro-pol | 30546653 | 15q13.2 | DKFZP434L187-intron | snRNA U6 |
| RNA | 1 | 5905 | Gag-pro-pol | 893541 | 19p13.3 | RNU6-9-exon | snRNA U6 |
| RNA | 1 | 5987 | Gag-pro-pol | 24560675 | Xp22.11 | PDK3-exon | / |
| RNA | 1 | 6728 | Env | 74622910 | 9q21.13 | / | / |
| RNA | 1 | 6896 | Env | 51728140 | 15q21.2 | / | L1MA1 |
| RNA | 1 | 7354 | Env | 118799 | Un_gl000220 | LOC100507412-intron | / |
| RNA | 1 | 7662 | Env | 22851667 | 1p36.12 | ZBTB-intron | Charlie15a |
| RNA | 1 | 8009 | Env | 145349364 | 2q22.3 | / | / |
| RNA | 1 | 8031 | Env | 32677706 | 20q11.22 | EIF2S2-exon | / |
| RNA | 1 | 8080 | Env | 34939048 | 5p13.2 | DNAJC21-exon | / |
| RNA | 1 | 8188 | / | 1153 | M | / | / |
| RNA | 1 | 845 | / | 144624402 | 8q24.3 | / | 7SK repeat |
| DNA | 1 | 1014 | / | 53986953 | 20q13.2 | / | MIR |
| DNA | 1 | 3973 | Gag-pro-pol | 56999943 | 1p32.2 | PLPP3-intron | / |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| DNA | 1 | 6059 | Gag-pro-pol | 58280138 | 6p11.2 | LINC00680-intron | AluSq2 |
| DNA | 1 | 7380 | Env | 104488072 | 7q22.2 | LHFLP3-intron | AluSz |

**Supplementary Data 4.1: Fastq file containing all raw DNA-seq reads aligning on the Bxv1 (JF908815) genome.**

**Supplementary Data 4.2: Fastq file containing all raw RNA-seq reads aligning on the Bxv1 (JF908815) genome.**

**Supplementary Data 4.3: Alignment of DNA-seq reads supporting the Bxv1 proviral locus characterized in this study.** Aligned sequences are numbered from 1 to 6. 1: the sequence of the human genome flanking the Bxv1 proviral locus, comprising 100 bp upstream and downstream of the provirus, located in the second intron of the *PUS1* gene. 2: full length Bxv1 genome sequenced from the VCaP prostate cancer cell line. 3 – 5: three reads supporting the virus-cell 5' junction. 6: Read supporting the virus-cell 3' junction. Note that the 3' junction is also supported by one RNA-seq read. The integration generated a 5-bp target site duplication (AAACC). The alignment is provided in docx format. It can be pasted and visualized in any alignment viewer such as BioEdit or Geneious.

**Supplementary Data 4.4: Alignment of Bxv1 (JF908815) with Sanger-sequencing reads produced during this study.** The name of each read begins with a number corresponding to the region illustrated on Figure 4.2A. F: forward read. R: reverse read. The name of the read also contains the name of the primers used to PCR-amplify the four Bxv1 regions.

# Discussion

# *Discussion générale et perspectives*

Les transferts horizontaux (TH) d'éléments transposables (ET) chez les métazoaires sont de plus en plus étudiés et leur rôle dans l'évolution des espèces apparaît évident. Les mécanismes responsables de ces TH entre animaux sont mal connus et différentes hypothèses ont été évoquées dans la littérature, comme nous l'avons vu en introduction. L'une de ces hypothèses invoque les virus comme potentiels vecteurs d'ADN entre espèces. Bien qu'il n'y ait à ce jour aucune démonstration formelle du rôle des virus comme vecteurs de matériel génétique entre animaux, des études ont par le passé mis en évidence la présence d'ADN non viral présent dans les génomes ou les capsides de virus. L'objectif de cette thèse était ainsi d'approfondir et de décortiquer les événements génomiques ayant lieu au cours d'une infection virale. C'est pourquoi nous avons étudié tout d'abord l'activité des ET chez la fausse arpenteuse du chou (*Trichoplusia ni*) lors d'une infection par le baculovirus *Autographa californica* multiple nucleopolyhedrovirus (AcMNPV). Nous avons également pu analyser les ET cotranscrits avec les gènes viraux (chapitre 1). Nous nous sommes également intéressés à 11 systèmes hôte-virus différents à travers 35 jeux de données de courtes lectures de séquençage, afin d'élargir nos connaissances sur la diversité des systèmes dans lesquels des ET de l'hôte pouvaient sauter dans des génomes viraux. Cela nous a permis d'avoir une idée de la fréquence d'insertion d'ET pour différents systèmes. Chose inattendue, nous avons pu mettre en évidence un phénomène de transposition d'ET virus-virus, appuyant le fait que la présence d'ET dans les génomes viraux n'est peut-être pas simplement une phase de transition pour un potentiel TH d'ET entre hôtes, mais pourrait aussi constituer une étape importante dans la persistance de certains ET, pouvant être « latents » au sein d'une population virale (chapitre 2). Afin d'en savoir davantage sur la présence des ET au sein des génomes viraux, nous avons utilisé une technologie de séquençage de longues lectures sur des génomes viraux, ce qui nous a permis d'identifier des séquences d'ET complètes insérées au sein des génomes de baculovirus. Nous avons aussi profité de la richesse de ces données pour caractériser l'ensemble des variants structuraux (VS) génomiques (comprenant les insertions, délétions, inversions et duplications >50 pb) au sein de quatre populations virales (chapitre 3). Enfin, l'étude de données de séquençage visant à découvrir un nouveau pathogène dans une lignée cellulaire humaine a abouti à la mise en

évidence d'un cas de TH d'un rétrovirus murin dans la lignée cellulaire humaine Hep2-clone 2B (chapitre 4).

Divers éléments permettent de penser que les virus peuvent être des vecteurs de matériel génétique entre métazoaires. Cela tient à la caractéristique des virus à être infectieux et à pouvoir être transmis à la fois horizontalement et verticalement. Un TH réalisé par l'intermédiaire d'un virus peut être vu comme le résultat de deux étapes successives. Il y a tout d'abord l'acquisition d'un morceau du génome hôte par un virus, puis l'intégration de ce fragment génomique dans le génome d'un autre hôte. La première étape est soutenue par plusieurs études ayant mis en avant la présence d'ET intégrés dans des génomes viraux après infection (Fraser, Smith, et Summers 1983; Gilbert et al. 2016; Loiseau et al. 2020). Concernant la seconde étape, la possibilité que du matériel génétique transporté par un virus puisse s'intégrer dans le génome d'un des hôtes de ce virus est soutenue par la découverte de nombreux éléments viraux endogènes (EVE) dans les génomes cellulaires (eucaryotes et procaryotes; Katzourakis et Gifford 2010; Holmes 2011). L'étude effectuée dans le chapitre 4 concernant l'intégration d'un rétrovirus de souris dans une lignée cellulaire humaine contribue à mieux caractériser la seconde étape d'un TH (TH de matériel génétique du virus vers l'hôte). La thématique des TH chez les animaux, étudiée en biologie de l'évolution s'inscrit dans une perspective évolutive, c'est-à-dire ayant un impact évolutif potentiel. De nombreux exemples ont été révélés dans la littérature scientifique sur ce sujet, dont certains ont été explicités dans l'introduction de ce manuscrit. Les lignées cellulaires utilisées par les scientifiques pour des facilités d'usage en comparaison d'organismes vivants entiers ont peu de chances d'être impliquées dans des phénomènes impactant l'évolution du vivant à long terme. Néanmoins, cette étude a le mérite d'avoir caractérisé finement grâce à des données de séquençage haut débit et pour la première fois dans la lignée cellulaire Hep2-clone 2B, un événement d'intégration du rétrovirus murin Bxv1. Elle vient ainsi s'ajouter à la littérature des contaminations de lignées cellulaires couramment utilisées de par le monde et appelle à tester chacune d'elles avant utilisation pour des expériences.

Concernant le chapitre 1, bien que nous n'ayons pas trouvé un patron d'expression global permettant de conclure à une augmentation globale de l'expression des ET au cours d'une infection virale, nous avons trouvé certains ET surexprimés pendant l'infection et certains ET cotranscrits avec des gènes viraux. Cependant, les données utilisées n'étaient peut-être pas idéales. L'utilisation de lectures faisant 51 pb de long limite leur assignation à une copie précise

d'ET, et entrave également la détection de cotranscrits ET-gènes viraux. L'idéal aurait été de faire une expérience et de générer nous-mêmes des données de séquençage. Cette expérience aurait pu être réalisée comme indiqué en figure 5.1. De même que pour les jeux de données analysés, on pourrait infecter plusieurs larves d'un papillon sensible à AcMNPV avec ce virus. À différents pas de temps, des extractions d'ARN et d'ADN totaux seraient effectuées. Concernant l'ARN, un séquençage à la fois des longs ARN correspondant aux ARN messagers et des courts ARN impliqués dans les voies des siARN et piARN permettrait de mettre en relation l'activité des ET (détectée par l'analyse des longs ARN) avec une possible régulation de ces ET par l'intermédiaire des petits ARN (ces patrons de régulation seraient détectés par l'analyse des courts ARN). Un séquençage permettant d'obtenir des lectures plus longues serait un plus, également pour la détection de lectures chimériques ET-virus (on pourrait envisager un séquençage Illumina permettant d'obtenir des lectures de 150 pb). Un autre point important qui pourrait être soulevé au cours d'une telle expérience serait le séquençage d'ADN circulaire extrachromosomique. Des kits permettent de réaliser l'extraction de ce type d'ADN. Un séquençage de ces ADN pour obtenir des courtes (séquençage Illumina 150 pb) et des longues lectures (séquençage Pacbio Sequel II) serait alors envisagé. L'intérêt de l'étude de ces ADN circulaires serait de voir si, outre la modulation de la transcription de certains ET, l'activité de transposition serait plus importante. Ce résultat serait évidemment à mettre en perspective avec les résultats de l'expression des ET. Ainsi, cela permettrait de mieux évaluer le lien entre expression et transposition et d'avoir une vision plus complète des mécanismes à l'œuvre sur l'activité des ET au cours d'un stress provoqué par une infection virale.

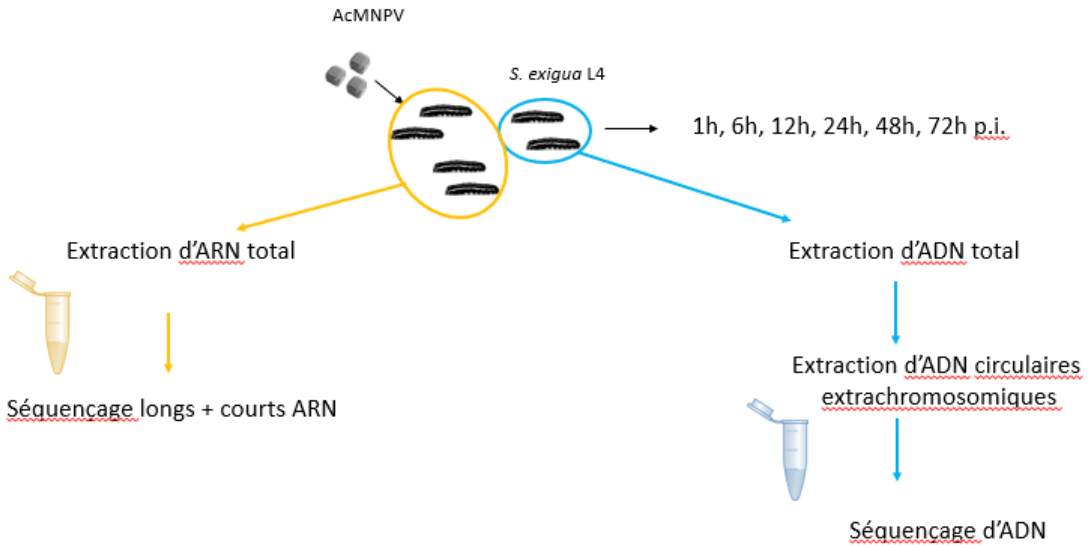



**Figure 5.1: Schéma d'une expérience permettant d'avoir une idée plus précise de l'impact d'une infection à AcMNPV sur l'activité des ET d'une espèce sensible à AcMNPV (dans**

**cet exemple, la noctuelle exiguë *Spodoptera exigua*).** Tous les types de séquençages seraient effectués à chaque pas de temps.

Un point important émergeant de ce travail de thèse concerne l'impact évolutif des insertions d'ET dans les génomes viraux au cours d'une infection. Le paradigme actuel considère la dérépression des ET, lorsque l'hôte est confronté au stress, comme probablement bénéfique grâce à la mutagenèse insertionnelle des ET et au recâblage du réseau de régulation des gènes (McClintock 1984 ; Capy et al. 2000 ; Negi, Rai, and Suprasanna 2016). Cependant, la génétique des populations affirme qu'une telle stratégie ne peut pas être viable, car la plupart des nouveaux événements d'insertion d'ET sont délétères (Le Rouzic, Boutin et Capy 2007). Dit autrement, l'avantage tiré d'une éventuelle insertion d'ET bénéfique pendant un stress est largement contrebalancé par toutes les autres insertions délétères survenant pendant et en dehors des périodes de stress, qui entraînent une diminution de la valeur sélective de l'hôte. En revanche, dans le cas d'une infection virale, il est tentant de penser que la dérépression des ET pourrait alors agir comme une arme qui augmenterait l'activité de certains ET qui pourraient transposer dans les génomes viraux et ainsi diminuer l'efficacité de réplication de ces derniers. Une telle stratégie pourrait conduire à une destruction retardée des cellules hôtes et donner plus de temps au système immunitaire de l'hôte pour entraver la réplication virale. Poussé à l'extrême, un tel mécanisme (associé aux autres mécanismes de défense antiviraux) pourrait permettre la survie de l'hôte à la suite de l'infection et aboutir à l'insertion de nombreuses séquences d'ET dans les génomes viraux, rendant les particules virales moins virulentes lors de prochaines infections. Cependant, l'hypothèse selon laquelle les insertions d'ET dans les génomes viraux lors d'une infection en condition *in vivo* retarderaient la mort de l'hôte n'a jamais été formellement testée. Comme nous l'avons vu dans le chapitre 2, les fréquences d'insertions des ET dans les virus sont généralement <10%. Il faudrait d'une part qu'il y ait suffisamment de génomes viraux portant une insertion d'ET pour que ces insertions puissent avoir un effet à l'échelle de l'hôte, et ces insertions devraient être délétères et ralentir la réplication virale. Ainsi, plus d'études sont nécessaires pour savoir si une telle stratégie pourrait être à l'œuvre. Un tel travail pourrait ouvrir des pistes complètement nouvelles dans notre compréhension des interactions hôte-virus-ET.

Le chapitre 2 reprend l'ensemble des jeux de données hôte-virus étudiés chez des invertébrés. Si des insertions d'ET ont été mises en évidence dans neuf systèmes hôte-virus, aucun ET de cloporte n'a été trouvé dans les génomes de l'iridovirus IIV31, virus infectant naturellement les

deux espèces de cloportes étudiées *Armadillidium vulgare* et *Porcellio dilatatus*. Trois réplicas étaient disponibles pour chacun des deux systèmes, soit six jeux de données, au sein desquels aucun ET n'a été trouvé inséré dans les génomes viraux. Ce résultat semblait étrange pour deux raisons. D'une part, Piegu et al. (2014) ont révélé la présence d'un ET de classe II non autonome (Miniature Inverted Transposable Element ou MITE) dans le génome consensus de IIV31. Ce MITE de 1049 pb, appelé IIV31-MITE, est un indice laissant penser que les génomes d'IIV31 pourraient être capables de porter des ET. D'autre part, l'étude du génome de l'armadille vulgaire *A. vulgare* a mis en évidence la présence de copies d'ET peu divergentes des séquences consensus de plusieurs familles d'ET identifiées par le programme RepeatMasker (Becking, Gilbert, et Cordaux 2020). Des données transcriptomiques ainsi qu'un génome assemblé de cette espèce étant disponibles, des transcrits d'ET ont pu être identifiés montrant que 5317 copies d'ET sont potentiellement transcrites. Toutes ces données suggèrent qu'au moins certaines familles d'ET sont capables de transposer actuellement dans le génome d'*A. vulgare*. L'absence de transposition d'ET dans les génomes viraux dans ces jeux de données est ainsi étonnante, d'autant que la profondeur de séquençage des génomes d'IIV31 est >110 000X. Bien que dans d'autres systèmes hôte-virus des ET ont été trouvés insérés dans les génomes viraux, dix et cinq jeux de données impliquant *Cydia pomonella* Granulovirus (CpGV) et *Agrotis segetum* Nucleopolyhedrovirus (AgseNPV), respectivement, se sont révélés négatifs quant à la présence d'ET dans les génomes viraux. Ces résultats négatifs posent la question des mécanismes déterminant l'intégration des ET dans les génomes viraux pendant l'infection. Les mécanismes d'infection et les interactions moléculaires de ces virus sont globalement mal connus, ce qui ne permet pas d'avoir une idée précise du mécanisme de transposition des ET hôtes dans les génomes viraux. Néanmoins l'analyse de deux systèmes différents impliquant le même hôte, mais des virus différents, à savoir *S. nonagrioides*-AcMNPV et *S. nonagrioides*-IIV6, laisse penser que l'absence d'intégration d'ET dans les génomes de certains virus proviendrait de mécanismes impliquant principalement le virus. En effet, des ET de la sésamie du maïs ont été détectés dans les génomes d'AcMNPV, mais pas dans ceux d'IIV6. On peut spéculer qu'un mécanisme au cours de la réplication virale chez certains virus (ici chez IIV31, mais pas chez AcMNPV), ou dans certaines conditions, entraîne une répression des ET (effets secondaires d'une compaction chromatinienne par exemple), ayant pour conséquence l'impossibilité d'intégration de ces derniers dans les génomes viraux. Pour finir sur ce point, les interactions hôte-virus sont complexes et une meilleure connaissance de ces dernières pourrait permettre d'apporter un éclairage nouveau sur ces questions.

D'autres jeux de données Illumina générés pour d'autres systèmes hôte-virus impliquant des mammifères ont été analysés au cours de cette thèse, mais la plupart n'ont pas été mentionnés ici. Ces jeux de données se sont tous révélés négatifs puisqu'aucun d'ET de l'hôte n'a été détecté dans les génomes viraux post-infection, bien que les génomes viraux aient été séquencés à des profondeurs allant de 843 X à 21000 X. L'un de ces jeux de données a été présenté dans le chapitre 3, correspondant à l'infection de cellules MRC-5 par le cytomégalovirus humain (ou HCMV). Neuf autres jeux de données ont été analysés, correspondant à cinq systèmes hôte-virus différents, à savoir homme-BK virus (1 jeu de données), homme-parvovirus B19 (PVB19, 1 jeu de données), chien de prairie (*Cynomys ludovicianus*)-Monkeypox virus (MPV, 3 jeux de données), rat de Gambie (*Cricetomys gambianus*)-MPV (2 jeux de données) et loir (*Graphiurus kelleni*)-MPV (2 jeux de données). Bien que ces différents virus soient tous des virus à ADN double brin, sauf le PVB19 (virus à ADN simple brin), on peut penser une fois encore que l'inadéquation entre les ET actifs des hôtes et le mode de réplication du virus est responsable de l'absence d'ET intégrés dans les génomes viraux. On peut également noter que les principaux ET actifs chez les mammifères sont des rétrotransposons non-LTR appelés LINE (pour Long interspersed Nuclear Elements, ET autonomes) et SINE (pour Short Interspersed Nuclear Elements, ET non autonomes). Une hypothèse possible serait que ces ET pourraient ne pas être les plus prompts à s'intégrer dans des génomes viraux. Cette hypothèse s'appuie simplement sur l'observation des résultats présentés dans le chapitre 2 dans lequel 37 ET différents ont été détectés dans des génomes viraux. Sur ces 37 ET, un seul était un SINE et aucun LINE n'a été détecté. De plus, un seul SINE (HaSE3) a été détecté à ce jour dans des génomes d'AcMNPV (Gilbert et al. 2014; Loiseau et al. 2020). Néanmoins, un biais possible pourrait être simplement que les ET actifs chez les espèces étudiées ne sont pas des rétrotransposons non-LTR.

Bien qu'évoqué dans la discussion du chapitre 3, il me paraît important de revenir sur l'absence de gène hôte intégré dans les génomes viraux d'AcMNPV et des autres baculovirus étudiés dans le chapitre 2. Il peut paraître paradoxal de détecter des ET insérés dans les génomes viraux, mais pas des gènes non-ET, alors que les génomes consensus des virus à grands génomes à ADN double brins sont porteurs de nombreux gènes provenant de métazoaires (i.e. Thézé et al. 2015), mais possèdent peu d'ET. Ce paradoxe semble trouver sa solution au regard des forces évolutives impliquées dans les interactions hôte-virus. Si les ET s'intègrent dans les génomes viraux au cours d'une infection, il est peu probable qu'ils confèrent un avantage sélectif au génome porteur. De plus, les génomes viraux étant denses en gènes et possédant peu de régions

non géniques, il apparaît improbable qu'un ET puisse se fixer par dérive génétique au sein d'un génome viral, son intégration risquant fort d'être délétère. À l'inverse, les gènes hôtes n'ont a priori pas de capacité de transposition leur permettant de s'intégrer de façon autonome dans un génome viral. En revanche, il est possible d'imaginer que la transposition d'ET puisse capturer un gène hôte en plus de la séquence de l'ET qui aille ensuite s'insérer dans un génome viral. On peut également penser qu'un gène hôte pourrait se retrouver intégré à un génome viral à la suite d'une réparation d'une cassure double-brin de l'ADN viral par recombinaison, par exemple. Un gène hôte ainsi intégré pourrait potentiellement apporter un avantage sélectif au génome porteur et se fixer dans la population virale. Un tel scénario expliquerait la différence d'insertion des gènes hôtes et des ET dans les génomes viraux individuels et consensus. Autrement dit, d'un côté les ET s'intégreraient souvent dans les génomes viraux, mais chaque nouvelle insertion atteindrait très rarement une fréquence élevée dans les populations virales et de l'autre côté, les gènes hôtes s'intégreraient très rarement dans les génomes viraux, mais atteindraient plus souvent des fréquences élevées dans les populations virales. Dans notre étude des VS dans les génomes d'une population d'AcMNPV par l'analyse de longues lectures, nous avions détecté une insertion de gènes mitochondriaux entourée à chaque extrémité de séquences virales au sein d'une seule lecture. Cette insertion comprenait deux gènes entiers codant pour des ARN de transfert (ARNt-Thréonine et ARNt-Proline) et deux gènes tronqués codant pour des sous-unités de la NADH déshydrogénase, l'un en 5' (gène ND6), l'autre en 3' (gène ND4L). Cette insertion était longue de 440 pb, au sein d'une lecture de 4 647 pb. Entre le gène ND4L tronqué en 3' et la séquence virale, une séquence de 130 pb d'origine incertaine était présente (peut-être bactérienne, une faible similarité avec des séquences bactériennes connues dans la base de données 'NR' de Genbank a été détectée). De plus, les régions virales en 5' et 3' de l'insertion n'étaient pas contiguës. La région virale en 5' de l'insertion correspondait au gène Ac-odv-ec27, alors que celle en 3' de l'insertion correspondait au gène Ac-pk-2, ces gènes étant non contigus dans le génome d'AcMNPV. Ces gènes étaient de plus tronqués par l'insertion dans la lecture. Un tel patron d'insertion nous laisse supposer une recombinaison ayant rassemblé deux séquences virales avec de l'ADN environnant, en l'occurrence des gènes mitochondriaux. Bien que cette unique insertion ne constitue pas une évidence forte, elle soutient une des hypothèses évoquées permettant l'intégration d'un gène hôte par recombinaison. L'étude de Sasani et al. (2018) a également montré que la recombinaison était un mécanisme important chez les poxvirus (virus à ADN double-brin) dans son adaptation à l'hôte. Leur étude met en avant des phénomènes de recombinaison entraînant des variations du nombre de copies d'un gène impliqué dans les interactions avec les défenses immunitaires de

l'hôte. Ces variations du nombre de copies géniques sont liées à l'apparition et l'augmentation en fréquence d'une mutation ponctuelle bénéfique de ce gène, toujours liées à des événements de recombinaison. Concernant le paradoxe mentionné plus haut, il est intéressant de noter la détection d'une insertion à haute fréquence (>26%) d'un ET appartenant à la superfamille piggybac dans des génomes d'AcMNPV (chapitre 2). Une telle augmentation en fréquence est intrigante, car c'est la seule détectée dans les jeux de données analysés et cet exemple semble aller à l'encontre des hypothèses évoquées dans ce paragraphe. Une possible raison expliquant cette augmentation de fréquence serait un bénéfice apporté par cette insertion lors de la réplication virale. Cela semble peu intuitif dans la mesure où une insertion d'ET dans un génome viral a une forte probabilité d'être délétère pour la réplication du virus. Une autre hypothèse serait une distorsion d'encapsidation des génomes viraux portant l'insertion d'ET, dont le mécanisme resterait à être étudié.

Le chapitre 3 illustre l'utilité des longues lectures dans l'analyse génomique puisqu'un tel patron d'insertion n'aurait pu être détecté dans de courtes lectures de 150 pb. En revanche, la technologie de séquençage utilisée (PacBio) ne nous a pas permis d'avoir une qualité de séquençage égale à celle des courtes séquences (Illumina), puisqu'environ 13% des nucléotides séquencés au sein des lectures étaient erronés. Non seulement le taux d'erreur était élevé, mais cette technologie génère aussi de courtes insertions et délétions (appelés indels), réduisant la qualité d'alignement de ces lectures sur des génomes. Cette difficulté a été assez importante quant à la détection des VS par différents programmes dans les génomes d'AcMNPV, basée à la fois sur l'analyse des courtes et des longues lectures. Le taux d'erreur et d'indels ne permet pas un alignement précis des longues lectures sur les génomes, contrairement aux courtes lectures. Il a donc fallu être flexible sur les positions des VS et ne pas considérer les coordonnées au nucléotide près. Le problème était alors de considérer ce qui était un même VS détecté par différents programmes, étant donné que les positions de début et fin de ces VS pouvaient être différentes entre les programmes. C'est pourquoi une étape de 'clustering' était nécessaire. La complexité des données à traiter nous a obligé à faire plusieurs étapes de clustering, comme expliqué dans la figure 3.1 du chapitre 3. Cette figure quelque peu complexe résume les multiples contraintes auxquelles nous devions faire face. Tout d'abord, comme expliqué juste avant, l'imprécision des positions des VS nous contraignait à faire un 'clustering' de ces positions pour identifier un même VS détecté par différents programmes et différentes technologies de séquençage. De plus, pour être conservatif dans notre approche, nous ne considérions que les VS détectés par au moins deux programmes et deux technologies de

séquençage différentes. Enfin, chaque VS détecté par un programme est soutenu par un certain nombre de lectures. Le nombre de VS détectés par les programmes était si élevé que les positions de ces VS se chevauchaient et les 'clusterings' risquaient d'éliminer un nombre important de VS. En jouant sur le nombre minimum de lectures soutenant un VS, nous avons pu récupérer davantage de VS, au prix d'un nombre important de 'clusterings' à réaliser. Cela nous a conduits à réaliser 308 'clusterings' sur les sorties des programmes de détection de VS des données concernant AcMNPV. Une telle approche ne nous permettait clairement pas de pouvoir être sûr de détecter l'intégralité les VS biologiques, ni d'être certains d'éliminer tous les VS artéfactuels. Le temps de calcul nécessaire pour faire tourner le script codant ces 'clusterings' est non négligeable puisqu'il fallait environ une heure pour le faire tourner sur un ordinateur de bureau. Enfin, notre approche de détection des VS uniquement à partir de courtes lectures chez les populations de HCMV, IIV6 et IIV31 n'est pas idéale. Au regard des VS détectés chez AcMNPV, 4,98% de ceux détectés avec les courtes lectures l'étaient aussi avec les longues lectures. Cette proportion a donc été utilisée pour déterminer le nombre de VS à considérer dans ces autres populations virales. Mais ces autres jeux de données ne correspondaient pas à la même profondeur de séquençage du génome viral, génome qui n'avait pas non plus la même taille ni la même architecture génétique que le génome d'AcMNPV. On peut penser que ces paramètres influent de façon importante sur la détection des VS par les programmes, rendant cette proportion de 4,98% assez approximative. De plus, ces 4,98% de VS ne correspondaient pas aux 4,98% de VS les plus fréquents détectés par les courtes séquences. À partir de ce constat, comment décider quels VS garder ? Nous avons fait le choix de garder les 4,98% de VS les plus fréquents, en supposant qu'ils avaient moins de chances d'être artéfactuels s'ils étaient soutenus par un nombre important de lectures. De plus, des filtres ont été utilisés pour minimiser le nombre de VS artéfactuels. Pour améliorer ce type d'analyse, je pense qu'il est judicieux, comme nous l'avons fait pour AcMNPV, d'utiliser à la fois des technologies de séquençage basées sur les courtes et les longues lectures, car elles ne souffrent pas des mêmes biais. D'autant plus qu'aujourd'hui la technologie de séquençage Sequel II de PacBio promet une qualité des lectures >99% et une longueur accrue de ces dernières. Il est possible que ce type de longues lectures associé aux courtes lectures permettent d'être plus précis dans les positions des VS détectés. Cela permettrait aussi de réduire le nombre de VS artéfactuels dans les données et de réduire le nombre de 'clusterings' nécessaires pour récupérer un maximum de VS, réduisant ainsi la complexité du script.

Enfin, une expérience d'évolution expérimentale était censée être effectuée au cours de cette thèse. Elle avait pour but de démontrer formellement un TH d'ET entre deux espèces de papillons par l'intermédiaire d'un virus (cf figure 0.6). Le virus choisi était AcMNPV et les deux hôtes étaient le sphinx du tabac (*Manduca sexta*) et la sésamie du maïs (*Sesamia nonagrioides*), tous les deux sensibles au virus. *M. sexta* est cependant connue pour être assez résistante à l'infection. Un élevage de sésamies du maïs était déjà présent au laboratoire, et des infections à AcMNPV de larves au troisième stade larvaire avaient été réalisées dans le cadre de l'étude présentée dans le chapitre 2. La technique d'infection était elle aussi maîtrisée. Le point inconnu de cette expérience était la mise en place d'un élevage de sphinx du tabac. Pour ce faire, des moyens logistiques étaient déjà présents au laboratoire ou ont été acquis pour cet élevage (étuve avec photopériode et température réglables, boîtes d'élevage de papillons, plants de tabac pour la ponte d'œufs, veilleuse imitant un clair de lune pour faciliter la reproduction). Malheureusement, les différentes tentatives d'élevage de *M. sexta* se sont toutes soldées par un échec, soit parce que presque toutes les larves arrivaient mortes au laboratoire, soit parce que notre protocole d'élevage n'a pas permis la survie des larves ou des pupes. Ces échecs illustrent à quel point il n'est pas aisé de mettre rapidement en place un élevage pour les besoins d'une expérience. L'idée était dans un premier temps d'infecter des larves de sésamie du maïs avec AcMNPV. Après infection, une partie des particules virales aurait été utilisée pour infecter les larves de sphinx du tabac, suffisamment résistantes pour survivre à l'infection et atteindre le stage imago. Des croisements d'adultes infectés auraient été effectués dans les cages à papillons pour leur reproduction. Les larves issues des accouplements et leurs parents auraient été testés par PCR puis vérifiés par séquençage Sanger pour rechercher la présence d'ET de la sésamie, préalablement caractérisés (chapitre 2). Ces tests auraient pu être effectués sur le corps entier des individus, sur les tissus somatiques et sur les tissus de la lignée germinale, séparément. Les ET auraient également été recherchés dans la population d'AcMNPV avant infection de *M. sexta*. Ces ET ne doivent bien sûr pas être naturellement présents dans le génome du sphinx du tabac. Certaines étapes techniques auraient nécessité une mise au point, notamment la détermination de la charge virale maximale non létale pouvant être utilisée sur les chenilles de *M. sexta*, selon le stade larvaire choisi pour l'infection. Il aurait été théoriquement possible de détecter la présence d'ET de *S. nonagrioides* dans tous les tissus testés, chez les individus infectés comme chez leurs descendants. Étant donné la prévalence des cellules somatiques comparée aux cellules germinales, la probabilité d'insertion dans des cellules germinales aurait été faible, mais non nulle, permettant alors à la descendance des individus de *M. sexta* infectés par AcMNPV de porter dans leur génome un ou plusieurs ET de la sésamie du maïs. Une étude

réalisée par Yamao et al. (1999), a démontré la faisabilité de ce genre d'expérience. Les auteurs ont cloné le gène codant la fibroïne du bombyx du mûrier (*Bombyx mori*) et inséré le gène codant la protéine fluorescente verte (ou GFP pour Green Fluorescent Protein) dans un des exons. Ce gène chimérique a ensuite été inséré dans des génomes d'AcMNPV en remplacement du gène codant la polyhédrine. Ce virus recombinant (qui correspondrait à la population virale portant des ET de *S. nonagrioides* dans notre expérience) a été injecté dans des larves femelles du bombyx du mûrier (cinquième stage larvaire) à une dose non létale, des croisements ont été effectués avec des mâles non infectés et les descendances (F1 et F2) ont été testées pour rechercher le gène de la GFP. Les auteurs ont retrouvé le gène chimérique dans le génome des descendants des deux générations et avancent que cette intégration du virus au génome des individus infectés s'est réalisée par recombinaison homologue avec le gène de la fibroïne naturellement présent dans le génome du bombyx du mûrier. Il est à noter que toutes les portées issues des femelles infectées n'étaient pas porteuses du gène de la GFP, seules 3% de ces portées étaient positives. Au sein des portées positives, 3% des œufs testés étaient positifs, soit 0,09% de l'ensemble des descendants des femelles infectées étaient porteurs du gène de la GFP. Malgré la faible proportion des descendants positifs, cette expérience est une preuve de principe que des transferts de gènes peuvent se produire entre AcMNPV et ses hôtes. Ces résultats nous encouragent à penser que notre expérience aurait pu aboutir à un résultat positif puisque l'ensemble des individus (mâles et femelles) auraient été infectés par AcMNPV et parce que des transferts d'ADN effectués par transposition pourraient se produire plus fréquemment que ceux effectués par recombinaison homologue. L'étude présentée dans le chapitre 2 de cette thèse a révélé que plus d'un quart des génomes d'AcMNPV pouvaient porter un ET, ce qui constitue une portion non négligeable de génomes viraux pouvant potentiellement infecter des cellules germinales d'un futur hôte.

Si l'expérience présentée ici devait être de nouveau tentée, il serait préférable d'utiliser comme espèce receveuse non pas le sphinx du tabac, mais un autre lépidoptère pouvant être facilement élevé au laboratoire, comme le foreur ponctué de graminées (*Chilo partellus*; milieu nutritif et conditions d'élevage similaires à ceux de la sésamie du maïs) et qui a déjà été étudié (cf. chapitre 2). Des infections à AcMNPV n'ont jamais été effectuées sur ce papillon, mais il est possible, étant un lépidoptère, qu'il soit sensible à ce virus. Des tests PCR devraient être réalisés sur des individus de *C. partellus* pour s'assurer que les ET recherchés par la suite ne sont pas naturellement présents dans le génome de cette espèce. Un tel système (sésamie du maïs-AcMNPV-lépidoptère) serait vraiment à considérer, car ce virus est celui pour lequel le plus de

données sont disponibles concernant les insertions d'ET dans son génome. De plus, la seule population d'AcMNPV possédant plus de 25% de génomes portant un ET a été obtenue après infection de larves de la sésamie du maïs. Enfin, d'autres systèmes basés sur IIV6 comme vecteur d'ET méritent d'être étudiés. Les résultats présentés au chapitre 2 révèlent que IIV6 peut infecter différentes espèces de drosophiles (adultes de *D. melanogaster* et *D. hydei*), ainsi que différentes espèces de lépidoptères (larves de *S. nonagrioides* et *C. partellus*) élevées au laboratoire. Là encore, il faudrait déterminer la charge virale maximale non létale à utiliser en fonction de l'espèce infectée et s'assurer que le tropisme cellulaire du virus inclut les gonades (notamment dans le cas des infections de drosophiles adultes dont les gonades sont déjà différenciées). L'inconvénient d'IIV6 réside dans la faible proportion de génomes viraux portant un ET (jusqu'à 6,5% après infection de *D. melanogaster*) en comparaison d'AcMNPV. De manière générale, il y a à ce jour assez peu de recul sur l'étude de ces différents systèmes : est-ce que les individus vont survivre à l'infection ? S'ils survivent, restera-t-il suffisamment de particules virales pour augmenter la probabilité de transposition virus-hôte des ET ? Les individus survivants à l'infection se reproduiront-ils ? Ces diverses questions restent sans réponse claire dans le cadre de cette thèse. Il faudrait également disposer de données plus fournies concernant les différents systèmes hôte-virus : peu de réplicas réalisés, compréhension incomplète des phénomènes de transposition hôte-virus et virus-hôte (cf. chapitre 2).

Ainsi, les projets initiés au cours de cette thèse auront permis de décortiquer différentes étapes d'un TH d'ET réalisé par l'intermédiaire d'un virus, d'apporter des résultats soutenant le rôle des virus comme vecteurs d'ET entre insectes et de soulever peut-être plus de questions encore, restant pour l'instant sans réponses précises et quantifiables.

# Bibliographie

# *Bibliographie*


Acevedo, A., and Andino, R. 2014. Library preparation for highly accurate population sequencing of RNA viruses. Nat. Protoc. *9*, 1760–1769.

Acevedo, A., Brodsky, L., and Andino, R. 2014. Mutational and fitness landscapes of an RNA virus revealed through population sequencing. Nature *505*, 686–690.

Ackermann, H.-W., and Smirnoff, W.A. 1983. A morphological investigation of 23 baculoviruses. J. Invertebr. Pathol. *41*, 269–280.

Adey, A. *et al.* 2013. The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature* **500**, 207–211.

Akagi, K. *et al.* 2014. Genome-wide analysis of HPV integration in human cancers reveals recurrent, focal genomic instability. *Genome Res.* **24**, 185–199.

Akhtar, L.N., Bowen, C.D., Renner, D.W., Pandey, U., Della Fera, A.N., Kimberlin, D.W., Prichard, M.N., Whitley, R.J., Weitzman, M.D., and Szpara, M.L. 2019. Genotypic and Phenotypic Diversity of Herpes Simplex Virus 2 within the Infected Neonatal Population. MSphere *4*.

Akiba, T., K. Koyama, Y. Ishiki, S. Kimura, and T. Fukushima. 1960. On the mechanism of the development of multiple-drug-resistant clones of Shigella. Jpn J Microbiol **4**:219-227.

Akkouche, A., R. Rebollo, N. Burlet, C. Esnault, S. Martinez, B. Viginier, C. Terzian, C. Vieira, et M. Fablet. 2012. « Tirant, a Newly Discovered Active Endogenous Retrovirus in Drosophila Simulans ». *Journal of Virology* 86 (7): 3675-81. https://doi.org/10.1128/JVI.07146-11.

Alkan, C., Coe, B.P., and Eichler, E.E. 2011. Genome structural variation discovery and genotyping. Nat. Rev. Genet. *12*, 363–376.

Alletti, G., Sauer, A., Weihrauch, B., Fritsch, E., Undorf-Spahn, K., Wennmann, J., Jehle, J., 2017. Using Next Generation Sequencing to Identify and Quantify the Genetic Composition of Resistance-Breaking Commercial Isolates of Cydia pomonella Granulovirus. Viruses 9, 250. https://doi.org/10.3390/v9090250

Angly, F.E., Willner, D., Rohwer, F., Hugenholtz, P., and Tyson, G.W. 2012. Grinder: a versatile amplicon and shotgun sequence simulator. Nucleic Acids Res. *40*, e94–e94.

Anwar, Sumadi, Wahyu Wulaningsih, and Ulrich Lehmann. 2017. "Transposable Elements in Human Cancer: Causes and Consequences of Deregulation." *International Journal of Molecular Sciences* 18 (5): 974. https://doi.org/10.3390/ijms18050974.

Arita, Shuji, Eishi Baba, Yoshihiro Shibata, Hiroaki Niiro, Shinji Shimoda, Taichi Isobe, Hitoshi Kusaba, Shuji Nakano, et Mine Harada. 2008. « B Cell Activation Regulates Exosomal HLA Production ». *European Journal of Immunology* 38 (5): 1423-34. https://doi.org/10.1002/eji.200737694.

Arkhipova, Irina R. 2018. « Neutral Theory, Transposable Elements, and Eukaryotic Genome Evolution ». Édité par Sudhir Kumar. *Molecular Biology and Evolution* 35 (6): 1332-37. https://doi.org/10.1093/molbev/msy083.

Balaj, Leonora, Ryan Lessard, Lixin Dai, Yoon-Jae Cho, Scott L. Pomeroy, Xandra O. Breakefield, et Johan Skog. 2011. « Tumour Microvesicles Contain Retrotransposon Elements and Amplified Oncogene Sequences ». *Nature Communications* 2 (1): 180. https://doi.org/10.1038/ncomms1180.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., et al. 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J. Comput. Biol. *19*, 455–477.

Bao, W., Kojima, K.K., and Kohany, O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob. DNA *6*, 11. https://doi.org/10.1186/s13100-015-0041-9

Barah, Pankaj, Naresh D. Jayavelu, John Mundy, and Atle M. Bones. 2013. “Genome Scale Transcriptional Response Diversity among Ten Ecotypes of Arabidopsis Thaliana during Heat Stress.” *Frontiers in Plant Science* 4. https://doi.org/10.3389/fpls.2013.00532.

Barrett, John W., Andy J. Brownwright, Mark J. Primavera, and Subba Reddy Palli. 1998. “Studies of the Nucleopolyhedrovirus Infection Process in Insects by Using the Green Fluorescence Protein as a Reporter.” *Journal of Virology* 72 (4): 3377–82. https://doi.org/10.1128/JVI.72.4.3377-3382.1998.

Barrón, Maite G., Anna-Sophie Fiston-Lavier, Dmitri A. Petrov, and Josefa González. 2014. “Population Genomics of Transposable Elements in *Drosophila*.” *Annual Review of Genetics* 48 (1): 561–81. https://doi.org/10.1146/annurev-genet-120213-092359.

Bartolomé, C., Bello, X., Maside, X., 2009. Widespread evidence for horizontal transfer of transposable elements across Drosophila genomes. Genome Biol. 10, R22. https://doi.org/10.1186/gb-2009-10-2-r22

Battaglia, R., S. Palini, M. E. Vento, A. La Ferlita, M. J. Lo Faro, E. Caroppo, P. Borzì, et al. 2019. « Identification of Extracellular Vesicles and Characterization of MiRNA Expression Profiles in Human Blastocoel Fluid ». *Scientific Reports* 9 (1): 84. https://doi.org/10.1038/s41598-018-36452-7.

Bauser, Christopher A., Teresa A. Elick, and M.J. Fraser. 1996. “Characterization Ofhitchhiker,a Transposon Insertion Frequently Associated with Baculovirus FP Mutants Derived upon Passage in the TN-368 Cell Line.” *Virology* 216 (1): 235–37. https://doi.org/10.1006/viro.1996.0053.

Becking, T., Gilbert, C., Cordaux, R., 2020. Impact of transposable elements on genome size variation between two closely related crustacean species. Anal. Biochem. 600, 113770. https://doi.org/10.1016/j.ab.2020.113770

Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discoveryrate: a practical and powerful approach to multiple testing,Journal ofthe Royal Statistical Society Series B.85: 289–300.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Wheeler, D.L. 2005. GenBank. Nucleic Acids Res. *33*, D34–D38.

Bézier, Annie, Juline Herbinière, Beatrice Lanzrein, et Jean-Michel Drezen. 2009. « Polydnavirus Hidden Face: The Genes Producing Virus Particles of Parasitic Wasps ». *Journal of Invertebrate Pathology* 101 (3): 194-203. https://doi.org/10.1016/j.jip.2009.04.006.

Biémont, Christian, et Cristina Vieira. 2006. « Junk DNA as an Evolutionary Force ». *Nature* 443 (7111): 521-24. https://doi.org/10.1038/443521a.

Blumenstiel, Justin P. 2019. “Birth, School, Work, Death, and Resurrection: The Life Stages and Dynamics of Transposable Element Proliferation.” *Genes* 10 (5): 336. https://doi.org/10.3390/genes10050336.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. “Trimmomatic: A Flexible Trimmer for Illumina Sequence Data.” *Bioinformatics* 30 (15): 2114–20. https://doi.org/10.1093/bioinformatics/btu170.

Bouallaga, I., Massicard, S., Yaniv, M. & Thierry, F. 2000. An enhanceosome containing the Jun B/Fra-2 heterodimer and the HMG-I(Y) architectural protein controls HPV18 transcription. *EMBO Rep.* **1**, 422–427.

Bouallègue, Maryem, Jacques-Deric Rouault, Aurélie Hua-Van, Mohamed Makni, and Pierre Capy. 2017. “Molecular Evolution of *PiggyBac* Superfamily: From Selfishness to Domestication.” *Genome Biology and Evolution*, January, evw292. https://doi.org/10.1093/gbe/evw292.

Bourque, G., Burns, K.H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H.L., Macfarlan, T.S., Mager, D.L., Feschotte, C., 2018. Ten things you should know about transposable elements. Genome Biol. 19, 199. https://doi.org/10.1186/s13059-018-1577-z

Brookfield, John F. Y., and Richard M. Badge. 1997. “Population Genetics Models of Transposable Elements.” In *Evolution and Impact of Transposable Elements*, edited by Pierre Capy, 6:281–94. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-011-4898-6_28.

Bucher, Etienne, Jon Reinders, and Marie Mirouze. 2012. “Epigenetic Control of Transposon Transcription and Mobility in Arabidopsis.” *Current Opinion in Plant Biology* 15 (5): 503–10. https://doi.org/10.1016/j.pbi.2012.08.006.

Buchfink, Benjamin, Chao Xie, and Daniel H Huson. 2015. “Fast and Sensitive Protein Alignment Using DIAMOND.” *Nature Methods* 12 (1): 59–60. https://doi.org/10.1038/nmeth.3176.

Bull, J.C., Godfray, H.C.J., and O’Reilly, D.R. 2003. A Few-Polyhedra Mutant and Wild-Type Nucleopolyhedrovirus Remain as a Stable Polymorphism during Serial Coinfection in Trichoplusia ni. Appl. Environ. Microbiol. *69*, 2052–2057.

Burke, Gaelen R., Sarah A. Thomas, Jai H. Eum, et Michael R. Strand. 2013. « Mutualistic Polydnaviruses Share Essential Replication Gene Functions with Pathogenic Ancestors ». Édité par David S. Schneider. *PLoS Pathogens* 9 (5): e1003348. https://doi.org/10.1371/journal.ppat.1003348.

Burns, Kathleen H. 2017. “Transposable Elements in Cancer.” *Nature Reviews Cancer* 17 (7): 415–24. https://doi.org/10.1038/nrc.2017.35.

Cadiñanos, Juan, and Allan Bradley. 2007. “Generation of an Inducible and Optimized PiggyBac Transposon System†.” *Nucleic Acids Research* 35 (12): e87. https://doi.org/10.1093/nar/gkm446.

Cai, Jin, Gengze Wu, Pedro A. Jose, et Chunyu Zeng. 2016. « Functional Transferred DNA within Extracellular Vesicles ». *Experimental Cell Research* 349 (1): 179-83. https://doi.org/10.1016/j.yexcr.2016.10.012.

Cai, Jin, Gengze Wu, Xiaorong Tan, Yu Han, Caiyu Chen, Chuanwei Li, Na Wang, et al. 2014. « Transferred BCR/ABL DNA from K562 Extracellular Vesicles Causes Chronic Myeloid Leukemia in Immunodeficient Mice ». Édité par Sonja Loges. *PLoS ONE* 9 (8): e105200. https://doi.org/10.1371/journal.pone.0105200.

Cai, Jin, Weiwei Guan, Xiaorong Tan, Caiyu Chen, Liangpeng Li, Na Wang, Xue Zou, et al. 2015. « SRY Gene Transferred by Extracellular Vesicles Accelerates Atherosclerosis by Promotion of Leucocyte Adherence to Endothelial Cells ». *Clinical Science* 129 (3): 259-69. https://doi.org/10.1042/CS20140826.

Cai, Jin, Yu Han, Hongmei Ren, Caiyu Chen, Duofen He, Lin Zhou, Gilbert M. Eisner, Laureano D. Asico, Pedro A. Jose, et Chunyu Zeng. 2013. « Extracellular Vesicle-Mediated Transfer of Donor Genomic DNA to Recipient Cells Is a Novel Mechanism for Genetic Influence between Cells ». *Journal of Molecular Cell Biology* 5 (4): 227-38. https://doi.org/10.1093/jmcb/mjt011.

Cantalupo, P. G., Katz, J. P. & Pipas, J. M. 2015. HeLa Nucleic Acid Contamination in The Cancer Genome Atlas Leads to the Misidentification of Human Papillomavirus 18. *J. Virol.* **89**, 4051–4057.

Cao, S. *et al.* 2015. High-Throughput RNA Sequencing-Based Virome Analysis of 50 Lymphoma Cell Lines from the Cancer Cell Line Encyclopedia Project. *J. Virol.* **89**, 713–729.

Capes-Davis, A. *et al.* 2010. Check your cultures! A list of cross-contaminated or misidentified cell lines. *Int. J. Cancer* **127**, 1–8.

Capy, Pierre, Giuliano Gasperi, Christian Biémont, and Claude Bazin. 2000. “Stress and Transposable Elements: Co-Evolution or Useful Parasites?” *Heredity* 85 (2): 101–6. https://doi.org/10.1046/j.1365-2540.2000.00751.x.

Carstens, E.B., 1987. Identification and nucleotide sequence of the regions of Autographa californica nuclear polyhedrosis virus genome carrying insertion elements derived from Spodoptera frugiperda. Virology 161, 8–17. https://doi.org/10.1016/0042-6822(87)90165-6

Casacuberta, Elena, and Josefa González. 2013. “The Impact of Transposable Elements in Environmental Adaptation.” *Molecular Ecology* 22 (6): 1503–17. https://doi.org/10.1111/mec.12170.

Chaisson, M.J., and Tesler, G. 2012. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics *13*.

Charif, D., Thioulouse, J., Lobry, J.R., Perrire, G., 2005. Online synonymous codon usage analyses with the ade4 and seqinR packages. Bioinformatics 21, 545–547. https://doi.org/10.1093/bioinformatics/bti037

Charlesworth, B., P. Sniegowski, and W. Stephan. 1994. The evolutionary dynamics of repetitive DNA in eukaryotes. Nature **371**:215-220.

Chateigner, A., Bézier, A., Labrousse, C., Jiolle, D., Barbe, V., and Herniou, E. 2015. Ultra Deep Sequencing of a Baculovirus Population Reveals Widespread Genomic Variations. Viruses *7*, 3625–3646.

Chebbi, M.A., Becking, T., Moumen, B., Giraud, I., Gilbert, C., Peccoud, J., and Cordaux, R. 2019. The Genome of *Armadillidium vulgare* (Crustacea, Isopoda) Provides Insights into Sex Chromosome Evolution in the Context of Cytoplasmic Sex Determination. Mol. Biol. Evol. *36*, 727–741.

Chen, W., Yang, X., Tetreau, G., Song, X., Coutu, C., Hegedus, D., Blissard, G., Fei, Z., Wang, P., 2019. A high-quality chromosome-level genome assembly of a generalist herbivore, *Trichoplusia ni*. Mol. Ecol. Resour. 19, 485–496. https://doi.org/10.1111/1755-0998.12966

Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, A.J., Kruglyak, S., and Saunders, C.T. 2016. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics *32*, 1220–1222.

Chen, Y.-R., S. Zhong, Z. Fei, S. Gao, S. Zhang, Z. Li, P. Wang, et G. W. Blissard. 2014. « Transcriptome Responses of the Host Trichoplusia Ni to Infection by the Baculovirus Autographa Californica Multiple Nucleopolyhedrovirus ». *Journal of Virology* 88 (23): 13781-97. https://doi.org/10.1128/JVI.02243-14.

Chen, Y.-R., S. Zhong, Z. Fei, Y. Hashimoto, J. Z. Xiang, S. Zhang, and G. W. Blissard. 2013. "The Transcriptome of the Baculovirus Autographa Californica Multiple Nucleopolyhedrovirus in Trichoplusia Ni Cells." *Journal of Virology* 87 (11): 6391–6405. https://doi.org/10.1128/JVI.00194-13.

Chénais, Benoît, Aurore Caruso, Sophie Hiard, et Nathalie Casse. 2012. « The Impact of Transposable Elements on Eukaryotic Genomes: From Genome Size Increase to Genetic Adaptation to Stressful Environments ». *Gene* 509 (1): 7-15. https://doi.org/10.1016/j.gene.2012.07.042.

Chenais, Benoit. 2015. "Transposable Elements in Cancer and Other Human Diseases." *Current Cancer Drug Targets* 15 (3): 227–42. https://doi.org/10.2174/1568009615666150317122506.

Chevignon, Germain, Georges Periquet, Gabor Gyapay, Nathalie Vega-Czarny, Karine Musset, Jean-Michel Drezen, et Elisabeth Huguet. 2018. « Cotesia Congregata Bracovirus Circles Encoding *PTP* and *Ankyrin* Genes Integrate into the DNA of Parasitized *Manduca Sexta* Hemocytes ». Édité par Jae U. Jung. *Journal of Virology* 92 (15): e00438-18, /jvi/92/15/e00438-18.atom. https://doi.org/10.1128/JVI.00438-18.

Chuong, Edward B., Nels C. Elde, et Cédric Feschotte. 2017. « Regulatory Activities of Transposable Elements: From Conflicts to Benefits ». *Nature Reviews Genetics* 18 (2): 71-86. https://doi.org/10.1038/nrg.2016.139.

Cmarik, J. L., Troxler, J. A., Hanson, C. A., Zhang, X. & Ruscetti, S. K. 2011. The Human Lung Adenocarcinoma Cell Line EKVX Produces an Infectious Xenotropic Murine Leukemia Virus. *Viruses* **3**, 2442–2461.

Coakley, Gillian, Rick M. Maizels, et Amy H. Buck. 2015. « Exosomes and Other Extracellular Vesicles: The New Communicators in Parasite Infections ». *Trends in Parasitology* 31 (10): 477-89. https://doi.org/10.1016/j.pt.2015.06.009.

Cocucci, Emanuele, Gabriella Racchetti, et Jacopo Meldolesi. 2009. « Shedding Microvesicles: Artefacts No More ». *Trends in Cell Biology* 19 (2): 43-51. https://doi.org/10.1016/j.tcb.2008.11.003.

Cordaux, R., Batzer, M.A., 2009. The impact of retrotransposons on human genome evolution. Nat. Rev. Genet. 10, 691–703. https://doi.org/10.1038/nrg2640

Cordaux, R., and C. Gilbert. 2017. « Evolutionary Significance of Wolbachia-to-Animal Horizontal Gene Transfer: Female Sex Determination and the f Element in the Isopod Armadillidium Vulgare ». *Genes* 8 (7): 186. https://doi.org/10.3390/genes8070186.

Cordaux, Richard, et Mark A. Batzer. 2009. « The Impact of Retrotransposons on Human Genome Evolution ». *Nature Reviews Genetics* 10 (10): 691-703. https://doi.org/10.1038/nrg2640.

Correa, Ricardo, Zuleima Caballero, Luis F. De León, et Carmenza Spadafora. 2020. « Extracellular Vesicles Could Carry an Evolutionary Footprint in Interkingdom Communication ». *Frontiers*

*in Cellular and Infection Microbiology* 10 (mars): 76. https://doi.org/10.3389/fcimb.2020.00076.

Cosby, R.L., Chang, N.-C., and Feschotte, C. 2019. Host–transposon interactions: conflict, cooperation, and cooption. Genes Dev. *33*, 1098–1116.

Cowley, Michael, and Rebecca J. Oakey. 2013. "Transposable Elements Re-Wire and Fine-Tune the Transcriptome." Edited by Elizabeth M. C. Fisher. *PLoS Genetics* 9 (1): e1003234. https://doi.org/10.1371/journal.pgen.1003234.

Craig NL, C. R., Gellert M, Lambowitz AM. 2002. Mobile DNA II. Washington (DC): American Society for Microbiology Press.

Craig, N.L., Chandler, M., Gellert, M., Lambowitz, A.M., Rice, P.A., Sandmeyer, S.B. (Eds.), 2015. Mobile DNA III. ASM Press, Washington, DC, USA. https://doi.org/10.1128/9781555819217

Crouch, E.A., and Passarelli, A.L. 2002. Genetic Requirements for Homologous Recombination in Autographa californica Nucleopolyhedrovirus. J. Virol. *76*, 9323–9334.

Cruz, Fernando de la, et Julian Davies. 2000. « Horizontal Gene Transfer and the Origin of Species: Lessons from Bacteria ». *Trends in Microbiology* 8 (3): 128-33. https://doi.org/10.1016/S0966-842X(00)01703-0.

Cudini, J., Roy, S., Houldcroft, C.J., Bryant, J.M., Depledge, D.P., Tutill, H., Veys, P., Williams, R., Worth, A.J.J., Tamuri, A.U., et al. 2019. Human cytomegalovirus haplotype reconstruction reveals high diversity due to superinfection and evidence of within-host recombination. Proc. Natl. Acad. Sci. *116*, 5693–5698.

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, 339–355.

Darwin, C. 1859. The Origin of Species by Means of Natural Selection. Murray, London, U.K.

De Gooijer, C.D., Koken, R.H.M., Van Lier, F.L.J., Kool, M., Vlak, J.M., and Tramper, J. 1992. A structured dynamic model for the baculovirus infection process in insect-cell reactor configurations. Biotechnol. Bioeng. *40*, 537–548.

De Rijck, J. *et al.* 2013. The BET Family of Proteins Targets Moloney Murine Leukemia Virus Integration near Transcription Start Sites. *Cell Rep.* **5**, 886–894.

Deniz, Özgen, Jennifer M. Frost, and Miguel R. Branco. 2019. "Regulation of Transposable Elements by DNA Modifications." *Nature Reviews Genetics*, March. https://doi.org/10.1038/s41576-019-0106-6.

Descartes, R., 1637. 'Discours de la méthode pour bien conduire sa raison et chercher la vérité dans les sciences, plus la dioptrique, les météores et la géométrie.*' 537 p.*

Dong, George, Alonso Lira Filho, et Martin Olivier. 2019. « Modulation of Host-Pathogen Communication by Extracellular Vesicles (EVs) of the Protozoan Parasite Leishmania ». *Frontiers in Cellular and Infection Microbiology* 9 (avril): 100. https://doi.org/10.3389/fcimb.2019.00100.

Dotto, B.R., Carvalho, E.L., da Silva, A.F., Dezordi, F.Z., Pinto, P.M., Campos, T. de L., Rezende, A.M., Wallau, G. da L., 2018. HTT-DB: new features and updates. Database 2018. https://doi.org/10.1093/database/bax102

Drexler, H. G. & Uphoff, C. C. 2002. Mycoplasma contamination of cell cultures: Incidence, sources, effects, detection, elimination, prevention. *Cytotechnology* **39**, 75–90.

Drezen, Jean-Michel, Matthieu Leobold, Annie Bézier, Elisabeth Huguet, Anne-Nathalie Volkoff, et Elisabeth A Herniou. 2017. « Endogenous Viruses of Parasitic Wasps: Variations on a Common Theme ». *Current Opinion in Virology* 25 (août): 41-48. https://doi.org/10.1016/j.coviro.2017.07.002.

Dubin, Manu J, Ortrun Mittelsten Scheid, and Claude Becker. 2018. "Transposons: A Blessing Curse." *Current Opinion in Plant Biology* 42 (April): 23–29. https://doi.org/10.1016/j.pbi.2018.01.003.

Duffy, S., Shackelton, L.A., and Holmes, E.C. 2008. Rates of evolutionary change in viruses: patterns and determinants. Nat. Rev. Genet. *9*, 267–276.

Dupressoir A., C. Vernochet, O. Bawa, F. Harper, G. Pierron, P. Opolon et T. Heidmann. 2009. "Syncytin-A knockout mice demonstrate the critical role in placentation of a fusogenic, endogenous retrovirus-derived, envelope gene A." *Proc. Natl. Acad. Sci.* doi:10.1073/pnas.0902925106

Eckwahl, M.J., Telesnitsky, A., Wolin, S.L., 2016. Host RNA Packaging by Retroviruses: A Newly Synthesized Story. mBio 7, e02025-02015. https://doi.org/10.1128/mBio.02025-15

Eichenberger, Ramon M., Md Hasanuzzaman Talukder, Matthew A. Field, Phurpa Wangchuk, Paul Giacomin, Alex Loukas, et Javier Sotillo. 2018. « Characterization of *Trichuris Muris* Secreted Proteins and Extracellular Vesicles Provides New Insights into Host–Parasite Communication ». *Journal of Extracellular Vesicles* 7 (1): 1428004. https://doi.org/10.1080/20013078.2018.1428004.

Elde, N.C., Child, S.J., Eickbush, M.T., Kitzman, J.O., Rogers, K.S., Shendure, J., Geballe, A.P., and Malik, H.S. 2012. Poxviruses Deploy Genomic Accordions to Adapt Rapidly against Host Antiviral Defenses. Cell *150*, 831–841.

Eme, Laura, Eleni Gentekaki, Bruce Curtis, John M. Archibald, et Andrew J. Roger. 2017. « Lateral Gene Transfer in the Adaptation of the Anaerobic Parasite Blastocystis to the Gut ». *Current Biology* 27 (6): 807-20. https://doi.org/10.1016/j.cub.2017.02.003.

Engelhard, E. K., L. N. Kam-Morgan, J. O. Washburn, and L. E. Volkman. 1994. "The Insect Tracheal System: A Conduit for the Systemic Spread of Autographa Californica M Nuclear Polyhedrosis Virus." *Proceedings of the National Academy of Sciences* 91 (8): 3224–27. https://doi.org/10.1073/pnas.91.8.3224.

English, A.C., Salerno, W.J., and Reid, J.G. 2014. PBHoney: identifying genomic variants via long-read discordance and interrupted mapping. BMC Bioinformatics *15*.

Eves-van den Akker, Sebastian, Dominik R. Laetsch, Peter Thorpe, Catherine J. Lilley, Etienne G. J. Danchin, Martine Da Rocha, Corinne Rancurel, et al. 2016. « The Genome of the Yellow Potato Cyst Nematode, Globodera Rostochiensis, Reveals Insights into the Basis of Parasitism and Virulence ». *Genome Biology* 17 (1): 124. https://doi.org/10.1186/s13059-016-0985-1.

Fablet, Marie, Angelo Jacquet, Rita Rebollo, Annabelle Haudry, Carine Rey, Judit Salces-Ortiz, Prajakta Bajad, et al. 2019. « Dynamic Interactions Between the Genome and an Endogenous Retrovirus: *Tirant* in *Drosophila Simulans* Wild-Type Strains ». *G3; Genes|Genomes|Genetics*, janvier, g3.200789.2018. https://doi.org/10.1534/g3.118.200789.

Fan, J., Wennmann, J.T., Wang, D., Jehle, J.A., 2019. Novel Diversity and Virulence Patterns Found in New Isolates of Cydia pomonella Granulovirus from China. Appl. Environ. Microbiol. 86, e02000-19, /aem/86/2/AEM.02000-19.atom. https://doi.org/10.1128/AEM.02000-19

Fan, J., Wennmann, J.T., Wang, D., Jehle, J.A., 2020. Single nucleotide polymorphism (SNP) frequencies and distribution reveal complex genetic composition of seven novel natural isolates of Cydia pomonella granulovirus. Virology 541, 32–40. https://doi.org/10.1016/j.virol.2019.11.016

Feschotte, Cédric, et Ellen J. Pritham. 2007. « DNA Transposons and the Evolution of Eukaryotic Genomes ». *Annual Review of Genetics* 41 (1): 331-68. https://doi.org/10.1146/annurev.genet.40.110405.090448.

Filée, J. 2013. Route of NCLDV evolution: the genomic accordion. Curr. Opin. Virol. *3*, 595–599.

Filée, J. 2015. Genomic comparison of closely related Giant Viruses supports an accordion-like model of evolution. Front. Microbiol. *6*.

Filée, J., 2018. Giant viruses and their mobile genetic elements: the molecular symbiosis hypothesis. Curr. Opin. Virol. 33, 81–88. https://doi.org/10.1016/j.coviro.2018.07.013

Fischer, Stefanie, Kerstin Cornils, Thomas Speiseder, Anita Badbaran, Rudolph Reimer, Daniela Indenbirken, Adam Grundhoff, Bärbel Brunswig-Spickenheier, Malik Alawi, et Claudia Lange. 2016. « Indication of Horizontal DNA Gene Transfer by Extracellular Vesicles ». Édité par Giovanni Camussi. *PLOS ONE* 11 (9): e0163665. https://doi.org/10.1371/journal.pone.0163665.

Flynn, J.M., Hubley, R., Goubert, C., Rosen, J., Clark, A.G., Feschotte, C., Smit, A.F., 2020. RepeatModeler2 for automated genomic discovery of transposable element families. Proc. Natl. Acad. Sci. 117, 9451–9457. https://doi.org/10.1073/pnas.1921046117

Flynn, Peter J., et Corrie S. Moreau. 2019. « Assessing the Diversity of Endogenous Viruses Throughout Ant Genomes ». *Frontiers in Microbiology* 10 (mai): 1139. https://doi.org/10.3389/fmicb.2019.01139.

Fouché, Simone, Thomas Badet, Ursula Oggenfuss, Clémence Plissonneau, Carolina Sardinha Francisco, and Daniel Croll. 2020. "Stress-Driven Transposable Element De-Repression Dynamics and Virulence Evolution in a Fungal Pathogen." Edited by Irina Arkhipova. *Molecular Biology and Evolution* 37 (1): 221–39. https://doi.org/10.1093/molbev/msz216.

Frank, John A, et Cédric Feschotte. 2017. « Co-Option of Endogenous Viral Sequences for Host Cell Function ». *Current Opinion in Virology* 25 (août): 81-89. https://doi.org/10.1016/j.coviro.2017.07.021.

Fraser, M. J. 1987. The FP mutation of nuclear polyhedrosis viruses: A novel system for the study of transposon- mediated mutagenesis. Pages 265-293. Biotechnology in Invertebrate Pathology and Cell.

Fraser, M. J., Gale E. Smith, and Max D. Summers. 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. https://doi.org/10.1128/JVI.47.2.287-300.1983.

Fraser, M.J., Brusca, J.S., Smith, G.E., Summers, M.D., 1985. Transposon-mediated mutagenesis of a baculovirus. Virology 145, 356–361. https://doi.org/10.1016/0042-6822(85)90172-2

Fraser, M.J., Cary, L., Boonvisudhi, K., Wang, H.G., 1995. Assay for movement of Lepidopteran transposon IFP2 in insect cells using a baculovirus genome as a target DNA. Virology 211, 397–407. https://doi.org/10.1006/viro.1995.1422

Fu, Yu, Yujing Yang, Han Zhang, Gwen Farley, Junling Wang, Kaycee A Quarles, Zhiping Weng, and Phillip D Zamore. 2018. "The Genome of the Hi5 Germ Cell Line from Trichoplusia Ni, an Agricultural Pest and Novel Model for Small RNA Biology." *ELife* 7 (January): e31628. https://doi.org/10.7554/eLife.31628.

Fujino, K., M. Horie, T. Honda, D. K. Merriman, et K. Tomonaga. 2014. « Inhibition of Borna Disease Virus Replication by an Endogenous Bornavirus-like Element in the Ground Squirrel Genome ». *Proceedings of the National Academy of Sciences* 111 (36): 13175-80. https://doi.org/10.1073/pnas.1407046111.

Fukaya, M., Nasu, S., 1966. A Chilo Iridescent Virus (CIV) from the Rice Stem Borer, Chilo suppressalis WALKER (Lepidoptera : Pyralidae). Appl. Entomol. Zool. 1, 69–72. https://doi.org/10.1303/aez.1.69

Galindo-González, Leonardo, Corinne Mhiri, Michael K. Deyholos, and Marie-Angèle Grandbastien. 2017. "LTR-Retrotransposons in Plants: Engines of Evolution." *Gene* 626 (August): 14–25. https://doi.org/10.1016/j.gene.2017.04.051.

Gartler, S. M. 1968. Apparent HeLa Cell Contamination of Human Heteroploid Cell Lines. *Nature* **217**, 750–751.

Gasmi, Laila, Helene Boulain, Jeremy Gauthier, Aurelie Hua-Van, Karine Musset, Agata K. Jakubowska, Jean-Marc Aury, et al. 2015. « Recurrent Domestication by Lepidoptera of Genes from Their Parasites Mediated by Bracoviruses ». Édité par Cédric Feschotte. *PLOS Genetics* 11 (9): e1005470. https://doi.org/10.1371/journal.pgen.1005470.

Gauthier, Jérémy, Jean-Michel Drezen, et Elisabeth A. Herniou. 2018. « The Recurrent Domestication of Viruses: Major Evolutionary Transitions in Parasitic Wasps ». *Parasitology* 145 (6): 713-23. https://doi.org/10.1017/S0031182017000725.

Gelvin, Stanton B. 2009. « *Agrobacterium* in the Genomics Age ». *Plant Physiology* 150 (4): 1665-76. https://doi.org/10.1104/pp.109.139873. Genomes 10:937–951.

Ghoshal, K., Theilmann, J., Reade, R., Maghodia, A., Rochon, D., 2015. Encapsidation of Host RNAs by Cucumber Necrosis Virus Coat Protein during both Agroinfiltration and Infection. J. Virol. 89, 10748–10761. https://doi.org/10.1128/JVI.01466-15

Gilbert, C., Chateigner, A., Ernenwein, L., Barbe, V., Bézier, A., Herniou, E.A., Cordaux, R., 2014. Population genomics supports baculoviruses as vectors of horizontal transfer of insect transposons. Nat. Commun. 5, 3348. https://doi.org/10.1038/ncomms4348

Gilbert, C., Cordaux, R., 2017. Viruses as vectors of horizontal transfer of genetic material in eukaryotes. Curr. Opin. Virol. 25, 16–22. https://doi.org/10.1016/j.coviro.2017.06.005

Gilbert, C., Feschotte, C., 2018. Horizontal acquisition of transposable elements and viral sequences: patterns and consequences. Curr. Opin. Genet. Dev. 49, 15–24. https://doi.org/10.1016/j.gde.2018.02.007

Gilbert, C., Peccoud, J., Chateigner, A., Moumen, B., Cordaux, R., Herniou, E.A., 2016. Continuous Influx of Genetic Material from Host to Virus Populations. PLoS Genet. 12, e1005838. https://doi.org/10.1371/journal.pgen.1005838

Gilbert, C., Schaack, S., Pace, J.K., Brindley, P.J., Feschotte, C., 2010. A role for host-parasite interactions in the horizontal transfer of transposons across phyla. Nature 464, 1347–1350. https://doi.org/10.1038/nature08939

Godfray, H.C.J., Reilly, D.R.O., and Briggs, C.J. 1997. A Model of Nucleopolyhedrovirus (NPV) Population Genetics Applied to Co-Occlusion and the Spread of the Few Polyhedra (FP) Phenotype. Proc. Biol. Sci. *264*, 315–322.

Gorphe, P. A 2019. comprehensive review of Hep-2 cell line in translational research for laryngeal cancer. *Am. J. Cancer Res.* **9**, 644–649.

Görzer, I., Guelly, C., Trajanoski, S., and Puchhammer-Stöckl, E. 2010. The impact of PCR-generated recombination on diversity estimation of mixed viral populations by deep sequencing. J. Virol. Methods *169*, 248–252.

Goulson, D., 2003. Can Host Susceptibility to Baculovirus Infection be Predicted from Host Taxonomy or Life History? Environ. Entomol. 32, 61–70. https://doi.org/10.1603/0046-225X-32.1.61

Grabherr, Manfred G, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn A Thompson, Ido Amit, Xian Adiconis, et al. 2011. "Full-Length Transcriptome Assembly from RNA-Seq Data without a Reference Genome." *Nature Biotechnology* 29 (7): 644–52. https://doi.org/10.1038/nbt.1883.

Graham, L. A., Lougheed S. C., Ewart K. V. et Davies P. L. 2008. "Lateral transfer of a lectin-like antifreeze protein gene in fishes". *PLoS One*, 9;3(7):e2616. https://doi.org/10.1371/journal.pone.0002616.

Graillot, B., Berling, M., Blachere-López, C., Siegwart, M., Besse, S., López-Ferber, M., 2014. Progressive Adaptation of a CpGV Isolate to Codling Moth Populations Resistant to CpGV-M. Viruses 6, 5135–5144. https://doi.org/10.3390/v6125135

Granados, Robert R., Anja C.G. Derksen, and Kathleen G. Dwyer. 1986. "Replication of the Trichoplusia Ni Granulosis and Nuclear Polyhedrosis Viruses in Cell Cultures." *Virology* 152 (2): 472–76. https://doi.org/10.1016/0042-6822(86)90150-9.

Granados, Robert R., Li Guoxun, Anja C.G. Derksen, and Kevin A. McKenna. 1994. "A New Insect Cell Line from Trichoplusia Ni (BTI-Tn-5B1-4) Susceptible to Trichoplusia Ni Single Enveloped Nuclear Polyhedrosis Virus." *Journal of Invertebrate Pathology* 64 (3): 260–66. https://doi.org/10.1016/S0022-2011(94)90400-6.

Griffith, P., Raley, C., Sun, D., Zhao, Y., Sun, Z., Mehta, M., Tran, B., and Wu, X. 2018. PacBio library preparation using blunt-end adapter ligation produces significant artefactual fusion DNA sequences. BioRxiv.

Groot, Michael, et Heedoo Lee. 2020. « Sorting Mechanisms for MicroRNAs into Extracellular Vesicles and Their Associated Diseases ». *Cells* 9 (4): 1044. https://doi.org/10.3390/cells9041044.

Gueli Alletti, G., Carstens, E.B., Weihrauch, B., Jehle, J.A., 2018. Agrotis segetum nucleopolyhedrovirus but not Agrotis segetum granulovirus replicate in AiE1611T cell line of Agrotisipsilon. J. Invertebr. Pathol. 151, 7–13. https://doi.org/10.1016/j.jip.2017.10.005

Gueli Alletti, G., Eigenbrod, M., Carstens, E.B., Kleespies, R.G., Jehle, J.A., 2017. The genome sequence of Agrotis segetum granulovirus, isolate AgseGV-DA, reveals a new Betabaculovirus species of a slow killing granulovirus. J. Invertebr. Pathol. 146, 58–68. https://doi.org/10.1016/j.jip.2017.04.008

Gundersen-Rindal, Dawn, Catherine Dupuy, Elisabeth Huguet, et Jean-Michel Drezen. 2013. « Parasitoid Polydnaviruses: Evolution, Pathology and Applications: Dedicated to the Memory of Nancy E. Beckage ». *Biocontrol Science and Technology* 23 (1): 1-61. https://doi.org/10.1080/09583157.2012.731497.

Guo, X., Gao, J., Li, F., Wang, J., 2014. Evidence of horizontal transfer of non-autonomous Lep1 Helitrons facilitated by host-parasite interactions. Sci. Rep. 4, 5119. https://doi.org/10.1038/srep05119

Haegeman, Annelies, John T. Jones, et Etienne G. J. Danchin. 2011. « Horizontal Gene Transfer in Nematodes: A Catalyst for Plant Parasitism? » *Molecular Plant-Microbe Interactions* 24 (8): 879-87. https://doi.org/10.1094/MPMI-03-11-0055.

Han, M.-J., Zhou, Q.-Z., Zhang, H.-H., Tong, X., Lu, C., Zhang, Z., Dai, F., 2016. iMITEdb: the genome-wide landscape of miniature inverted-repeat transposable elements in insects. Database 2016, baw148. https://doi.org/10.1093/database/baw148

Harrison, R.L., Lynn, D.E., 2008. New cell lines derived from the black cutworm, Agrotis ipsilon, that support replication of the A. ipsilon multiple nucleopolyhedrovirus and several group I nucleopolyhedroviruses. J. Invertebr. Pathol. 99, 28–34. https://doi.org/10.1016/j.jip.2008.02.015

Hazkani-Covo, Einat, Raymond M. Zeller, et William Martin. 2010. « Molecular Poltergeists: Mitochondrial DNA Copies (Numts) in Sequenced Nuclear Genomes ». Édité par Harmit S. Malik. *PLoS Genetics* 6 (2): e1000834. https://doi.org/10.1371/journal.pgen.1000834.

Hempel, H. A., Burns, K. H., De Marzo, A. M. & Sfanos, K. S. 2013. Infection of Xenotransplanted Human Cell Lines by Murine Retroviruses: A Lesson Brought Back to Light by XMRV. *Front. Oncol.* **3**.

Herniou, Elisabeth A., Elisabeth Huguet, Julien Thézé, Annie Bézier, Georges Periquet, et Jean-Michel Drezen. 2013. « When Parasitic Wasps Hijacked Viruses: Genomic and Functional Evolution of Polydnaviruses ». *Philosophical Transactions of the Royal Society B: Biological Sciences* 368 (1626): 20130051. https://doi.org/10.1098/rstb.2013.0051.

Hilgenboecker, Kirsten, Peter Hammerstein, Peter Schlattmann, Arndt Telschow, et John H. Werren. 2008. « How Many Species Are Infected with Wolbachia? – A Statistical Analysis of Current Data: Wolbachia Infection Rates ». *FEMS Microbiology Letters* 281 (2): 215-20. https://doi.org/10.1111/j.1574-6968.2008.01110.x.

Hink, W. F. and P. V. Vail. 1973. A plaque assay for titration of alfalfa looper nuclear polyhedrosis virus in a cabbage looper (TN-368) cell line. Journal of Invertebrate Pathology **22**:168-174.

Holmes, Edward C. 2011. « The Evolution of Endogenous Viral Elements ». *Cell Host & Microbe* 10 (4): 368-77. https://doi.org/10.1016/j.chom.2011.09.002.

Horbach, S. P. J. M. & Halffman, W. 2017. The ghosts of HeLa: How cell line misidentification contaminates the scientific literature. *PLOS ONE* **12**, e0186281.

Horváth, Vivien, Miriam Merenciano, and Josefa González. 2017. "Revisiting the Relationship between Transposable Elements and the Eukaryotic Stress Response." *Trends in Genetics* 33 (11): 832–41. https://doi.org/10.1016/j.tig.2017.08.007.

Hotopp, J. C. D., M. E. Clark, D. C. S. G. Oliveira, J. M. Foster, P. Fischer, M. C. M. Torres, J. D. Giebel, et al. 2007. « Widespread Lateral Gene Transfer from Intracellular Bacteria to Multicellular Eukaryotes ». *Science* 317 (5845): 1753-56. https://doi.org/10.1126/science.1142490.

Huang, Jianhua, Yushuai Wang, Wenwen Liu, Xu Shen, Qiang Fan, Shuguang Jian, and Tian Tang. 2017. "EARE-1, a Transcriptionally Active Ty1/Copia-Like Retrotransposon Has Colonized the Genome of Excoecaria Agallocha through Horizontal Transfer." *Frontiers in Plant Science* 8 (January). https://doi.org/10.3389/fpls.2017.00045.

Hughes, A.L. 2003. Genome-Wide Survey for Genes Horizontally Transferred from Cellular Organisms to Baculoviruses. Mol. Biol. Evol. *20*, 979–987.

Hummel, Barbara, Erik C Hansen, Aneliya Yoveva, Fernando Aprile-Garcia, Rebecca Hussong, and Ritwick Sawarkar. 2017. "The Evolutionary Capacitor HSP90 Buffers the Regulatory Effects of Mammalian Endogenous Retroviruses." *Nature Structural & Molecular Biology* 24 (3): 234–42. https://doi.org/10.1038/nsmb.3368.

Intrieri, Maria Carmela, et Marcello Buiatti. 2001. « The Horizontal Transfer of Agrobacterium Rhizogenes Genes and the Evolution of the Genus Nicotiana ». *Molecular Phylogenetics and Evolution* 20 (1): 100-110. https://doi.org/10.1006/mpev.2001.0927.

Ivancevic, A.M., Kortschak, R.D., Bertozzi, T., Adelson, D.L., 2018. Horizontal transfer of BovB and L1 retrotransposons in eukaryotes. Genome Biol. 19, 85. https://doi.org/10.1186/s13059-018-1456-7

Jabalee, James, Rebecca Towle, et Cathie Garnis. 2018. « The Role of Extracellular Vesicles in Cancer: Cargo, Function, and Therapeutic Implications ». *Cells* 7 (8): 93. https://doi.org/10.3390/cells7080093.

Jang, Hyo Sik, Nakul M. Shah, Alan Y. Du, Zea Z. Dailey, Erica C. Pehrsson, Paula M. Godoy, David Zhang, et al. 2019. "Transposable Elements Drive Widespread Expression of Oncogenes in Human Cancers." *Nature Genetics* 51 (4): 611–17. https://doi.org/10.1038/s41588-019-0373-3.

Jaworski, E., and Routh, A. 2017. Parallel ClickSeq and Nanopore sequencing elucidates the rapid evolution of defective-interfering RNAs in Flock House virus. PLOS Pathog. *13*, e1006365.

Jehle, J.A, I.F.A van der Linden, H Backhaus, et J.M Vlak. 1997. « Identification and Sequence Analysis of the Integration Site of Transposon TCp3.2 in the Genome of Cydia Pomonella Granulovirus ». *Virus Research* 50 (2): 151-57. https://doi.org/10.1016/S0168-1702(97)00066-X.

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, 369–379. https://doi.org/10.1006/viro.1995.1096

Jehle, J.A., Nickel, A., Vlak, J.M., Backhaus, H., 1998. Horizontal escape of the novel Tc1-like lepidopteran transposon TCp3.2 into Cydia pomonella granulovirus. J. Mol. Evol. 46, 215–224. https://doi.org/10.1007/pl00006296

Johnson, A. D. & Cohn, C. S. 2016. Xenotropic Murine Leukemia Virus-Related Virus (XMRV) and the Safety of the Blood Supply. *Clin. Microbiol. Rev.* **29**, 749–757.

Kale, Shubha, Lakisha Moore, Prescott Deininger, and Astrid Roy-Engel. 2005. "Heavy Metals Stimulate Human LINE-1 Retrotransposition." *International Journal of Environmental Research and Public Health* 2 (1): 14–23. https://doi.org/10.3390/ijerph2005010014.

Kalluri, Raghu, et Valerie S. LeBleu. 2016. « Discovery of Double-Stranded Genomic DNA in Circulating Exosomes ». *Cold Spring Harbor Symposia on Quantitative Biology* 81: 275-80. https://doi.org/10.1101/sqb.2016.81.030932.

Kamita, S.G., Maeda, S., and Hammock, B.D. 2003. High-Frequency Homologous Recombination between Baculoviruses Involves DNA Replication. J. Virol. *77*, 13053–13061.

Karamitros, T., Harrison, I., Piorkowska, R., Katzourakis, A., Magiorkinis, G., and Mbisa, J.L. 2016. De Novo Assembly of Human Herpes Virus Type 1 (HHV-1) Genome, Mining of Non-Canonical Structures and Detection of Novel Drug-Resistance Mutations Using Short- and Long-Read Next Generation Sequencing Technologies. PLOS ONE *11*, e0157600.

Karamitros, T., van Wilgenburg, B., Wills, M., Klenerman, P., and Magiorkinis, G. 2018. Nanopore sequencing and full genome de novo assembly of human cytomegalovirus TB40/E reveals clonal diversity and structural variations. BMC Genomics *19*.

Katoh, K., Standley, D.M., 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol. Biol. Evol. 30, 772–780. https://doi.org/10.1093/molbev/mst010

Katsuma, S., Fujii, T., Kawaoka, S., Shimada, T., 2008. Bombyx mori nucleopolyhedrovirus SNF2 global transactivator homologue (Bm33) enhances viral pathogenicity in B. mori larvae. J. Gen. Virol. 89, 3039–3046. https://doi.org/10.1099/vir.0.2008/004887-0

Katzourakis, Aris, et Robert J. Gifford. 2010. « Endogenous Viral Elements in Animal Genomes ». Édité par Harmit S. Malik. *PLoS Genetics* 6 (11): e1001191. https://doi.org/10.1371/journal.pgen.1001191.

Kawamura, Y., Sanchez Calle, A., Yamamoto, Y., Sato, T.-A., Ochiya, T., 2019. Extracellular vesicles mediate the horizontal transfer of an active LINE-1 retrotransposon. J. Extracell. Vesicles 8, 1643214. https://doi.org/10.1080/20013078.2019.1643214

Kidwell, Margaret G., and Damon R. Lisch. 2001. "Perspective: transposable elements, parasitic DNA, and genome evolution." *Evolution* 55 (1): 1–24. https://doi.org/10.1111/j.0014-3820.2001.tb01268.x.

Kilpatrick, B.A., and Huang, E.-S. 1977. Human Cytomegalovirus Genome: Partial Denaturation Map and Organization of Genome Sequences. J VIROL *24*, 16.

Kimura, M. 1983. The Neutral Theory of Molecular Evolution. Cambridge University Press,Cambridge.

Kirkegaard, J. S. *et al.* 2016. Xenotropic retrovirus Bxv1 in human pancreatic β cell lines. *J. Clin. Invest.* **126**, 1109–1113.

Kitts, P.A., Ayres, M.D., and Possee, R.D. 1990. Linearization of baculovirus DNA enhances the recovery of recombinant virus expression vectors. Nucleic Acids Res. *18*, 5667–5672.

Klump, Jennifer, Ulrike Phillipp, Marie Follo, Anna Eremin, Hannes Lehmann, Sigrun Nestel, Nikolas von Bubnoff, et Irina Nazarenko. 2018. « Extracellular Vesicles or Free Circulating DNA: Where to Search for BRAF and CKIT Mutations? » *Nanomedicine: Nanotechnology, Biology and Medicine* 14 (3): 875-82. https://doi.org/10.1016/j.nano.2017.12.009.

Kofler, R., Nolte, V., Schlötterer, C., 2015. Tempo and Mode of Transposable Element Activity in Drosophila. PLoS Genet. 11, e1005406. https://doi.org/10.1371/journal.pgen.1005406

Kondo, Natsuko, Naruo Nikoh, Nobuyuki Ijichi, Masakazu Shimada, et Takema Fukatsu. 2002. « Genome Fragment of Wolbachia Endosymbiont Transferred to X Chromosome of Host Insect ». *Proceedings of the National Academy of Sciences of the United States of America* 99 (22): 14280-85.

Kool, M., Ahrens, C.H., Vlak, J.M., and Rohrmann, G.F. 1995. Replication of baculovirus DNA. J. Gen. Virol. *76*, 2103–2118.

Kool, M., Voncken, J.W., Van Lier, F.L.J., Tramper, J., and Vlak, J.M. 1991. Detection and analysis of Autographa californica nuclear polyhedrosis virus mutants with defective interfering properties. Virology *183*, 739–746.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. 2017. Canu: scalable and accurate long-read assembly via adaptive *k* -mer weighting and repeat separation. Genome Res. *27*, 722–736.

Kozak, C. & Rowe, W. 1978. Genetic mapping of xenotropic leukemia virus-inducing loci in two mouse strains. *Science* **199**, 1448–1449.

Kozak, C. A. 2010. The mouse" xenotropic" gammaretroviruses and their XPR1 receptor. *Retrovirology* **7**, 101.

Kozak, C. 2014. Origins of the Endogenous and Infectious Laboratory Mouse Gammaretroviruses. *Viruses* **7**, 1–26.

Kremer, M. *et al.* 2006. Vaccinia virus replication is not affected by APOBEC3 family members. *Virol. J.* **3**, 86.

Kronenberg, Z.N., Osborne, E.J., Cone, K.R., Kennedy, B.J., Domyan, E.T., Shapiro, M.D., Elde, N.C., and Yandell, M. 2015. Wham: Identifying Structural Variants of Biological Consequence. PLOS Comput. Biol. *11*, e1004572.

Krupovic, Mart, et Eugene V. Koonin. 2015. « Evolution of Eukaryotic Single-Stranded DNA Viruses of the Bidnaviridae Family from Genes of Four Other Groups of Widely Different Viruses ». *Scientific Reports* 4 (1): 5347. https://doi.org/10.1038/srep05347.

Kulkarni, A.S., and Fortunato, E.A. 2011. Stimulation of Homology-Directed Repair at I-SceI-Induced DNA Breaks during the Permissive Life Cycle of Human Cytomegalovirus. J. Virol. *85*, 6049–6054.

Kumar, S., Stecher, G., Suleski, M., Hedges, S.B., 2017. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Mol. Biol. Evol. 34, 1812–1819. https://doi.org/10.1093/molbev/msx116

Kuraku, S., Qiu, H., Meyer, A., 2012. Horizontal Transfers of Tc1 Elements between Teleost Fishes and Their Vertebrate Parasites, Lampreys. Genome Biol. Evol. 4, 929–936. https://doi.org/10.1093/gbe/evs069

Kyndt, Tina, Dora Quispe, Hong Zhai, Robert Jarret, Marc Ghislain, Qingchang Liu, Godelieve Gheysen, et Jan F. Kreuze. 2015. « The Genome of Cultivated Sweet Potato Contains *Agrobacterium* T-DNAs with Expressed Genes: An Example of a Naturally Transgenic Food Crop ». *Proceedings of the National Academy of Sciences* 112 (18): 5844-49. https://doi.org/10.1073/pnas.1419685112.

Lacroix, Benoît, et Vitaly Citovsky. 2016. « Transfer of DNA from Bacteria to Eukaryotes ». *MBio* 7 (4): e00863-16, /mbio/7/4/e00863-16.atom. https://doi.org/10.1128/mBio.00863-16.

Lanciano, Sophie, and Marie Mirouze. 2018. “Transposable Elements: All Mobile, All Different, Some Stress Responsive, Some Adaptive?” *Current Opinion in Genetics & Development* 49 (April): 106–14. https://doi.org/10.1016/j.gde.2018.04.002.

Lander et al. 2001. « Initial Sequencing and Analysis of the Human Genome ». *Nature* 409 (6822): 860-921. https://doi.org/10.1038/35057062.

Langmead, Ben, and Steven L Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9 (4): 357–59. https://doi.org/10.1038/nmeth.1923.

Lauring, A.S., and Andino, R. 2010. Quasispecies Theory and the Behavior of RNA Viruses. PLoS Pathog. *6*, e1001005.

Lauring, A.S., Frydman, J., and Andino, R. 2013. The role of mutational robustness in RNA virus evolution. Nat. Rev. Microbiol. *11*, 327–336.

Layer, R.M., Chiang, C., Quinlan, A.R., and Hall, I.M. 2014. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. *15*, R84.

Lázaro-Ibáñez, Elisa, Andres Sanz-Garcia, Tapio Visakorpi, Carmen Escobedo-Lucea, Pia Siljander, Ángel Ayuso-Sacido, et Marjo Yliperttula. 2014. « Different GDNA Content in the Subpopulations of Prostate Cancer Extracellular Vesicles: Apoptotic Bodies, Microvesicles, and Exosomes: Different GDNA Content in PCa EV Subpopulations ». *The Prostate* 74 (14): 1379-90. https://doi.org/10.1002/pros.22853.

Lázaro-Ibáñez, Elisa, Cecilia Lässer, Ganesh Vilas Shelke, Rossella Crescitelli, Su Chul Jang, Aleksander Cvjetkovic, Anaís García-Rodríguez, et Jan Lötvall. 2019. « DNA Analysis of Low- and High-Density Fractions Defines Heterogeneous Subpopulations of Small Extracellular Vesicles Based on Their DNA Cargo and Topology ». *Journal of Extracellular Vesicles* 8 (1): 1656993. https://doi.org/10.1080/20013078.2019.1656993.

Le Rouzic, A., T. S. Boutin, and P. Capy. 2007. “Long-Term Evolution of Transposable Elements.” *Proceedings of the National Academy of Sciences* 104 (49): 19375–80. https://doi.org/10.1073/pnas.0705238104.

Leblanc, P., S. Desset, F. Giorgi, A. R. Taddei, A. M. Fausto, M. Mazzini, B. Dastugue, et C. Vaury. 2000. « Life Cycle of an Endogenous Retrovirus,ZAM, in Drosophila Melanogaster ». *Journal of Virology* 74 (22): 10658-69. https://doi.org/10.1128/JVI.74.22.10658-10669.2000.

Leclercq, Sébastien, Julien Thézé, Mohamed Amine Chebbi, Isabelle Giraud, Bouziane Moumen, Lise Ernenwein, Pierre Grève, Clément Gilbert, et Richard Cordaux. 2016. « Birth of a W Sex Chromosome by Horizontal Transfer of *Wolbachia* Bacterial Symbiont Genome ». *Proceedings of the National Academy of Sciences* 113 (52): 15036-41. https://doi.org/10.1073/pnas.1608979113.

Legrand, Jean-Jacques, et Pierre Juchault. 1984. « Nouvelles données sur le déterminisme génétique et épigénétique de la monogénie chez le crustacé isopode terrestre Armadillidium vulgare Latr. » 16: 57-84.

Lerat, Emmanuelle, Marie Fablet, Laurent Modolo, Hélène Lopez-Maestre, and Cristina Vieira. 2016. “TEtools Facilitates Big Data Expression Analysis of Transposable Elements and Reveals an Antagonism between Their Activity and That of PiRNA Genes.” *Nucleic Acids Research*, October, gkw953. https://doi.org/10.1093/nar/gkw953.

Li, D., Lott, W.B., Lowry, K., Jones, A., Thu, H.M., and Aaskov, J. 2011. Defective Interfering Viral Particles in Acute Dengue Infections. PLoS ONE *6*, e19447.

Li, H. 2015. FermiKit: assembly-based variant calling for Illumina resequencing data. Bioinformatics btv440.

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

Lin, Z. *et al.* Detection of Murine Leukemia Virus in the Epstein-Barr Virus-Positive Human B-Cell Line JY, Using a Computational RNA-Seq-Based Exogenous Agent Detection Pipeline, PARSES. *J. Virol.* **86**, 2970–2977 2012.

Loiseau, V., Herniou, E.A., Moreau, Y., Lévêque, N., Meignin, C., Daeffler, L., Federici, B., Cordaux, R., Gilbert, C., 2020. Wide spectrum and high frequency of genomic structural variation, including transposable elements, in large double-stranded DNA viruses. Virus Evol. 6, vez060. https://doi.org/10.1093/ve/vez060

Lopez, C.B. 2014. Defective Viral Genomes: Critical Danger Signals of Viral Infections. J. Virol. *88*, 8720–8723.

López-Ferber, M., Simón, O., Williams, T., and Caballero, P. 2003. Defective or effective? Mutualistic interactions between virus genotypes. Proc. R. Soc. Lond. B Biol. Sci. *270*, 2249–2255.

Loreto, E.L.S., Carareto, C.M.A., Capy, P., 2008. Revisiting horizontal transfer of transposable elements in Drosophila. Heredity 100, 545–554. https://doi.org/10.1038/sj.hdy.6801094

Love, Michael I, Wolfgang Huber, and Simon Anders. 2014. "Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2." *Genome Biology* 15 (12): 550. https://doi.org/10.1186/s13059-014-0550-8.

Lupetti, P., Montesanto, G., Ciolfi, S., Marri, L., Gentile, M., Paccagnini, E., and Lombardo, B.M. 2013. Iridovirus infection in terrestrial isopods from Sicily (Italy). Tissue Cell *45*, 321–327.

Macchietto, Marissa G, Ryan A Langlois, and Steven S Shen. 2020. "Virus-Induced Transposable Element Expression upregulation in Human and Mouse Host Cells." *Life Science Alliance* 3 (2): e201900536. https://doi.org/10.26508/lsa.201900536.

Maghodia, A.B., Jarvis, D.L., and Geisler, C. 2014. Complete Genome Sequence of the Autographa californica Multiple Nucleopolyhedrovirus Strain E2. Genome Announc. *2*.

Mahiet, C., Ergani, A., Huot, N., Alende, N., Azough, A., Salvaire, F., Bensimon, A., Conseiller, E., Wain-Hobson, S., Labetoulle, M., et al. 2012. Structural Variability of the Herpes Simplex Virus 1 Genome *In Vitro* and *In Vivo*. J. Virol. *86*, 8592–8601.

Manzoni, T.B., and López, C.B. 2018. Defective (interfering) viral genomes re-explored: impact on antiviral immunity and virus persistence. Future Virol. *13*, 493–503.

Marriott, A.C., and Dimmock, N.J. 2010. Defective interfering viruses and their potential as antiviral agents. Rev. Med. Virol. *20*, 51–62.

Matveeva, Tatiana V., Denis I. Bogomaz, Olga A. Pavlova, Eugene W. Nester, et Ludmila A. Lutova. 2012. « Horizontal Gene Transfer from Genus *Agrobacterium* to the Plant *Linaria* in Nature ». *Molecular Plant-Microbe Interactions* 25 (12): 1542-51. https://doi.org/10.1094/MPMI-07-12-0169-R.

McAllister, R. M. *et al.* 1972. C-Type Virus Released from Cultured Human Rhabdomyosarcoma Cells. *Nature. New Biol.* **235**, 3–6.

McClintock, B. 1950. "The Origin and Behavior of Mutable Loci in Maize." *Proceedings of the National Academy of Sciences* 36 (6): 344–55. https://doi.org/10.1073/pnas.36.6.344.

McClintock, B. 1984. "The Significance of Responses of the Genome to Challenge." *Science* 226 (4676): 792–801. https://doi.org/10.1126/science.15739260.

Mendel, G. 1866. Versuche über Pflanzenhybriden. Verhandlungen des naturforschenden Vereines in Brünn, Bd. IV für das Jahr 1865, Abhandlungen, 3–47.

Menees, T. M., and S. B. Sandmeyer. 1996. "Cellular Stress Inhibits Transposition of the Yeast Retrovirus-like Element Ty3 by a Ubiquitin-Dependent Block of Virus-like Particle Formation." *Proceedings of the National Academy of Sciences* 93 (11): 5629–34. https://doi.org/10.1073/pnas.93.11.5629.

Merten, O.-W. 2002. Virus contaminations of cell cultures - A biotechnological view. *Cytotechnology* **39**, 91–116.

Miller, D.W., Miller, L.K., 1982. A virus mutant with an insertion of a copia-like transposable element. Nature 299, 562–564. https://doi.org/10.1038/299562a0

Miousse, Isabelle R., Marie-Cecile G. Chalbot, Annie Lumen, Alesia Ferguson, Ilias G. Kavouras, and Igor Koturbash. 2015. "Response of Transposable Elements to Environmental Stressors." *Mutation Research/Reviews in Mutation Research* 765 (July): 19–39. https://doi.org/10.1016/j.mrrev.2015.05.003.

Mirjalili, A., Parmoor, E., Moradi Bidhendi, S. & Sarkari, B. 2005. Microbial contamination of cell cultures: A 2 years study. *Biologicals* **33**, 81–85.

Mita, Paolo, and Jef D Boeke. 2016. "How Retrotransposons Shape Genome Regulation." *Current Opinion in Genetics & Development* 37 (April): 90–100. https://doi.org/10.1016/j.gde.2016.01.001.

Modolo, Laurent, and Emmanuelle Lerat. 2015. "UrQt: An Efficient Software for the Unsupervised Quality Trimming of NGS Data." *BMC Bioinformatics* 16 (1): 137. https://doi.org/10.1186/s12859-015-0546-8.

Mohiyuddin, M., Mu, J.C., Li, J., Bani Asadi, N., Gerstein, M.B., Abyzov, A., Wong, W.H., and Lam, H.Y.K. 2015. MetaSV: an accurate and integrative structural-variant caller for next generation sequencing. Bioinformatics *31*, 2741–2744.

Moldovan, Leni, Kara Batte, Yijie Wang, Jon Wisler, et Melissa Piper. 2013. « Analyzing the Circulating MicroRNAs in Exosomes/Extracellular Vesicles from Serum or Plasma by QRT-PCR ». In *Circulating MicroRNAs*, édité par Nobuyoshi Kosaka, 1024:129-45. Totowa, NJ: Humana Press. https://doi.org/10.1007/978-1-62703-453-1_10.

Mönchgesang, S., Strehmel, N., Schmidt, S., Westphal, L., Taruttis, F., Müller, E., Herklotz, S., Neumann, S., and Scheel, D. 2016. Natural variation of root exudates in Arabidopsis thaliana-linking metabolomic and genomic data. Sci. Rep. *6*.

Moore, A. E., Sabachewsky, L. & Toolan, H. W. 1955. Culture characteristics of four permanent lines of human cancer cells. *Cancer Res.* **15**, 598–602.

Mortazavi, Ali, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. 2008. "Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq." *Nature Methods* 5 (7): 621–28. https://doi.org/10.1038/nmeth.1226.

Müllner, D. 2013. **fastcluster** : Fast Hierarchical, Agglomerative Clustering Routines for *R* and *Python*. J. Stat. Softw. *53*.

Nardone, R. M. 2007. Eradication of cross-contaminated cell lines: A call for action. *Cell Biol. Toxicol.* **23**, 367–372.

Naseer, A. *et al.* 2015. Frequent Infection of Human Cancer Xenografts with Murine Endogenous Retroviruses in Vivo. *Viruses* **7**, 2014–2029.

Negi, Pooja, Archana N. Rai, and Penna Suprasanna. 2016. "Moving through the Stressed Genome: Emerging Regulatory Roles for Transposons in Plant Stress Response." *Frontiers in Plant Science* 7 (October). https://doi.org/10.3389/fpls.2016.01448.

Nelson-Rees, W., Daniels, D. & Flandermeyer, R. 1981. Cross-contamination of cells in culture. *Science* **212**, 446–452.

Niederhuth, Chad E., Adam J. Bewick, Lexiang Ji, Magdy S. Alabady, Kyung Do Kim, Qing Li, Nicholas A. Rohr, et al. 2016. "Widespread Natural Variation of DNA Methylation within Angiosperms." *Genome Biology* 17 (1): 194. https://doi.org/10.1186/s13059-016-1059-0.

Nosenko, Tetyana, et Debashish Bhattacharya. 2007. « Horizontal Gene Transfer in Chromalveolates ». *BMC Evolutionary Biology* 7 (1): 173. https://doi.org/10.1186/1471-2148-7-173.

O'Reilly, D.R., Miller, L., Luckow, V.A., 1992. Baculovirus expression vectors: a laboratory manual. W.H. Freeman, New York.

Oakes, B. *et al.* 2010. Contamination of human DNA samples with mouse DNA can lead to false detection of XMRV-like sequences. *Retrovirology* **7**, 109.

Oger, P., Petit, A., et Dessaux, Y. 1997. Genetically engineered plants producing opines alter their biological environment. Nat. Biotechnol. 15, 369–372. doi: 10.1038/nbt0497-369

Ohshima, K., et N. Okada. 2005. « SINEs and LINEs: Symbionts of Eukaryotic Genomes with a Common Tail ». *Cytogenetic and Genome Research* 110 (1-4): 475-90. https://doi.org/10.1159/000084981.

Ono, R., Yasuhiko, Y., Aisaki, K.-I., Kitajima, S., Kanno, J., Hirabayashi, Y., 2019. Exosome-mediated horizontal gene transfer occurs in double-strand break repair during genome editing. Commun. Biol. 2, 57. https://doi.org/10.1038/s42003-019-0300-2

Ono, Y., Asai, K., and Hamada, M. 2013. PBSIM: PacBio reads simulator—toward accurate genome assembly. Bioinformatics *29*, 119–121.

Otten, Leon, Jean-Yves Salomone, Anne Helfer, Julien Schmidt, Philippe Hammann, et Patrice De Ruffray. 1999. « Sequence and Functional Analysis of the Left-Hand Part of the T-Region from the Nopaline-Type Ti Plasmid, PTiC58 », 12.

Paganini, Julien, Amandine Campan-Fournier, Martine Da Rocha, Philippe Gouret, Pierre Pontarotti, Eric Wajnberg, Pierre Abad, et Etienne G. J. Danchin. 2012. « Contribution of Lateral Gene

Transfers to the Genome Composition and Parasitic Ability of Root-Knot Nematodes ». Édité par Carlos Eduardo Winter. *PLoS ONE* 7 (11): e50875. https://doi.org/10.1371/journal.pone.0050875.

Paprotka, T. *et al.* 2011. Recombinant Origin of the Retrovirus XMRV. *Science* **333**, 97–101.

Parikh, H., Mohiyuddin, M., Lam, H.Y.K., Iyer, H., Chen, D., Pratt, M., Bartha, G., Spies, N., Losert, W., Zook, J.M., et al. 2016. svclassify: a method to establish benchmark structural variant calls. BMC Genomics *17*.

Pascual, Laura, Agata K. Jakubowska, Jose M. Blanca, Joaquin Cañizares, Juan Ferré, Gernot Gloeckner, Heiko Vogel, et Salvador Herrero. 2012. « The Transcriptome of Spodoptera Exigua Larvae Exposed to Different Types of Microbes ». *Insect Biochemistry and Molecular Biology* 42 (8): 557-70. https://doi.org/10.1016/j.ibmb.2012.04.003.

Pavlova, O. A., T. V. Matveeva, et L. A. Lutova. 2014. « Genome of Linaria Dalmatica Contains Agrobacterium Rhizogenes RolC Gene Homolog ». *Russian Journal of Genetics: Applied Research* 4 (5): 461-65. https://doi.org/10.1134/S2079059714050116.

Pearson, M., Bjornson, R., Pearson, G., and Rohrmann, G. 1992. The Autographa californica baculovirus genome: evidence for multiple replication origins. Science *257*, 1382–1384.

Peccoud, J., Cordaux, R., Gilbert, C., 2018. Analyzing Horizontal Transfer of Transposable Elements on a Large Scale: Challenges and Prospects. BioEssays 40, 1700177. https://doi.org/10.1002/bies.201700177

Peccoud, J., Loiseau, V., Cordaux, R., and Gilbert, C. 2017. Massive horizontal transfer of transposable elements in insects. Proc. Natl. Acad. Sci. *114*, 4721–4726. https://doi.org/10.1073/pnas.1621178114

Peccoud, Jean, Sébastian Lequime, Isabelle Moltini-Conclois, Isabelle Giraud, Louis Lambrechts, and Clément Gilbert. 2018. "A Survey of Virus Recombination Uncovers Canonical Features of Artificial Chimeras Generated During Deep Sequencing Library Preparation." *G3: Genes|Genomes|Genetics* 8 (4): 1129–38. https://doi.org/10.1534/g3.117.300468.

Pfeifer, Philipp, Nikos Werner, et Felix Jansen. 2015. « Role and Function of MicroRNAs in Extracellular Vesicles in Cardiovascular Biology ». *BioMed Research International* 2015: 1-11. https://doi.org/10.1155/2015/161393.

Piégu, Benoît, Sébastien Guizard, Tan Yeping, Corinne Cruaud, Sassan Asgari, Dennis K. Bideshi, Brian A. Federici, et Yves Bigot. 2014. « Genome Sequence of a Crustacean Iridovirus, IIV31, Isolated from the Pill Bug, Armadillidium Vulgare ». *Journal of General Virology* 95 (7): 1585-90. https://doi.org/10.1099/vir.0.066076-0.

Piégu, Benoît, Sébastien Guizard, Tatsinda Spears, Corinne Cruaud, Arnault Couloux, Dennis K. Bideshi, Brian A. Federici, et Yves Bigot. 2013. « Complete Genome Sequence of Invertebrate Iridescent Virus 22 Isolated from a Blackfly Larva ». *Journal of General Virology* 94 (9): 2112-16. https://doi.org/10.1099/vir.0.054213-0.

Piskurek, O., et N. Okada. 2007. « Poxviruses as Possible Vectors for Horizontal Transfer of Retroposons from Reptiles to Mammals ». *Proceedings of the National Academy of Sciences* 104 (29): 12046-51. https://doi.org/10.1073/pnas.0700531104.

Poirier, E.Z., Goic, B., Tomé-Poderti, L., Frangeul, L., Boussier, J., Gausson, V., Blanc, H., Vallet, T., Loyd, H., Levi, L.I., et al. 2018. Dicer-2-Dependent Generation of Viral DNA from Defective Genomes of RNA Viruses Modulates Antiviral Immunity in Insects. Cell Host Microbe *23*, 353-365.e8.

Quinlan, A.R., and Hall, I.M. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Quispe-Huamanquispe, Dora G., Godelieve Gheysen, et Jan F. Kreuze. 2017. « Horizontal Gene Transfer Contributes to Plant Evolution: The Case of Agrobacterium T-DNAs ». *Frontiers in Plant Science* 8 (novembre): 2015. https://doi.org/10.3389/fpls.2017.02015.

R Core Team. 2019. R: A language and environment for statistical computing. R Foundation for Statisical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rahman, Md.Masmudur, and Karumathil P Gopinathan. 2004. "Systemic and in Vitro Infection Process of Bombyx Mori Nucleopolyhedrovirus." *Virus Research* 101 (2): 109–18. https://doi.org/10.1016/j.virusres.2003.12.027.

Ramshani, Zeinab, Chenguang Zhang, Katherine Richards, Lulu Chen, Geyang Xu, Bangyan L. Stiles, Reginald Hill, Satyajyoti Senapati, David B. Go, et Hsueh-Chia Chang. 2019. « Extracellular Vesicle MicroRNA Quantification from Plasma Using an Integrated Microfluidic Device ». *Communications Biology* 2 (1): 189. https://doi.org/10.1038/s42003-019-0435-1.

Rausch, T., Zichner, T., Schlattl, A., Stutz, A.M., Benes, V., and Korbel, J.O. 2012. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics *28*, i333–i339.

Reiss, D., Mialdea, G., Miele, V., de Vienne, D.M., Peccoud, J., Gilbert, C., Duret, L., Charlat, S., 2019. Global survey of mobile DNA horizontal transfer in arthropods reveals Lepidoptera as a prime hotspot. PLoS Genet. 15, e1007965. https://doi.org/10.1371/journal.pgen.1007965

Renzette, N., Gibson, L., Bhattacharjee, B., Fisher, D., Schleiss, M.R., Jensen, J.D., and Kowalik, T.F. 2013. Rapid Intrahost Evolution of Human Cytomegalovirus Is Shaped by Demography and Positive Selection. PLoS Genet. *9*, e1003735.

Renzette, N., Pfeifer, S.P., Matuszewski, S., Kowalik, T.F., and Jensen, J.D. 2017. On the Analysis of Intrahost and Interhost Viral Populations: Human Cytomegalovirus as a Case Study of Pitfalls and Expectations. J. Virol. *91*.

Renzette, N., Pokalyuk, C., Gibson, L., Bhattacharjee, B., Schleiss, M.R., Hamprecht, K., Yamamoto, A.Y., Mussi-Pinhata, M.M., Britt, W.J., Jensen, J.D., et al. 2015. Limits and patterns of cytomegalovirus genomic diversity in humans. Proc. Natl. Acad. Sci. *112*, E4120–E4128.

Rognes, Torbjørn, Tomáš Flouri, Ben Nichols, Christopher Quince, and Frédéric Mahé. 2016. "VSEARCH: A Versatile Open Source Tool for Metagenomics." *PeerJ* 4 (October): e2584. https://doi.org/10.7717/peerj.2584.

Rohrmann, G. F. 2011. Baculovirus Molecular Biology 2nd edn. NCBI.

Rohrmann, G.F. 2014. Baculovirus nucleocapsid aggregation (MNPV vs SNPV): an evolutionary strategy, or a product of replication conditions? Virus Genes *49*, 351–357.

Rohrmann, G.F., 2019. Baculovirus Molecular Biology, 4th ed. National Center for Biotechnology Information (US), Bethesda (MD).

Romero-Soriano, Valèria, and Maria Pilar Garcia Guerreiro. 2016. "Expression of the Retrotransposon Helena Reveals a Complex Pattern of TE Deregulation in Drosophila Hybrids." Edited by Pawel Michalak. *PLOS ONE* 11 (1): e0147903. https://doi.org/10.1371/journal.pone.0147903.

Routh, A., Head, S.R., Ordoukhanian, P., and Johnson, J.E. 2015. ClickSeq: Fragmentation-Free Next-Generation Sequencing via Click Ligation of Adaptors to Stochastically Terminated 3′-Azido cDNAs. J. Mol. Biol. *427*, 2610–2616.

Routh, A., T. Domitrovic, and J. E. Johnson. 2012. "Host RNAs, Including Transposons, Are Encapsidated by a Eukaryotic Single-Stranded RNA Virus." *Proceedings of the National Academy of Sciences* 109 (6): 1907–12. https://doi.org/10.1073/pnas.1116168109.

Roy, Marlène, Barbara Viginier, Édouard Saint-Michel, Frédérick Arnaud, Maxime Ratinier, and Marie Fablet. 2020. "Viral Infection Impacts Transposable Element Transcript Amounts in *Drosophila*." *Proceedings of the National Academy of Sciences* 117 (22): 12249–57. https://doi.org/10.1073/pnas.2006106117.

Ryan, Calen P., Jeremy C. Brownlie, and Steve Whyard. 2017. "*Hsp90 and Physiological Stress Are Linked to Autonomous Transposon Mobility and Heritable Genetic Change in Nematodes*." *Genome Biology and Evolution*, January, evw284. https://doi.org/10.1093/gbe/evw284.

Saksouk, Nehmé, Elisabeth Simboeck, and Jérôme Déjardin. 2015. "Constitutive Heterochromatin Formation and Transcription in Mammals." *Epigenetics & Chromatin* 8 (1): 3. https://doi.org/10.1186/1756-8935-8-3.

Sambrook, J., H. Westphal, P. R. Srinivasan, and R. Dulbecco. 1968. The integrated state of viral DNA in SV40-transformed cells. Proc Natl Acad Sci U S A **60**:1288-1295.

Sampaio, Natalia Guimaraes, Lesley Cheng, et Emily M. Eriksson. 2017. « The Role of Extracellular Vesicles in Malaria Biology and Pathogenesis ». *Malaria Journal* 16 (1): 245. https://doi.org/10.1186/s12936-017-1891-z.

Sanjuán, R., and Domingo-Calap, P. 2016. Mechanisms of viral mutation. Cell. Mol. Life Sci. *73*, 4433–4448.

Sasani, Thomas A, Kelsey R Cone, Aaron R Quinlan, et Nels C Elde. 2018. « Long Read Sequencing Reveals Poxvirus Evolution through Rapid Homogenization of Gene Arrays ». *ELife* 7 (août): e35453. https://doi.org/10.7554/eLife.35453.

Schaack, S., Gilbert, C., Feschotte, C., 2010. Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. Trends Ecol. Evol. 25, 537–546. https://doi.org/10.1016/j.tree.2010.06.001

Schnable, P. S., D. Ware, R. S. Fulton, J. C. Stein, F. Wei, S. Pasternak, C. Liang, et al. 2009. "The B73 Maize Genome: Complexity, Diversity, and Dynamics." *Science* 326 (5956): 1112–15. https://doi.org/10.1126/science.1178534.

Schneider, Sean E., et James H. Thomas. 2014. « Accidental Genetic Engineers: Horizontal Sequence Transfer from Parasitoid Wasps to Their Lepidopteran Hosts ». Édité par Richard Cordaux. *PLoS ONE* 9 (10): e109446. https://doi.org/10.1371/journal.pone.0109446.

Sedlazeck, F.J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A., and Schatz, M.C. 2018. Accurate detection of complex structural variations using single-molecule sequencing. Nat. Methods *15*, 461–468.

Sfanos, K. S. *et al.* 2011. Identification of Replication Competent Murine Gammaretroviruses in Commonly Used Prostate Cancer Cell Lines. *PLoS ONE* **6**, e20874.

Sheehy, A. M., Gaddis, N. C., Choi, J. D. & Malim, M. H. 2002. Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. *Nature* **418**, 646–650.

Shioda, S. *et al.* 2018. Screening for 15 pathogenic viruses in human cell lines registered at the JCRB Cell Bank: characterization of *in vitro* human cells by viral infection. *R. Soc. Open Sci.* **5**, 172472.

Shrestha, Anita, Kan Bao, Wenbo Chen, Ping Wang, Zhangjun Fei, and Gary W. Blissard. 2019. "Transcriptional Responses of the *Trichoplusia Ni* Midgut to Oral Infection by the Baculovirus Autographa Californica Multiple Nucleopolyhedrovirus." Edited by Joanna L. Shisler. *Journal of Virology* 93 (14): e00353-19, /jvi/93/14/JVI.00353-19.atom. https://doi.org/10.1128/JVI.00353-19.

Shrestha, Anita, Kan Bao, Yun-Ru Chen, Wenbo Chen, Ping Wang, Zhangjun Fei, and Gary W. Blissard. 2018. "Global Analysis of Baculovirus Autographa Californica Multiple Nucleopolyhedrovirus Gene Expression in the Midgut of the Lepidopteran Host *Trichoplusia Ni*." Edited by Joanna L. Shisler. *Journal of Virology* 92 (23): e01277-18, /jvi/92/23/e01277-18.atom. https://doi.org/10.1128/JVI.01277-18.

Sijmons, S., Thys, K., Mbong Ngwese, M., Van Damme, E., Dvorak, J., Van Loock, M., Li, G., Tachezy, R., Busson, L., Aerssens, J., et al. 2015. High-Throughput Analysis of Human Cytomegalovirus Genome Diversity Highlights the Widespread Occurrence of Gene-Disrupting Mutations and Pervasive Recombination. J. Virol. *89*, 7673–7695.

Silva, J.C., Loreto, E.L., Clark, J.B., 2004. Factors that affect the horizontal transfer of transposable elements. Curr. Issues Mol. Biol. 6, 57–71.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., Zdobnov, E.M., 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinforma. Oxf. Engl. 31, 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Simón, O., Williams, T., Caballero, P., and López-Ferber, M. 2006. Dynamics of deletion genotypes in an experimental insect virus population. Proc. R. Soc. B Biol. Sci. *273*, 783–790.

Sinzelle, L., V. V. Kapitonov, D. P. Grzela, T. Jursch, J. Jurka, Z. Izsvak, and Z. Ivics. 2008. "Transposition of a Reconstructed Harbinger Element in Human Cells and Functional Homology with Two Transposon-Derived Cellular Genes." *Proceedings of the National Academy of Sciences* 105 (12): 4715–20. https://doi.org/10.1073/pnas.0707746105.

Skog, Johan, Tom Würdinger, Sjoerd van Rijn, Dimphna H. Meijer, Laura Gainche, William T. Curry, Bob S. Carter, Anna M. Krichevsky, et Xandra O. Breakefield. 2008. « Glioblastoma Microvesicles Transport RNA and Proteins That Promote Tumour Growth and Provide Diagnostic Biomarkers ». *Nature Cell Biology* 10 (12): 1470-76. https://doi.org/10.1038/ncb1800.

Slack, J., and Arif, B.M. 2006. The Baculoviruses Occlusion-Derived Virus: Virion Structure and Function. In Advances in Virus Research, (Elsevier), pp. 99–165.

Sliva, K., Erlwein, O., Bittner, A. & Schnierle, B. S. 2004. Murine leukemia virus (MLV) replication monitored with fluorescent proteins. *Virol. J.* **1**, 14.

Slotkin, R. Keith, et Robert Martienssen. 2007. « Transposable Elements and the Epigenetic Regulation of the Genome ». *Nature Reviews Genetics* 8 (4): 272-85. https://doi.org/10.1038/nrg2072.

Sommer, D.D., Delcher, A.L., Salzberg, S.L., and Pop, M. 2007. Minimus: a fast, lightweight genome assembler. BMC Bioinformatics *8*, 64.

Song, Michael J., and Sarah Schaack. 2018. "Evolutionary Conflict between Mobile DNA and Host Genomes." *The American Naturalist* 192 (2): 263–73. https://doi.org/10.1086/698482.

Sotero-Caio, Cibele G., Roy N. Platt, Alexander Suh, and David A. Ray. 2017. "Evolution and Diversity of Transposable Elements in Vertebrate Genomes." *Genome Biology and Evolution* 9 (1): 161–77. https://doi.org/10.1093/gbe/evw264.

Sparks, W., Li, H., Bonning, B., 2008. Protocols for Oral Infection of Lepidopteran Larvae with Baculovirus. J. Vis. Exp. JoVE. https://doi.org/10.3791/888

Stöcker, B.K., Köster, J., and Rahmann, S. 2016. SimLoRD: Simulation of Long Read Data. Bioinformatics *32*, 2704–2706.

Strand, Michael R, et Gaelen R Burke. 2013. « Polydnavirus-Wasp Associations: Evolution, Genome Organization, and Function ». *Current Opinion in Virology* 3 (5): 587-94. https://doi.org/10.1016/j.coviro.2013.06.004.

Suh, A., Witt, C.C., Menger, J., Sadanandan, K.R., Podsiadlowski, L., Gerth, M., Weigert, A., McGuire, J.A., Mudge, J., Edwards, S.V., Rheindt, F.E., 2016. Ancient horizontal transfers of retrotransposons between birds and ancestors of human pathogenic nematodes. Nat. Commun. 7, 11396. https://doi.org/10.1038/ncomms11396

Sultana, T., Zamborlini, A., Cristofari, G. & Lesage, P. 2017. Integration site selection by retroviruses and transposable elements in eukaryotes. *Nat. Rev. Genet.* **18**, 292–308.

Sun, C., Feschotte, C., Wu, Z., Mueller, R.L., 2015. DNA transposons have colonized the genome of the giant virus Pandoravirus salinus. BMC Biol. 13, 38. https://doi.org/10.1186/s12915-015-0145-1

Sun, Dongchang. 2018. « Pull in and Push Out: Mechanisms of Horizontal Gene Transfer in Bacteria ». *Frontiers in Microbiology* 9 (septembre): 2154. https://doi.org/10.3389/fmicb.2018.02154.

Szak, S.T., Pickeral, O.K., Makalowski, W., Boguski, M.S., Landsman, D., Boeke, J.D., 2002. Molecular archeology of L1 insertions in the human genome. Genome Biol. 3, research0052. https://doi.org/10.1186/gb-2002-3-10-research0052

Szitenberg, Amir, Soyeon Cha, Charles H. Opperman, David M. Bird, Mark L. Blaxter, et David H. Lunt. 2016. « Genetic Drift, Not Life History or RNAi, Determine Long-Term Evolution of Transposable Elements ». *Genome Biology and Evolution* 8 (9): 2964-78. https://doi.org/10.1093/gbe/evw208.

Szpara, M.L., Gatherer, D., Ochoa, A., Greenbaum, B., Dolan, A., Bowden, R.J., Enquist, L.W., Legendre, M., and Davison, A.J. 2014. Evolution and Diversity in Human Herpes Simplex Virus Genomes. J. Virol. *88*, 1209–1227.

Takeuchi, Y., McClure, M. O. & Pizzato, M. 2008. Identification of Gammaretroviruses Constitutively Released from Cell Lines Used for Human Immunodeficiency Virus Research. *J. Virol.* **82**, 12585–12588.

Tallon, L.J., Liu, X., Bennuru, S., Chibucos, M.C., Godinez, A., Ott, S., Zhao, X., Sadzewicz, L., Fraser, C.M., Nutman, T.B., et al. 2014. Single molecule sequencing and genome assembly of a clinical specimen of Loa loa, the causative agent of loiasis. BMC Genomics *15*, 788.

Talsania, K., Mehta, M., Raley, C., Kriga, Y., Gowda, S., Grose, C., Drew, M., Roberts, V., Cheng, K.T., Burkett, S., Oeser, S., Stephens, R., Soppet, D., Chen, X., Kumar, P., German, O., Smirnova, T., Hautman, C., Shetty, J., Tran, B., Zhao, Y., Esposito, D., 2019. Genome Assembly and Annotation of the Trichoplusia ni Tni-FNL Insect Cell Line Enabled by Long-Read Technologies. Genes 10, 79. https://doi.org/10.3390/genes10020079

Tao, X. Y., J. Y. Choi, W. J. Kim, J. H. Lee, Q. Liu, S. E. Kim, S. B. An, et al. 2013. "The Autographa Californica Multiple Nucleopolyhedrovirus ORF78 Is Essential for Budded Virus Production and General Occlusion Body Formation." *Journal of Virology* 87 (15): 8441–50. https://doi.org/10.1128/JVI.01290-13.

Tcherepanov, V., Ehlers, A., Upton, C., 2006. Genome Annotation Transfer Utility (GATU): rapid annotation of viral genomes using a closely related reference genome. BMC Genomics 7, 150. https://doi.org/10.1186/1471-2164-7-150

Telesnitsky, A., Wolin, S.L., 2016. The Host RNAs in Retroviral Particles. Viruses 8. https://doi.org/10.3390/v8080235

Temin, H. M. 1964. Homology between Rna from Rous Sarcoma Virous and DNA from Rous Sarcoma Virus-Infected Cells. Proc Natl Acad Sci U S A **52**:323-329.

Théry, Clotilde, Matias Ostrowski, et Elodie Segura. 2009. « Membrane Vesicles as Conveyors of Immune Responses ». *Nature Reviews Immunology* 9 (8): 581-93. https://doi.org/10.1038/nri2567.

Theze, J., A. Bezier, G. Periquet, J.-M. Drezen, et E. A. Herniou. 2011. « Paleozoic Origin of Insect Large DsDNA Viruses ». *Proceedings of the National Academy of Sciences* 108 (38): 15931-35. https://doi.org/10.1073/pnas.1105580108.

Thézé, Julien, Jun Takatsuka, Madoka Nakai, Basil Arif, et Elisabeth Herniou. 2015. « Gene Acquisition Convergence between Entomopoxviruses and Baculoviruses ». *Viruses* 7 (4): 1960-74. https://doi.org/10.3390/v7041960.

Thomas, Christopher M., et Kaare M. Nielsen. 2005. « Mechanisms of, and Barriers to, Horizontal Gene Transfer between Bacteria ». *Nature Reviews Microbiology* 3 (9): 711-21. https://doi.org/10.1038/nrmicro1234.

Toolan, H. W. 1954. Transplantable human neoplasms maintained in cortisone-treated laboratory animals: H.S. No. 1; H.Ep. No. 1; H.Ep. No. 2; H.Ep. No. 3; and H.Emb.Rh. No. 1. *Cancer Res.* **14**, 660–666.

Toussaint, E.F.A., Condamine, F.L., Kergoat, G.J., Capdevielle-Dulac, C., Barbut, J., Silvain, J.-F., and Le Ru, B.P. 2012. Palaeoenvironmental Shifts Drove the Adaptive Radiation of a Noctuid Stemborer Tribe (Lepidoptera, Noctuidae, Apameini) in the Miocene. PLoS ONE *7*, e41377.

Trajkovic, K., C. Hsu, S. Chiantia, L. Rajendran, D. Wenzel, F. Wieland, P. Schwille, B. Brugger, et M. Simons. 2008. « Ceramide Triggers Budding of Exosome Vesicles into Multivesicular Endosomes ». *Science* 319 (5867): 1244-47. https://doi.org/10.1126/science.1153124.

Treangen, T.J., Sommer, D.D., Angly, F.E., Koren, S., and Pop, M. 2011. Next Generation Sequence Assembly with AMOS. Curr. Protoc. Bioinformatics *33*, 11.8.

Trojer, Patrick, and Danny Reinberg. 2007. "Facultative Heterochromatin: Is There a Distinctive Molecular Signature?" *Molecular Cell* 28 (1): 1–13. https://doi.org/10.1016/j.molcel.2007.09.011.

Tsai, I.J., Hunt, M., Holroyd, N., Huckvale, T., Berriman, M., and Kikuchi, T. (2014). Summarizing Specific Profiles in Illumina Sequencing from Whole-Genome Amplified DNA. DNA Res. *21*, 243–254.

Tzfira, Tzvi, et Vitaly Citovsky. 2006. « Agrobacterium-Mediated Genetic Transformation of Plants: Biology and Biotechnology ». *Current Opinion in Biotechnology* 17 (2): 147-54. https://doi.org/10.1016/j.copbio.2006.01.009.

Uphoff, C. C., Denkmann, S. A., Steube, K. G. & Drexler, H. G. 2010. Detection of EBV, HBV, HCV, HIV-1, HTLV-I and -II, and SMRV in Human and Other Primate Cell Lines. *J. Biomed. Biotechnol.* **2010**, 1–23.

Uphoff, C. C., Lange, S., Denkmann, S. A., Garritsen, H. S. P. & Drexler, H. G. 2015. Prevalence and Characterization of Murine Leukemia Virus Contamination in Human Cell Lines. *PLOS ONE* **10**, e0125622.

Vagner, Tatyana, Cristiana Spinelli, Valentina R. Minciacchi, Leonora Balaj, Mandana Zandian, Andrew Conley, Andries Zijlstra, et al. 2018. « Large Extracellular Vesicles Carry Most of the Tumour DNA Circulating in Prostate Cancer Patient Plasma ». *Journal of Extracellular Vesicles* 7 (1): 1505403. https://doi.org/10.1080/20013078.2018.1505403.

Valadi, Hadi, Karin Ekström, Apostolos Bossios, Margareta Sjöstrand, James J Lee, et Jan O Lötvall. 2007. « Exosome-Mediated Transfer of MRNAs and MicroRNAs Is a Novel Mechanism of Genetic Exchange between Cells ». *Nature Cell Biology* 9 (6): 654-59. https://doi.org/10.1038/ncb1596.

Van Meter, Michael, Mehr Kashyap, Sarallah Rezazadeh, Anthony J. Geneva, Timothy D. Morello, Andrei Seluanov, and Vera Gorbunova. 2014. “SIRT6 Represses LINE1 Retrotransposons by Ribosylating KAP1 but This Repression Fails with Stress and Age.” *Nature Communications* 5 (1): 5011. https://doi.org/10.1038/ncomms6011.

van Niel, G., D’Angelo, G., Raposo, G., 2018. Shedding light on the cell biology of extracellular vesicles. Nat. Rev. Mol. Cell Biol. 19, 213–228. https://doi.org/10.1038/nrm.2017.125

van Oers, M., and Vlak, J. 2007. Baculovirus Genomics. Curr. Drug Targets *8*, 1051–1068.

Vancaester, Emmelien, Thomas Depuydt, Cristina Maria Osuna-Cruz, et Klaas Vandepoele. 2020. « Systematic and Functional Analysis of Horizontal Gene Transfer Events in Diatoms ». Preprint. Evolutionary Biology. https://doi.org/10.1101/2020.01.24.918219.

Vasilijevic, J., Zamarreño, N., Oliveros, J.C., Rodriguez-Frandsen, A., Gómez, G., Rodriguez, G., Pérez-Ruiz, M., Rey, S., Barba, I., Pozo, F., et al. 2017. Reduced accumulation of defective viral genomes contributes to severe outcome in influenza virus infected patients. PLOS Pathog. *13*, e1006650.

Venner, Samuel, Vincent Miele, Christophe Terzian, Christian Biémont, Vincent Daubin, Cédric Feschotte, and Dominique Pontier. 2017. “Ecological Networks to Unravel the Routes to Horizontal Transposon Transfers.” *PLOS Biology* 15 (2): e2001536. https://doi.org/10.1371/journal.pbio.2001536.

Volff, Jean-Nicolas. 2006. “Turning Junk into Gold: Domestication of Transposable Elements and the Creation of New Genes in Eukaryotes.” *BioEssays* 28 (9): 913–22. https://doi.org/10.1002/bies.20452.

Voronova A, Belevich V, Jansons A, Rungis D. 2014. Stress-induced transcriptional activation of retrotransposon-like sequences in the Scots pine (*Pinus sylvestris L.*) genome. Tree Genet.

Wagih, Omar. 2017. “Ggseqlogo: A Versatile R Package for Drawing Sequence Logos.” Edited by John Hancock. *Bioinformatics* 33 (22): 3645–47. https://doi.org/10.1093/bioinformatics/btx469.

Wagner, Alexander, Rachel J. Whitaker, David J. Krause, Jan-Hendrik Heilers, Marleen van Wolferen, Chris van der Does, et Sonja-Verena Albers. 2017. « Mechanisms of Gene Flow in Archaea ». *Nature Reviews Microbiology* 15 (8): 492-501. https://doi.org/10.1038/nrmicro.2017.41.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., et al. 2014. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLoS ONE *9*, e112963.

Wallau, G.L., Ortiz, M.F., Loreto, E.L.S., 2012. Horizontal Transposon Transfer in Eukarya: Detection, Bias, and Perspectives. Genome Biol. Evol. 4, 801–811. https://doi.org/10.1093/gbe/evs055

Walsh, A. M., R. D. Kortschak, M. G. Gardner, T. Bertozzi, and D. L. Adelson. 2013. “Widespread Horizontal Transfer of Retrotransposons.” *Proceedings of the National Academy of Sciences* 110 (3): 1012–16. https://doi.org/10.1073/pnas.1205856110.

Wang, Hwei-gene Heidi, et M. J. Fraser. 1993. « TTAA Serves as the Target Site for TFP3 Lepidopteran Transposon Insertions in Both Nuclear Polyhedrosis Virus and *Trichoplusia Ni* Genomes ». *Insect Molecular Biology* 1 (3): 109-16. https://doi.org/10.1111/j.1365-2583.1993.tb00111.x.

Wang, X., Meyers, C., Wang, H.-K., Chow, L. T. & Zheng, Z.-M. 2011. Construction of a Full Transcription Map of Human Papillomavirus Type 18 during Productive Viral Infection. *J. Virol.* **85**, 8080–8092.

Ward, J.H. 1963. Hierarchical Grouping to Optimize an Objective Function. J. Am. Stat. Assoc. *58*, 236–244.

Waterhouse, R.M., Seppey, M., Simão, F.A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E.V., Zdobnov, E.M., 2018. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. Mol. Biol. Evol. 35, 543–548. https://doi.org/10.1093/molbev/msx319

Watson, J. D. et Crick, F. H. C. 1953. Molecular structure of nucleic acids. Nature 171, 737.

Wei, Z.-G., and Zhang, S.-W. 2018. NPBSS: a new PacBio sequencing simulator for generating the continuous long reads with an empirical model. BMC Bioinformatics *19*.

Wennmann, J.T., Fan, J., Jehle, J.A., 2020. Bacsnp: Using Single Nucleotide Polymorphism (SNP) Specificities and Frequencies to Identify Genotype Composition in Baculoviruses. Viruses 12, 625. https://doi.org/10.3390/v12060625

Wetterwald, C., T. Roth, M. Kaeslin, M. Annaheim, G. Wespi, M. Heller, P. Maser, et al. 2010. « Identification of Bracovirus Particle Proteins and Analysis of Their Transcript Levels at the Stage of Virion Formation ». *Journal of General Virology* 91 (10): 2610-19. https://doi.org/10.1099/vir.0.022699-0.

White, F. F., Garfinkel, D. J., Huffman, G. A., Gordon, M. P., and Nester, E. W. 1983. Sequences homologous to Agrobacterium rhizogenes T-DNA in the genomes of uninfected plants. Nature 301, 348–350. doi: 10.1038/30 1348a0

Wicker, Thomas, François Sabot, Aurélie Hua-Van, Jeffrey L. Bennetzen, Pierre Capy, Boulos Chalhoub, Andrew Flavell, et al. 2007. « A Unified Classification System for Eukaryotic Transposable Elements ». *Nature Reviews Genetics* 8 (12): 973-82. https://doi.org/10.1038/nrg2165.

Wiedenbeck, Jane, et Frederick M. Cohan. 2011. « Origins of Bacterial Diversity through Horizontal Genetic Transfer and Adaptation to New Ecological Niches ». *FEMS Microbiology Reviews* 35 (5): 957-76. https://doi.org/10.1111/j.1574-6976.2011.00292.x.

Williams, T., Barbosa-Solomieu, V., Chinchar, V.G., 2005. A Decade of Advances in Iridovirus Research, in: Advances in Virus Research. Elsevier, pp. 173–248. https://doi.org/10.1016/S0065-3527(05)65006-3

Wong, K., Keane, T.M., Stalker, J., and Adams, D.J. 2010. Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. Genome Biol. *11*, R128.

Wood, H.A. 1980. "Isolation and Replication of an Occlusion Body-Deficient Mutant of the Autographa Californica Nuclear Polyhedrosis Virus." *Virology* 105 (2): 338–44. https://doi.org/10.1016/0042-6822(80)90035-5.

Wu, Zhenyu, Lingling Wang, Jiaying Li, Lifu Wang, Zhongdao Wu, et Xi Sun. 2019. « Extracellular Vesicle-Mediated Communication Within Host-Parasite Interactions ». *Frontiers in Immunology* 9 (janvier): 3066. https://doi.org/10.3389/fimmu.2018.03066.

Xiaofei, E., and Kowalik, T. 2014. The DNA Damage Response Induced by Infection with Human Cytomegalovirus and Other Viruses. Viruses *6*, 2155–2185.

Yamao, M., N. Katayama, H. Nakazawa, M. Yamakawa, Y. Hayashi, S. Hara, K. Kamei, et H. Mori. 1999. « Gene Targeting in the Silkworm by Use of a Baculovirus ». *Genes & Development* 13 (5): 511-16. https://doi.org/10.1101/gad.13.5.511.

Yan, Y. *et al.* 2010. Evolution of Functional and Sequence Variants of the Mammalian XPR1 Receptor for Mouse Xenotropic Gammaretroviruses and the Human-Derived Retrovirus XMRV. *J. Virol.* **84**, 11970–11980.

Ye, K., Schulz, M.H., Long, Q., Apweiler, R., and Ning, Z. 2009. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics *25*, 2865–2871.

Yoder, J. A., Walsh, C. P. & Bestor, T. H. 1997. Cytosine methylation and the ecology of intragenomic parasites. Trends Genet. 13: 335–340.

Yoshikawa, Fabio Seiti Yamada, Franciane Mouradian Emidio Teixeira, Maria Notomi Sato, et Luanda Mara da Silva Oliveira. 2019. « Delivery of MicroRNAs by Extracellular Vesicles in Viral Infections: Could the News Be Packaged? » *Cells* 8 (6): 611. https://doi.org/10.3390/cells8060611.

Zarate, S., Carroll, A., Krasheninina, O., Sedlazeck, F.J., Jun, G., Salerno, W., Boerwinkle, E., and Gibbs, R. 2018. Parliament2: Fast Structural Variant Calling Using Optimized Combinations of Callers. BioRxiv.

Zhang, H.-H., Peccoud, J., Xu, M.-R.-X., Zhang, X.-G., Gilbert, C., 2020. Horizontal transfer and evolution of transposable elements in vertebrates. Nat. Commun. 11, 1362. https://doi.org/10.1038/s41467-020-15149-4

Zhang, Lin, Siyuan Zhang, Jun Yao, Frank J. Lowery, Qingling Zhang, Wen-Chien Huang, Ping Li, et al. 2015. « Microenvironment-Induced PTEN Loss by Exosomal MicroRNA Primes Brain Metastasis Outgrowth ». *Nature* 527 (7576): 100-104. https://doi.org/10.1038/nature15376.

Zhang, W., Jia, B., and Wei, C. 2019. PaSS: a sequencing simulator for PacBio sequencing. BMC Bioinformatics *20*.

Zhang, Y., Harris, C.J., Liu, Q., Liu, W., Ausin, I., Long, Y., Xiao, L., Feng, L., Chen, X., Xie, Y., et al. 2018. Large-scale comparative epigenomics reveals hierarchical regulation of non-CG methylation in *Arabidopsis*. Proc. Natl. Acad. Sci. *115*, E1069–E1074.

Zhang, Y.-A. *et al.* 2011. Frequent detection of infectious xenotropic murine leukemia virus (XMLV) in human cultures established from mouse xenografts. *Cancer Biol. Ther.* **12**, 617–628.

Zhang, Yi, Chil-Woo Lee, Nora Wehner, Fabian Imdahl, Veselova Svetlana, Christoph Weiste, Wolfgang Dröge-Laser, et Rosalia Deeken. 2015. « Regulation of Oncogene Expression in T-DNA-Transformed Host Plant Cells ». Édité par Darrell Desveaux. *PLOS Pathogens* 11 (1): e1004620. https://doi.org/10.1371/journal.ppat.1004620.

Zhao, X. & Yoshimura, F. K. 2008. Expression of Murine Leukemia Virus Envelope Protein Is Sufficient for the Induction of Apoptosis. *J. Virol.* **82**, 2586–2589.

Zilberman, Daniel, Mary Gehring, Robert K Tran, Tracy Ballinger, and Steven Henikoff. 2007. "Genome-Wide Analysis of Arabidopsis Thaliana DNA Methylation Uncovers an Interdependence between Methylation and Transcription." *Nature Genetics* 39 (1): 61–69. https://doi.org/10.1038/ng1929.

Zimin, A. V. *et al.* 2013. The MaSuRCA genome assembler. *Bioinformatics* **29**, 2669–2677.

Zingg, D., Züger, M., Bollhalder, F., Andermatt, M., 2011. Use of resistance overcoming CpGV isolates and CpGV resistance situation of the codling moth in Europe seven years after the first discovery of resistance to CpGV-M 3.

Zomer, Anoek, Carrie Maynard, Frederik Johannes Verweij, Alwin Kamermans, Ronny Schäfer, Evelyne Beerling, Raymond Michel Schiffelers, et al. 2015. « In Vivo Imaging Reveals Extracellular Vesicle-Mediated Phenocopying of Metastatic Behavior ». *Cell* 161 (5): 1046-57. https://doi.org/10.1016/j.cell.2015.04.042.

Zovoilis, Athanasios, Catherine Cifuentes-Rojas, Hsueh-Ping Chu, Alfredo J. Hernandez, and Jeannie T. Lee. 2016. "Destabilization of B2 RNA by EZH2 Activates the Stress Response." *Cell* 167 (7): 1788-1802.e13. https://doi.org/10.1016/j.cell.2016.11.041.

Zwart, M.P., Tromas, N., and Elena, S.F. 2013. Model-Selection-Based Approach for Calculating Cellular Multiplicity of Infection during Virus Colonization of Multi-Cellular Hosts. PLoS ONE *8*, e64657.

université PARIS-SACLAY
ÉCOLE DOCTORALE
Structure et dynamique des systèmes vivants (SDSV)

**Titre** : Etude de transferts horizontaux de matériel génétique entre virus et animaux

**Mots clefs** : transfert horizontal, virus, élément transposable, baculovirus, iridovirus

**Résumé** : Les transferts horizontaux (TH) d'ADN sont de plus en plus reconnus comme un phénomène important dans l'évolution des métazoaires. La grande majorité des TH entre animaux implique des éléments transposables (ET), séquences génomiques égoïstes capables de se déplacer par transposition dans le génome et générer ainsi de multiples copies. Si ces TH d'ET semblent avoir un rôle prépondérant dans l'évolution des métazoaires, les mécanismes sous-tendant ces TH restent mal connus. Un des mécanismes possibles implique les virus qui pourraient être des vecteurs d'ADN entre les hôtes qu'ils infectent. Cette thèse contribue à évaluer cette hypothèse.
Dans le premier Chapitre, nous avons étudié l'impact du stress provoqué par une infection virale sur l'activité des ET. Certains ET sont surexprimés au cours de l'infection, et certains sont aussi exprimés après leur insertion dans des génomes viraux.
Dans le deuxième Chapitre, nous avons élargi le spectre des systèmes hôte-virus connus à ce jour dont les ET de l'hôte transposent dans les génomes viraux. L'analyse de 11 de ces systèmes a permis de découvrir neuf nouveaux systèmes dont des ET hôtes sont retrouvés dans les génomes du virus. Cette étude nous a permis d'inférer la capacité des ET portés par les génomes viraux à transposer vers d'autres génomes viraux, ce qui constitue un des résultats majeurs de cette thèse.
Dans le troisième Chapitre, la construction d'un pipeline bioinformatique a permis de caractériser de nombreuses insertions complètes d'ET dans les génomes viraux, ainsi que la diversité des variants structuraux génomiques présents dans quatre populations de grands virus à ADN double-brin.
Dans le dernier Chapitre, l'intégration d'un génome complet de rétrovirus murin dans le génome d'une lignée cellulaire humaine a été caractérisée, fournissant des résultats supplémentaires sur les TH de virus dans l'hôte.
Dans l'ensemble ces travaux apportent un éclairage nouveau sur le rôle des virus dans les TH d'ADN chez les métazoaires.

**Title** : Study of horizontal transfers of genetic material between viruses and animals

**Keywords** : horizontal transfer, virus, transposable element, baculovirus, iridovirus

**Abstract** : Horizontal transfers (HTs) of DNA are increasingly recognized as an important phenomenon in the evolution of metazoans. The vast majority of HTs between animals involves transposable elements (TEs), selfish genomic sequences capable of moving by transposition in the genome and thus generating multiple copies. While these HTs of TEs seem to have a major role in the evolution of metazoans, the mechanisms underlying these HTs remain poorly understood. One possible mechanism involves viruses which could be vectors of DNA between the hosts they infect. This thesis helps to evaluate this hypothesis.
In the first chapter, we studied the impact of stress caused by a viral infection on the activity of TEs. Some TEs are overexpressed during infection, and some are also expressed after their insertion into viral genomes.
In the second chapter, we have broadened the spectrum of host-virus systems known to date, in which host TEs transpose into viral genomes. Analysis of 11 of these systems revealed nine new systems in which host TEs were found in the genomes of the virus. This study allowed us to infer the capacity of TEs carried by viral genomes to transpose to other viral genomes, which constitutes one of the major results of this thesis.
In the third chapter, the construction of a bioinformatics pipeline has made it possible to characterize many complete TE insertions in viral genomes, as well as the diversity of structural genomic variants present in four populations of large double-stranded DNA viruses.
In the final chapter, the integration of a complete murine retrovirus genome into the genome of a human cell line was characterized, providing additional results on the virus HTs in the host.
Overall, this work sheds new light on the role of viruses in DNA HTs in metazoans.

**Université Paris-Saclay**
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France