Mechanisms and factors underlying horizontal transfers of genetic material between animals

Héloïse Muller

université PARIS-SACLAY

Mechanisms and factors underlying horizontal transfers of genetic material between animals

Mécanismes et facteurs impliqués dans les transferts horizontaux de matériel génétique entre animaux

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 577, Structure et dynamique des systèmes vivants (SDSV)
Spécialité de doctorat: Évolution
Graduate School : Life Sciences and Health
Référent : Faculté des sciences d'Orsay

Thèse préparée dans l'UMR **Évolution, Génomes, Comportement et Écologie (Université Paris-Saclay, CNRS, IRD)**,
sous la direction de Clément GILBERT (CR),
et la co-direction d'Élisabeth HUGUET (Professeure)

Thèse soutenue à Paris-Saclay, le 28 septembre 2023, par

Héloïse MULLER

Composition du jury

Membres du jury avec voix delibérative

Aurélie HUA-VAN Professeure, laboratoire Evolution, Génomes, Comportement, Ecologie, UMR CNRS 9191, Université Paris-Saclay	Présidente
Etienne DANCHIN DR INRAE, Institut Sophia Agrobiotech, UMR INRAE 1355, CNRS 7254, Univ. Côte d'Azur	Rapporteur & Examinateur
Gwenaël PIGANEAU DR CNRS, laboratoire Biologie Intégrative des Organismes Marins, UMR 7232 CNRS, Sorbonne Université, Banyuls sur Mer	Rapportrice & Examinatrice
Julien VARALDI MCF, laboratoire de biométrie et biologie évolutive, UMR CNRS 5558, Université Lyon 1	Examinateur
Cristina VIEIRA Professeure, laboratoire de biométrie et biologie évolutive, UMR CNRS 5558, Université Lyon 1	Examinatrice

THESE DE DOCTORAT
NNT : 2023UPASL076

université PARIS-SACLAY

ÉCOLE DOCTORALE
Structure et dynamique des systèmes vivants (SDSV)

Titre: Mécanismes et facteurs impliqués dans les transferts horizontaux de matériel génétique entre animaux

Mots clés: transferts horizontaux, virus, éléments transposables, insectes, relations hôtes-parasites, génomique

Résumé: Le transfert horizontal (TH) est la transmission de matériel génétique indépendamment de la reproduction, éventuellement entre espèces génétiquement éloignées. Chez les eucaryotes multicellulaires l'impact des THs sur leur évolution est mal connu ainsi que les mécanismes et facteurs impliqués. Cette thèse se concentre sur les animaux, en particulier les insectes, afin d'apporter des réponses à ces interrogations. Pour les mécanismes, j'ai étudié si les virus pourraient agir comme vecteurs de TH, en transportant du matériel génétique d'une espèce à une autre. Pour cela, j'ai étudié deux types de virus : un virus libre et des polydnavirus. Nous avons montré qu'une infection par un baculovirus (virus libre), impacte modérément l'activité des éléments transposables (ET) de l'hôte, ce qui pourrait augmenter les chances d'une transposition dans ce virus. Nous avons également montré qu'un ET inséré dans le virus était exprimé, et pourrait donc être capable de transposer de nouveau dans le génome d'une autre espèce lors d'une deuxième infection. Les polydnavirus, quant à eux, sont des virus domestiqués encodés dans le génome de guêpes parasitoïdes. Ces guêpes injectent les polydnavirus dans leurs hôtes, souvent des lépidoptères, en même temps que leurs œufs. Nous avons montré que ces polydnavirus s'intègrent massivement dans plusieurs tissus hôtes. Bien que la transmission par les hôtes survivants semble limitée dans notre système d'étude (*Cotesia typhae* [guêpe] –*Sesamia nonagrioides* [lépidoptère]), nous avons trouvé de nombreuses traces de transmission dans d'autres espèces de lépidoptères. Enfin, nous avons étudié quatre facteurs pouvant possiblement favoriser les THs, bien que ces résultats soient encore préliminaires : la proximité géographique, la proximité phylogénétique, l'habitat aquatique et le mode de fécondation.

Title: Mechanisms and factors underlying horizontal transfers of genetic material between animals

Keywords: horizontal transfers, viruses, transposable elements, hosts-parasites relationships, genomics

Abstract: Horizontal transfer (HT) is the transmission of genetic material by means other than reproduction, possibly between species which are genetically distant. In multicellular eukaryotes the impact of HT on their evolution is poorly understood, as well as the mechanisms and factors which are involved. In this thesis I focus on animals, with an emphasis on insects, to bring some insights on some of the mechanisms and factors that have been proposed. For mechanisms, I investigated whether viruses could act as vectors of HT, *i.e.* transport genetic material from one species to another. For this, I investigated two kinds of viruses: a free virus and polydnavirus. We showed that an infection by a baculovirus, a free virus, moderately impacts the activity of transposable elements (TE) of the host, which might increase the chance of a transposition in the virus during infections. We also showed that a TE that was inserted in the virus was expressed, which means that it might be able to transpose again in the genome of another species during a second infection. Regarding polydnaviruses, they are very particular viruses which are found in the genome of some parasitoid wasps that domesticated them. These parasitoids inject polydnaviruses in their host at the same time as their eggs. We showed that these polydnaviruses were able to integrate massively in several tissues of the hosts. Although the transmission to the next generation of surviving hosts seems quite limited in the system we investigated in detail (*Cotesia typhae* [wasp] – *Sesamia nonagrioides* [lepidoptera]), we found many traces of polydnavirus integrations in the genomes of many lepidopteran species, the main hosts of parasitoid wasps. Finally, we investigated four possible factors that might promote HT, although the results are still preliminary: the geographical proximity, the phylogenetic proximity, the aquatic habitat and the mode of fertilization.

Contents

Introduction 8

1 Preamble 8

2 Genetic sources of horizontal transfers 12
- 2.1 Transposable elements . . . 12
- 2.2 Polydnaviruses . . . 14
- 2.3 Other sources . . . 17

3 Putative mechanisms for horizontal transfers 20
- 3.1 Exiting the donor cell and reaching the recipient cell . . . 20
  - 3.1.1 Viruses as vectors of HT . . . 20
  - 3.1.2 Other intracellular parasites as vectors of HT . . . 23
  - 3.1.3 Extracellular vesicles as vectors of HT . . . 23
- 3.2 Reaching the nucleus . . . 24
- 3.3 Integrating the recipient genome . . . 25
- 3.4 Success of HT . . . 25

4 Putative factors promoting horizontal transfers 27

5 Methods to detect horizontal transfers and limitations 29
- 5.1 The problem of contamination . . . 29
- 5.2 Methods to identify HT . . . 30
- 5.3 The case of large-scale studies . . . 31
- 5.4 Dating HT events . . . 31
- 5.5 The case of *de novo* HT . . . 32
- 5.6 The direction of the transfer . . . 32

6 Goals and context of the thesis 34

I Free viruses as vectors of horizontal transfers 36

7 Article n°1: Assessing the Impact of a Viral Infection on the Expression of Transposable Elements in the Cabbage Looper Moth (*Trichoplusia ni*) 37

II Domesticated viruses as vectors of horizontal transfers from parasitoid wasps 55

8 Article n°2: Draft nuclear genome and complete mitogenome of the Mediterranean corn borer, *Sesamia nonagrioides*, a major pest of maize 56

9 Article n°3: Genome-Wide Patterns of Bracovirus Chromosomal Integration into Multiple Host Tissues during Parasitism 66

10 Article n°4: Investigating bracovirus chromosomal integration and inheritance in lepidopteran host and non-target species 89

11 Article n°5: Massive Somatic and Germline Chromosomal Integrations of Polydnaviruses in Lepidopterans 134

III Factors promoting horizontal transfers 158

12 Introduction 159

13 The aquatic habitat and the mode of fertilization as factors promoting HTT 161
- 13.1 Designing the dataset . . . 161
- 13.2 Phylogeny . . . 164
- 13.3 Core genes dS distribution . . . 164
- 13.4 TE annotation . . . 169
- 13.5 Similarity searches between TE copies . . . 171
- 13.6 Identification of TE-TE hits resulting from HT . . . 173
- 13.7 Refining hits . . . 173
- 13.8 Clustering . . . 174
- 13.9 Testing false positives . . . 176
- 13.10 Counting independent HTT events . . . 179
- 13.11 Statistical analysis . . . 181

14 The phylogenetic and geographical proximity as factors promoting HTT 184
- 14.1 Designing the dataset . . . 184
- 14.2 Genome sequencing and assembly . . . 187
- 14.3 Cleaning assemblies . . . 189
- 14.4 Check for contamination . . . 190
- 14.5 Taxonomy and phylogeny . . . 192
- 14.6 Rest of the pipeline . . . 195
- 14.7 Perspectives . . . 196

General discussion 198

15 Free viruses as vectors of horizontal transfers 198

16 Domesticated viruses as vectors of horizontal transfers from parasitoid wasps 199

17 Impact of horizontal transfers on evolution 201

Appendix 205

A Résumé détaillé en français 205
- A.1 Introduction . . . 205
  - A.1.1 Mécanismes lors d'un TH . . . 206
  - A.1.2 Facteurs influençant le nombre de TH . . . 208

A.1.3 Organisation de la thèse 209
A.2 Résultats et discussion 209
A.2.1 Partie I: Virus libres comme vecteurs de transferts horizontaux 209
A.2.2 Partie II: Virus domestiqués des guêpes parasitoïdes comme vecteurs de transferts horizontaux 210
A.2.3 Partie III: Facteurs influençant les transferts horizontaux 213
A.2.4 Impact évolutif des transferts horizontaux 213

Bibliography 215

Acknowledgements

I would firstly like to thank the members of my committee, Aurélie Hua-Van, Julien Varaldi, Cristina Vieira, and more specifically my two rapporteurs, Etienne Danchin and Gwenaël Pignaneau, who took the time to review my manuscript. I would also like to thank Cristina Vieira a second time for the L2 internship opportunity you offered me, opening the doors of research to me, and for your guidance during the rest of my formation.

To my thesis director, Clément Gilbert, I would like to express my gratitude for your mentorship, and for your support, but also for being so complaisant with my professional/personal life balance. This made my 3 years of PhD, 3 years and a half with the M2 internship, really fantastic. I really loved working with you, and I believe that the end of my PhD will not mark the end of our work together.

To my thesis co-director, Elisabeth Huguet, thank you for your precious discussions on polydnaviruses and for the numerous comments on my manuscript. I would also like to thank the other members of my "comité de suivie": Julien Varaldi and Richard Cordaux.

I would also like to thank Sylvain Charlat and Jean Peccoud with whom I collaborated on the large-scale study on the 247 genomes. Our regular meetings were very insightful and greatly improved the quality of this study. On the same project, I would also like to thank Antoine Fouquet for his expertise on Amphibian, identifying their habitat and their mode of fertilization.

I would like to thank all the members of the laboratory, and more specifically the members of pôle génome. I enjoyed our weekly lab meetings and our daily lunch breaks, firstly at the CNRS cafeteria and then on the Plateau. A big thank you for David Ogereau, without whom MinION sequencing and software installations would have been a real nightmare. I would also like to thank the team raising *C. typhae* and *S. nonagrioides* from pôle écologie (Claire Capdevielle, Taiadjana Fortuna, Rémi Jeannette, Laure Kaiser, and Florence Mougel), with whom I worked at the beginning of my PhD; I felt part of both pôles at that time.

To Lug and Toutatis, you are not perfect, but you work great most of the time, thank you for existing. Thank you to Jean-Bernard for taking care of them. I would also like to thank the other computing infrastructures I used, without which the study on the 247 genomes would not have been possible: genotoul, GenOuest, IFB, and ABiMS.

I would also like to thank Cheryl Andam and all her team, from the University at Albany (NY, USA), for welcoming me so warmly for 3 months in their laboratory.

Merci à mes amis pour ces soirées et week-ends qui m'ont permis de décompresser.

A ma famille, merci pour le soutien que vous m'avez apporté tout au long de ma vie. Un remerciement tout particulier à maman pour ses nombreuses relectures et corrections. Mamie, te rappelles-tu l'expérience que je décris au début de mon manuscrit ? (indice : le cours de génétique auquel je t'avais amenée à la fac).

Et enfin, à Hunter Belanger, merci d'avoir été à mes côtés tout au long de ma thèse, de m'avoir aidée en informatique, mais surtout de m'avoir tant soutenue. Tu as même réussi à rendre le travail en confinement super agréable.

Introduction

1 - Preamble

Genetic material is commonly transmitted from parents to offspring by vertical transmission (reproduction), following Mendelian genetics. However, in 1928, Griffith did an experiment in which mice infected by both a non-virulent bacteria and a dead virulent bacteria surprisingly died. In 1944, Avery and McCart understood that the DNA of the dead virulent bacteria had been transmitted to the non-virulent bacteria, becoming virulent. This was the first evidence of a horizontal DNA transfer (HT). Horizontal transfers can be defined as the transmission of genetic material by other means than reproduction, possibly between species which are genetically distant, as opposed to vertical transmission.

Studies have shown that HT are frequent and common in prokaryotes, in which about 81% of their genes would have been acquired horizontally at some point in their evolution (Dagan *et al.* 2008). In bacteria, three mechanisms leading to HT are clearly described: conjugation (transmission of DNA via their pilus, which requires close contact between the donor and the recipient cells), transformation (acquirement of foreign DNA from the environment), and transduction (transfer of DNA by the intermediate of a virus) (Thomas and Nielsen 2005). HT between bacteria have an important impact on their evolution, such as the transmission of genes implicated in antibiotic resistance (Gyles and Boerlin 2014; Kay *et al.* 2002). Numerous HT were also reported in archaea, including genes that likely provide a selective advantage for adaptation to new environments (Wagner *et al.* 2017). The same three mechanisms as in bacteria were identified in some archaeal species, but also mechanisms that rely on transfers via vesicles, cell fusion and other archaeal specific mechanisms (Wagner *et al.* 2017). It was shown that although most of the inheritance in prokaryotes is vertical, HT can be prominent between closely related species, and between distantly related species sharing a similar environment (Beiko *et al.* 2005). Horizontal transfer is thus a very important evolutionary force in prokaryotes.

The rapid improvements of sequencing tools allow researchers to report more and more HT, in prokaryotes, but also in eukaryotes. In eukaryotes, the distinction between unicellular and multicelluar organisms is important, since in addition to reaching the nucleus, HT DNA (horizontally transferred DNA) also need to reach specific cells in the latter in order to be transmitted to offspring. These cells are the germinal cells in sexual organisms, or the cells with the ability to dedifferentiate and/or regenerate to a functional organism in asexual organisms. All these barriers led researchers to firstly believe that HT was a prokaryote feature. However, it is now well known that unicelullar organisms are quite prone to HT, with for example 1% of the protist genes that originate from HT on average (Van Etten and Bhattacharya 2020).

In multicellular organisms, researchers keep discovering more and more examples of HT, despite the barriers which were initially thought to be insurmountable. Most examples of HT to multicellular organisms originate from bacteria or viruses and can sometimes bring important functional novelty. Some noteworthy examples of HT from bacteria are (i) a horizontal gene transfer (HGT) that enables an alga to survive hot, metal-rich and acidic environments (Schönknecht *et al.* 2013), (ii) a HGT that enables some insects to feed on plant tissues despite the production by these plants of toxic cyanide (Wybouw *et al.* 2014), (iii) a HGT that enables a tick to be protected against a pathogenic bacteria

(Chou *et al.* 2015), and (iv) a HGT bringing a key innovation in the evolution of the vertebrate eye (Kalluraya *et al.* 2023). The amount of bacterial genes in multicellular organisms can be quite high, like it is the case for *Drosophila ananassae* whose genome includes almost the entire genome of *Wolbachia* (Hotopp *et al.* 2007), or in the common pillbug *Armadillidium vulgare* which acquired a new W sex chromosome after the integration of the entire genome of *Wolbachia* (Leclercq *et al.* 2016). In addition, the evolution of endosymbionts into organelles was accompanied by massive transfers of genes to the host nucleus (Archibald 2015). While these examples demonstrate the existence of HT from bacteria to eukaryotes in the course of evolution, they do not inform about the pathways and mechanisms by which these sequences have been transferred. This information comes from well studied systems, mostly the bacteria *Agrobacterium*, but also *Escherichia coli* and *Rhizobium* species. *Agrobacterium* infects plants and transfers a segment of its DNA, called T-DNA, in its host, causing uncontrolled cell division, which results in crown galls or in proliferating roots (Lacroix and Citovsky 2016; Quispe-Huamanquispe *et al.* 2017), potentially leading to HGT (White *et al.* 1983; Intrieri and Buiatti 2001; Matveeva *et al.* 2012). *Agrobacterium* has been studied and used for years in an agronomic context, in order to transform plants to introduce genes of interest (Tzfira and Citovsky 2006; Gelvin 2009). Its mechanisms to transfer DNA to eukaryotic cells is the only one that has been demonstrated (see figure 1.1): the T-DNA leaves the bacterial cell thanks to the type IV secretion system, a conjugation-like mechanism, and then the T-DNA is imported to the plant nucleus thanks to interactions with host factors (Lacroix and Citovsky 2016). Experiments on cultivated cells of many species, from fungi to humans, showed that the host factors used by *Agrobacterium* are not specific of plant cells, rather they are found in diverse eukaryotic species (Lacroix and Citovsky 2016). However, the steps in between *i.e.* crossing the eukaryotic recipient cell wall and membrane, remains obscure.

Moreover, about 1800 HT from viruses (HVT) to multicellular organisms were reported in the database HVT-DB by 2023 (Dotto *et al.* 2018).The most famous example of HVT is probably the endogeneization of retroviruses, followed by the domestication of some *env* genes, allowing mammals to produce *env-like* proteins which are crucial mediators during the formation of the placenta. Other interesting examples of HVT are (i) the endogeneization of viruses in some parasitoid wasps allowing them to release viral particles in their host (see figure 2.2) (Bézier *et al.* 2009; Volkoff *et al.* 2010), and (ii) the endogeneization of a virus in some lepidoptera, allowing them to encode the protein *pfk*, a killing factor granting them a resistance against some parasitoid wasps (Gasmi *et al.* 2021). These three examples each took place several times independently in the course of evolution. A recent large-scale study performed a systematic evaluation of HT between eukaryotes and viruses, and they found that all lineages are impacted, although they confirmed that unicellular organisms are more involved (Irwin *et al.* 2022). They also showed that some groups of viruses transfer more often, mostly the double stranded DNA viruses, whose HT represent 97.7% of all the HT they reported. Guinet *et al.* (2023) analyzed 124 hymenopteran genomes to test whether the endoparatic lifestyle promotes endogenization and domestication of viruses. In addition to validating their hypothesis, they also found that double stranded viruses are more often endogenized than others, but also that they are more often domesticated. However, 45% of all HVT reported in HVT-DB are from type II viruss (ssDNA), and 37% are from type V viruses (-ssRNA).

At last, HT *between* multicellular organisms, *i.e.* *from* multicellular organisms *to* multicellular organisms, seem to be the transfers with the lowest probability to take place. Examples of horizontal

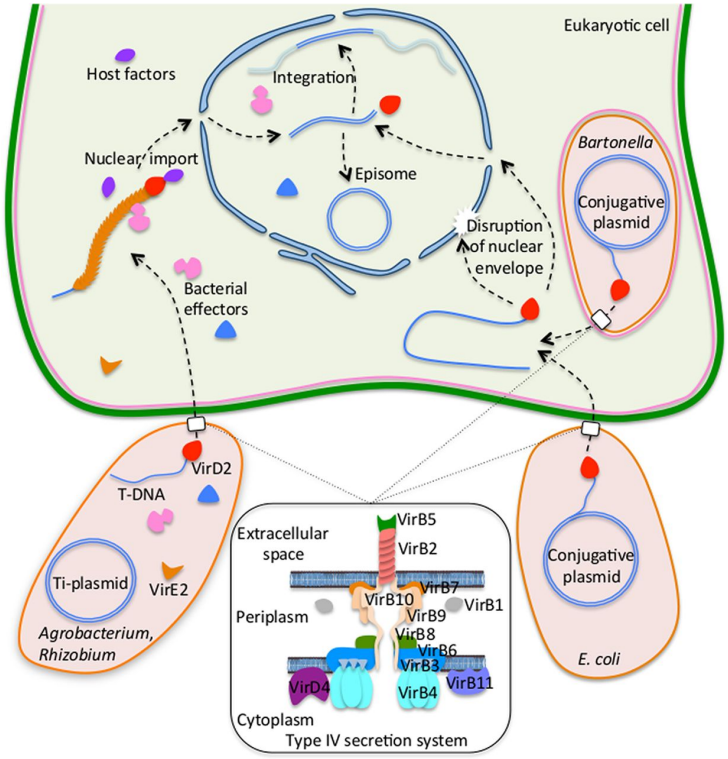


Figure 1.1: **Schematic summary of known natural and experimental pathways for DNA transfer from bacteria to eukaryotic cells**. *Agrobacterium* and related bacteria, *E. coli*, and *Bartonella henselae* can transfer DNA to different types of eukaryotic cells via the activity of their type IV secretion systems composed of VirD4/VirB proteins. Inside the host eukaryotic cell, the bacterial transferred DNA, usually a single-stranded molecule packaged into a nucleoprotein complex, is imported into the host nucleus. Nuclear import and further DNA processing, *i.e.*, conversion to a double-stranded form, integration into the recipient cell genome, or formation of an episome, depend on interactions of the transferred DNA and its associated proteins with numerous host cell factors that represent different types of cellular machineries, such as nuclear import machinery, the ubiquitin/proteasome system, and DNA repair machinery. Figure from Lacroix and Citovsky (2016).

gene transfers (HGT) between multicellular organisms are very anecdotal, but can obviously have major evolutionary impact (Moran and Jarvik 2010; Altincicek *et al.* 2012; Gasmi *et al.* 2015; Graham and Davies 2021). To my knowledge, there are only three large-scale studies that attempted to recover several HT between all multicellular species of a dataset, all focusing on HT of transposable elements (HTT). These three studies were performed on plants (Baidouri *et al.* 2014), insects (Peccoud *et al.* 2017) or vertebrates (Zhang *et al.* 2020), and they shed light not only on the fact that HTT are possible between multicellular organisms, but also on how numerous they are. Two other studies, performed at a smaller scale, were able to respectively estimate that the three species of analyzed *Drosophila* exchange TE at a frequency of 0.04 HT per TE family and per million years (Bartolomé *et al.* 2009), and that 24 of the 26 tested lines of *mariner* (a family of TE) are involved in HTT between 20 genomes of *Drosophia* of their dataset (Wallau *et al.* 2016). Additional large-scale studies are necessary to really understand the importance of HTT between multicellular organisms in evolution, looking at frequencies but also consequences. I will discuss more into details these studies, and also the putative mechanisms and factors underlying HTT, all along the introduction of this manuscript,

which will only deal with HT *between* multicellular organisms from now on.

Note: the drawings at the bottom of some pages are there just for a decorative purpose. I did these drawings following observations of insects I captured for my PhD or insects that were captured by the children at the MISS, where I did scientific mediation in my second year of PhD.

2 - Genetic sources of horizontal transfers

Genomes contain genes, but also many other sequences sometimes referred to as "junk" DNA. In this section, I go through several types of DNA sequences and discuss their potential as source of horizontal transfer, with an emphasis on the two sources I focused on over the course of my PhD: transposable elements and polydnaviruses.

2.1 . Transposable elements

A transposable element (TE) is a selfish DNA sequence which is mobile and can multiply within a genome. Such elements are present in all living organisms. Wicker *et al.* (2007) proposed a classification of TE based on their transposition mechanisms (see figure 2.1). They used a hierarchical classification: class, subclass, order, superfamily, family, and subfamily. The class I are the retro-elements, also known as the copy-and-paste elements. Their transposition requires an RNA intermediate, which is then reverse-transcribed to DNA, and inserted into a new site. These class I elements include long terminal repeat (LTR) elements, and non-LTR elements such as LINEs (long interspersed nuclear elements) and SINEs (short interspersed nuclear elements). Class II TE are DNA elements, also known as cut-and-paste TE. They are directly excised from one genomic locus and inserted into another locus. Both classes contain autonomous elements and non-autonomous elements. Autonomous elements have all the necessary machinery to transpose, whereas non-autonomous elements lack at least a part of the machinery, and sometimes do not even encode any proteins.

One of the first cases of HTT reported in eukaryotes is that of the P element that was transferred from *Drosophila willistoni* to *Drosophila melanogaster* (Daniels *et al.* 1990), before transferring again from the former to *Drosophila simulans* (Kofler *et al.* 2015). *Drosophila* is the genus with the highest number of HTT recorded: 80.8% of the total number of HTT recorded in eukaryotes (Wallau *et al.* 2018; Dotto *et al.* 2018). However, this number is strongly biased since *Drosophila* is the main model which has been investigated to study TE (Mérel *et al.* 2020). All species combined, we went from 200 documented cases of HTT to 5600 in only 10 years (Dotto *et al.* 2018). To my knowledge, only three major large-scale studies looked for HTT between multicellular organisms. The first one focused on class I TE and they estimated that at least 65% of the 40 plants they studied harbor at least one HTT (Baidouri *et al.* 2014). Another one, found 2,248 new HT events across the 195 insects they studied (Peccoud *et al.* 2017). They also estimated that on average, 2.08% of the nucleotides of insect genomes result from the activity of horizontally acquired TE, with a maximum of 24% for the barn fly. The third one inferred 975 HTT events among 307 vertebrates (Zhang *et al.* 2020). Nonetheless, very few studies have quantified HTT at large taxonomic scales, because of the many challenges these studies have to face, such as the quality of the genome assemblies and annotations, and the risk of contamination that could lead to false positives (see chapter 5). Yet, several features can explain why TE are so prone to HT.

First of all, TE are a massive source of raw genetic material since TE are the most abundant entity of large eukaryotic genomes (Schaack *et al.* 2010). For example, the proportion of TE is about 45% of the human genome (Lander *et al.* 2001), 40% of the *Mus musculus* genome and near 80% in the *Rana esculenta* (Biémont and Vieira 2006) and maize genomes (Schnable *et al.* 2009). This

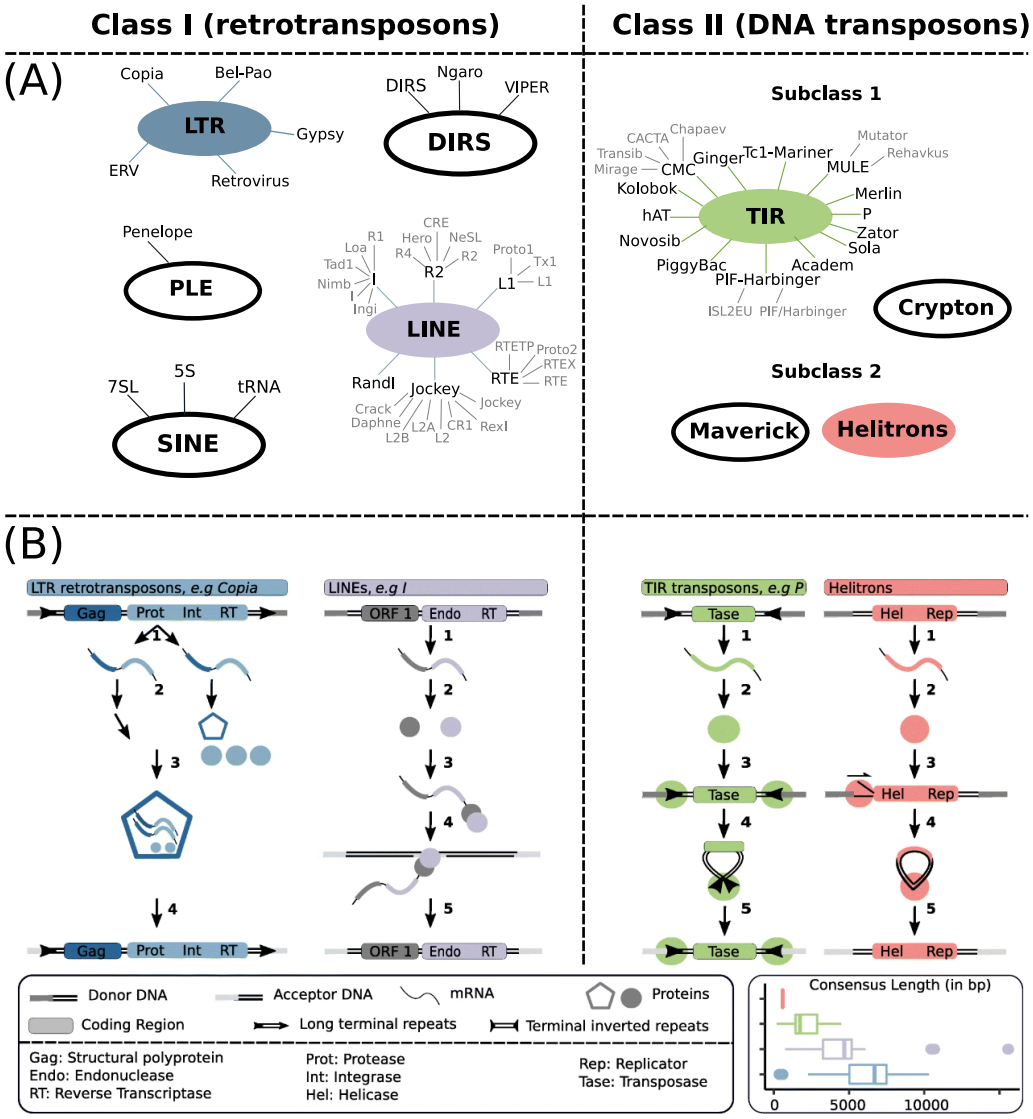


Figure 2.1: **TE classification, structure and transposition mechanisms.**
A. Classification system for TE. The classification is hierarchical and divides TE into two classes based on the transposition intermediate (RNA or DNA). They are further subdivided into subclasses, orders (ovals in figure) and superfamilies (black names linked to ovals).
B. Figure from Mérel *et al.* (2020) illustrating the structure and transposition mechanisms of four TE orders. **LTR retrotransposons**: 1. Transcription. 2. Translation of one part of the transcripts. The protease (Prot) cleaves pol polyprotein. 3. gag proteins assemble around untranslated transcripts, the integrase (Int), reverse transcriptase (RT) and a tRNA. 4. Reverse transcription and integration. **LINE retrotransposons**: 1. Transcription. 2. Translation. 3. Protein(s) bind to the transcript. 4. A strand of donor DNA is cut, target-primed reverse transcription starts at the exposed 3' extremity. 5. The TE is integrated. **TIR DNA transposons**: 1. Transcription. 2. Translation. 3. Two transposases bind to the TIRs. 4. Transposases dimerize and cut TIR extremities forming a free complex. 5. The complex binds to donor DNA and is integrated. **Helitron DNA transposons**: 1. Transcription. 2. Translation. 3. At the donor site, the plus strand is cut. A replication fork is formed. 4. Replication results in a double stranded transposon circle. 5. Integration. The bottom right panel represents the distribution of the lengths of *D. melanogaster* consensus sequences. The same colors are used in all the panels.

proportion is usually lower in smaller genomes, with 15-22% in *Drosophila melanogaster* genomes for example (Biémont and Vieira 2006).

Secondly, TE have the ability to cross the nuclear envelope and to insert into genomes. This mobility may also facilitate movements across individuals – in addition to along genomes - potentially leading to HTT.

Thirdly, once in a new and naive genome, TE can rapidly increase in copy number. Since naive genomes lack the appropriate repression tools, TE can massively transpose upon arrival, through an initial transposition burst (Le Rouzic and Capy 2005). It is then more difficult for the genome to get rid of all these copies, increasing the chance that this TE stays in this genome.

Finally, HT is actually very important for the persistence of TE (Schaack *et al.* 2010). Although TE can be a source for genetic novelty, their transposition can have deleterious effects on the host genome, such as chromosome breakages, mutations, ectopic recombination, and genetic rearrangements (Yoth *et al.* 2022). All these deleterious effects lead to the development of mechanisms to control TE in the hosts (Mérel *et al.* 2020). TE can be repressed thanks to epigenetic marks or RNA pathways such as siRNA and piRNA. Yet, too strict a control can lead to the loss of TE, which would deprive the host of the potential genetic novelty brought by TE. This might be why some eukaryotes temporarily relax their TE silencing machinery in their germline, such as the "Piwiless pocket" in *Drosophila* or the relaxation during epigenetic reprogramming in mammals (Mérel *et al.* 2020). Thus, TE can persist in a genome with the adequate balance between TE expression and TE repression (Bourque *et al.* 2018). Otherwise, TE can also persist through evolution by transposing to a naive genome by horizontal transmission, which allows it to escape vertical extinction (Schaack *et al.* 2010).

Because of their specific features, TE of different classes do not all have the same chance to undergo a HT. Regardless of the proportion of each class in a genome, both Peccoud *et al.* (2017) and Zhang *et al.* (2020) found an over-representation of HT of DNA elements over RNA elements in insects and vertebrates, respectively. This prevalence of DNA elements might partly be explained by a better stability for double-stranded DNA intermediates over RNA intermediates (Schaack *et al.* 2010). In addition, we expect autonomous elements to be more successful in HT across distant taxa since they carry their own transposition machinery, as opposed to non-automous ones (Schaack *et al.* 2010). In this sense, Tc1-Mariner, an autonomous DNA element, is the TE super-family for which both Peccoud *et al.* (2017) and Zhang *et al.* (2020) found the most HTT, and across more distantly related taxa. Interestingly, Tc1-Mariner element has a “blurry” promoter, which is able to activate transcription of a reporter gene in very distant taxa (in metazoan species but also yeast and bacteria), making this element compatible with many genomes (Palazzo *et al.* 2019).

2.2 . Polydnaviruses

Another mobile element, specific to some parasitoid wasp genomes, are polydnaviruses. Parasitoid wasps are a paraphyletic group of Hymenoptera, and can be ectoparasite or endoparasite insects, depending on whether they develop on or within their host (Beckage and Drezen 2011). To ensure the developmental success and survival of their eggs and larvae within hosts, many endoparasitoid wasps inject viral-like particles and venom in their host, at the same time as their eggs (see step (1) figure 2.3) (Herniou *et al.* 2013). Once in the host, the content of the viral particles and the venom repress the immune response of the host (Beckage and Gelman 2004). Viral particles of parasitoid

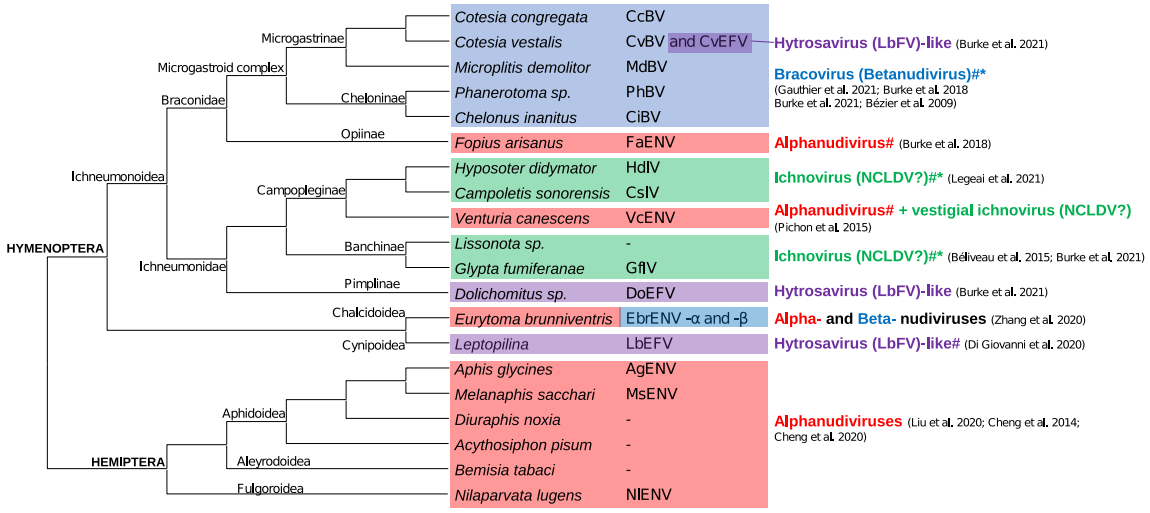


Figure 2.2: **Diversity of endogenous viral elements (EVE) derived from large double-stranded DNA viruses.** Branch lengths are not proportional to divergence. References in which each EVE was characterized are indicated. * indicates that EVE present in the group of species delineated by colored rectangles are orthologous or likely orthologous. # indicates domesticated EVE. The endogeneization event leading to bracoviruses in the Microgastrinae complex took place about 100 Myrs ago. Figure from Gilbert and Belliardo (2022).

wasps appeared several times independently in the evolution of wasps (see figure 2.2), yielding viral particles that differ in terms of structure, function and content. Remarkably, viral particles of the group polydnavirus (PDV) contain DNA circles, whose virulence genes are expressed once injected in the hosts. PDV are composed of two genus: the bracoviruses (BVs) and the ichnoviruses (IVs), respectively carried by wasps of the Braconidae and Ichneumonidae families, which both belong to the super-family Ichneumonoidea (see figure 2.2). In both genus, polydnaviruses are composed of two elements, both found in the wasp genome: (i) proviral segments that contain genes of wasp and unknown origin involved in the virulence against the host, and (ii) genes of viral origins. In the calyx cells of the ovaries, the former element (the proviral segments) is amplified, excised from the wasp genome and circularized into DNA circles (see steps (1a) and (1b) figure 2.4), thanks to a conserved motif called DRJs (Direct Repeat Junctions) which delimits each proviral segment (Desjardins *et al.* 2008; Burke *et al.* 2015; Legeai *et al.* 2020). The second element (the genes of viral origins) encodes viral particles (see step (1c) figure 2.4), in which are packaged the DNA circles (see step (3) figure 2.4). These assembled particles are then injected in the host.

The genes of viral origins were acquired several times independently in the course of the Ichneumonoidea wasps evolution, following the endogeneization of a virus and the domestication of some of its viral genes. The expression of these domesticated genes allow the Ichneumonoidea wasps to form viral particles. One of the events took place about 100 million years ago in the common ancestor of the microgastroid complex, which belongs to the Braconidae wasp family (see figure 2.2). The endogenized virus was a nudivirus, and all the descending wasp species, the microgastroid complex, now form a hyperdiversified monophyletic group, estimated to contain at least 46,000 species, which are all thought to harbor polydnaviruses (Bézier *et al.* 2009). The polydnaviruses resulting from this major event are called bracoviruses. Although there was only one event of endogeneizaion leading to

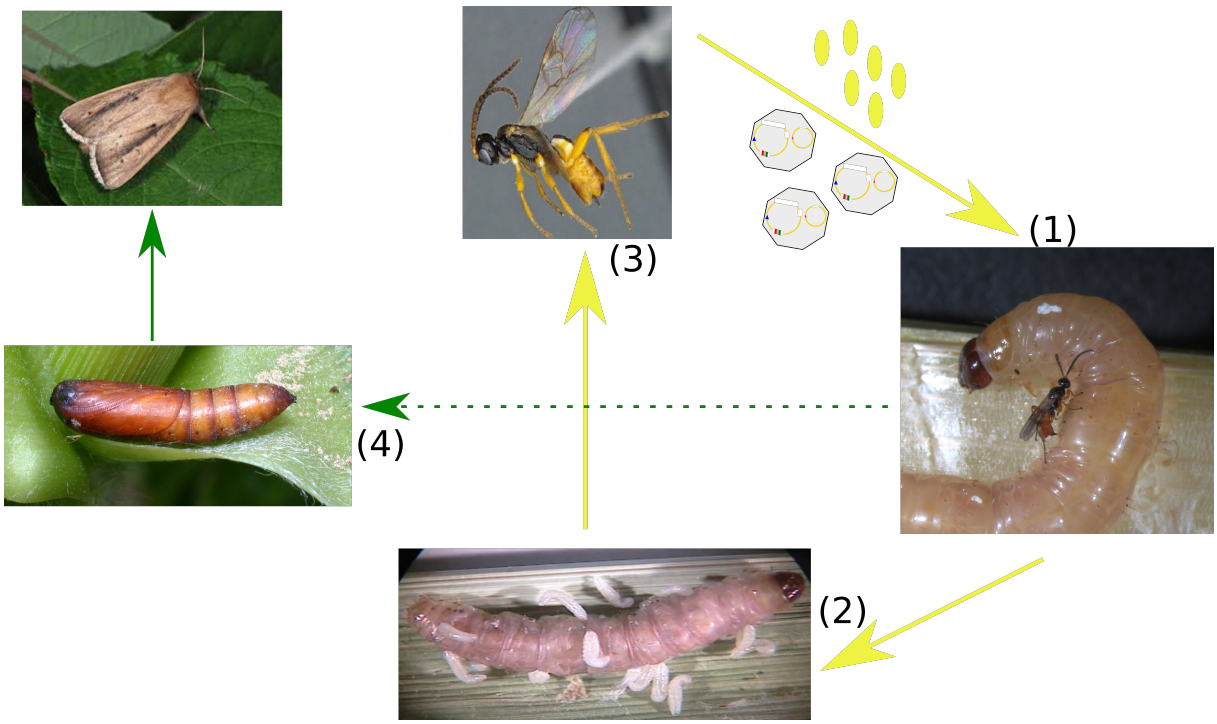


Figure 2.3: **Life cycle of parasitoid wasps**. As an example, we show the life cycle of *C. typhae*, represented by yellow arrows, with its natural host *S. nonagrioides*. In (1), *C. typhae* injects its eggs (yellow ovals) in the caterpillar and polydnaviruses (same representation as in figure 2.4). Parasitism can be successfull (2) or not (4). In case (2), the host dies a short time after the wasp larvae emerge from its body, whereas wasp larvae continue their development to cocoons and adult stage in (3). In case (4), the caterpillar can continue its life cycle, represented by green arrows.

polydnaviruses in Braconidae wasps, at least two events took place in Ichneumonidae, leading to several groups of polydnaviruses called ichnoviruses (see figure 2.2). However, in the case of ichnoviruses, the origins of the endogenized viruses are unknown, yet they both seem to derive from related viral progenitors (Béliveau *et al.* 2015).

Since the genes of viral origins are not injected in the host, polydnaviruses are only able to replicate in the wasps. This is why, this "virus" is really part of the wasp genome. Viruses being often referred as infectious particles, one can wonder whether polydnaviruses are really viruses, or whether one should consider them as transposable elements instead (mostly if we discover that proviral segments arise from TE, yet their origin is unknown for now), or even as a third category. The boundary of these two (three?) elements is as blurry as their origins, the distinction between retroviruses and LTR elements being unclear too (Hayward and Gilbert 2022).

Interestingly, several studies have shown that at least some DNA circles were somehow able to integrate in the genome of the host, in cell culture but also *in vivo* (McKelvey *et al.* 1996; Gundersen-Rindal and Lynn 2003; Beck *et al.* 2011; Chevignon *et al.* 2018; Wang *et al.* 2021b). It is not clear what role, if any, is fulfilled by integration, but it was proposed that it might enable a better persistence of the virulence genes throughout the wasp development (whereas the genes located on free DNA circles might be degraded after some time), and/or a better efficiency for gene expression for genes whose products operate on signaling pathways within each cell (Chevignon *et al.* 2018). Whatever the role played by integration, the level of conservation of the motif that was identified for

the integration of the DNA circles suggests a strong selection pressure for integration. This motif, named HIM for Host Integration Motif, was firstly identified in the Braconidae *Microplitis demolitor* (Beck *et al.* 2011). A latter study on the Braconidae *Cotesia congregata* found that at least eight of its DNA circles were able to integrate in the host, and that all these circles harbored HIMs, which is conserved between circles but also with *M. demolitor*'s circles (Chevignon *et al.* 2018). More precisely, the integration involves the motifs J1 and J2, which are the most conserved part of HIMs. It was shown that a ≈50bp sequence located between J1 and J2 is lost during integration, and that the integrated segments end up with J1 and J2 at their extremities in the host genome, instead of DRJs like in the wasp genome (see step (4) figure 2.4). Because of the new configuration of the DRJ (a single hybrid DRJ in the inner segment), integrated segments are not able to circularize again once in the caterpillar genome.

Despite their independent origins, bracoviruses and ichnoviruses harbor remarkable similarities. In addition to a similar architecture (PDV are composed of proviral segments within viral particles, and produced in the calyx before being injected in the host), ichnoviruses also have DRJs and HIMs that work the same way as for bracoviruses, although their nucleotidic sequences are clearly different (Wang *et al.* 2021b). In addition, some DNA circles of ichnoviruses were also found to integrate in the genome of the host (Wang *et al.* 2021b).

Because of their mobility, their capacity to enter cells, and their own mechanism for integration via HIM, PDV are an interesting source of DNA for HT. Despite the fact that integration is often a dead-end since most parasitized hosts die, several examples of HT from PDV were reported. The first study recovering HT from PDV to its hosts found 105 regions in two Lepidopteran genomes that derived from such HT, including two regions encoding for a BEN domain, known to be associated with polydnaviruses and transcriptional regulation (Schneider and Thomas 2014). In addition Gasmi *et al.* (2015) recovered bracoviral sequences in several lepidopteran genomes: in the monarch (*Danaus plexippus*), in the silkworm (*Bombyx mori*), in the beet armyworm (*Spodoptera exigua*), and in the fall armyworm (*Spodoptera frugiperda*). *Spodoptera littoralis* even domesticated a bracoviral gene, *SI gasmin*, which now plays a role in anti-bacterial immune response (Lelio *et al.* 2019).

2.3 . Other sources

Theoretically, any DNA sequences could transfer to a recipient species, even if they do not have inherent abilities to move across a genome like TE or PDV. Indeed, several horizontal transfers of genes (HGT) were reported, although examples between multicellular organisms are quite anecdotal, contrary to examples of HGT from bacteria to multicellular organisms which are more numerous. Although most of the massive HGT in bdelloid rotifers originated from bacteria, some from fungi were reported (Gladyshev *et al.* 2008). Another large-scale study found multiple horizontally acquired genes in both vertebrate and invertebrate genomes (Crisp *et al.* 2015). Some of these genes came from plants and fungi, although most of them came from bacteria and protists. Also from fungi, aphids and spider mites acquired carotenoid biosynthesis genes (Moran and Jarvik 2010; Altincicek *et al.* 2012). Carotenoids are pigments that protect plants and fungi from photo damage by reactive oxygen species. In animals, carotenoids can be involved in night vision and coloration. However, before this discovery it was assumed that animals were unable to produce their own carotenoids, but that they could sequester such pigments from their diet. Another example is the one of the whitefly

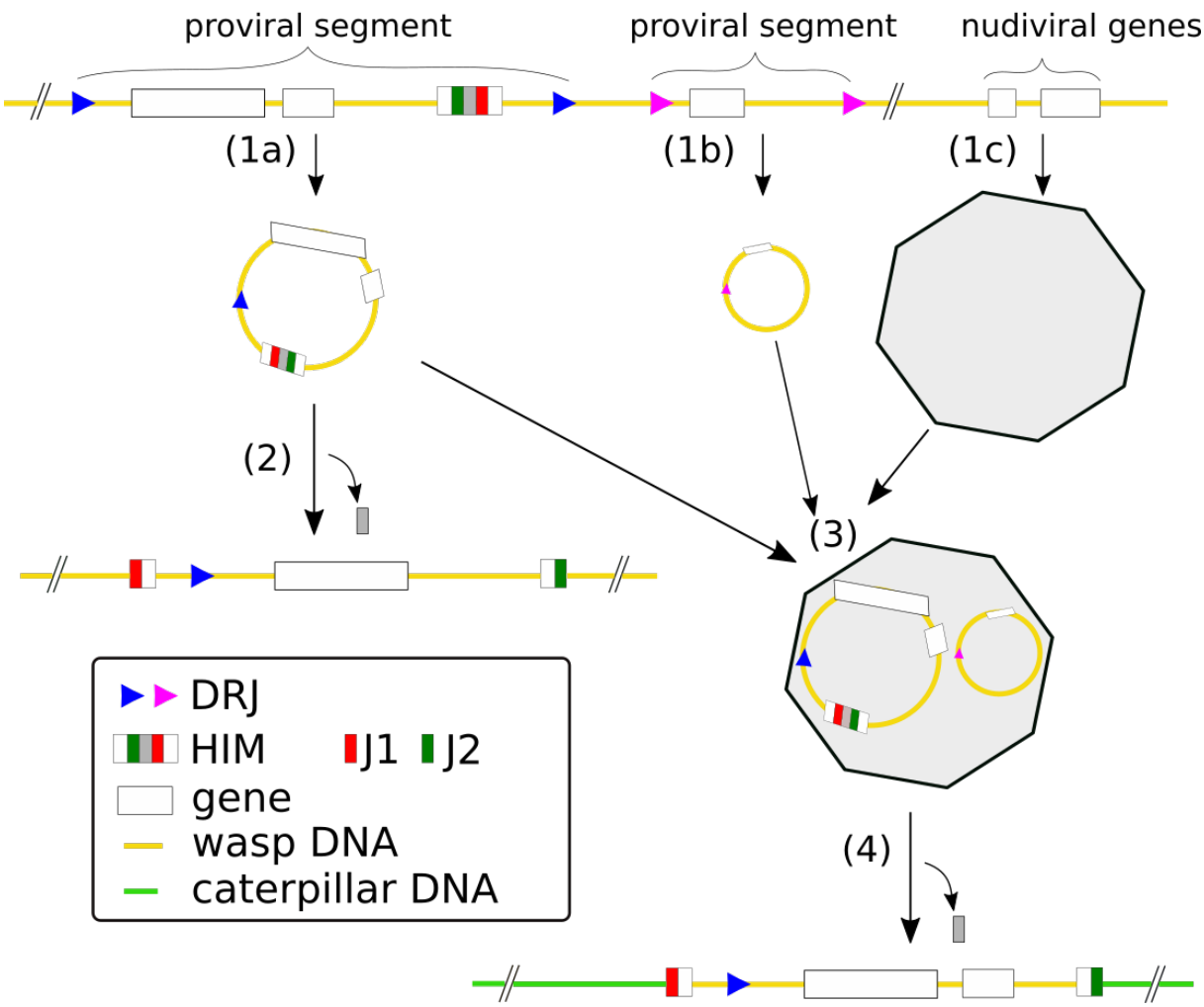
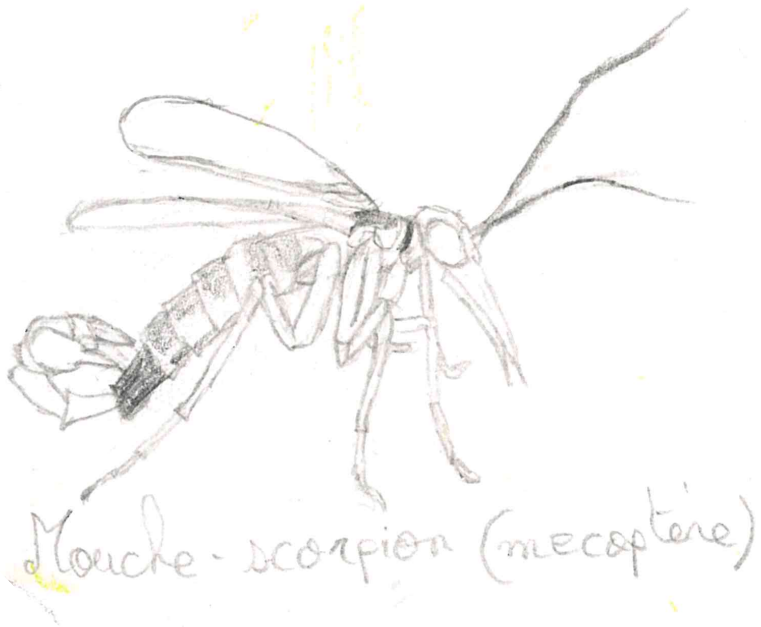


Figure 2.4: **Structure of bracoviruses.** In the wasp calyx, (1a) is a segment containing a HIM delimited by two blue DRJ, (1b) is a segment without HIM delimited by two pink DRJ and (1c) corresponds to genes of viral origin, more precisely a nudivirus in the case of bracoviruses. (1a) and (1b) form DNA circles, whereas (1c) form viral particles. In (2), DNA circle reintegrates into the wasp genome, loosing about 50bp, in gray, located between J1 and J2. In (3), DNA circles are packaged into viral particles, which are injected in the caterpillar host at the same time as the wasp eggs. In (4), the HIM-containing segment integrates into the caterpillar genome, loosing the same 50bp as in (2).

Bemisia tabaci that acquired a plant gene conferring a resistance to phenolic glucosides, defensive toxins produced by many plants (Xia *et al.* 2021). Gilbert and Maumus (2022) then recovered 49 plant-like genes in the genome of *B. tabaci*, deriving from at least 24 independent HGT events. Thus, substantial plant-to-insect HGT may have facilitated the evolution of *B. tabaci* toward adaptation to a large host spectrum. Furthermore, Li *et al.* (2022) achieved a large-scale study among 218 insects in which they show that most HGT come from bacteria (1,115 events, *i.e.* 79.0% of the total number of events they recovered), but they also found many HGT from multicellular organisms: 194 (13.8%) genes from fungi, and 43 (3.0%) genes from plants. Some of these genes are involved in important insect adaptation, such as courtship. All these examples show that genes can sometimes cross kingdoms of life in multicellular organisms. To my knowledge, only one example of HGT *between* metazoa was reported: a gene coding an antifreeze protein between two species of fishes, allowing them to live in cooler water (Graham and Davies 2021).

Studies on HGT between multicellular organisms, and especially between metazoa, being quite recent, it is not clear whether such a little amount of cases reflects the reality or is simply due to a lack of studies. Nonetheless, even a single HGT can have major consequences on the evolution of an organism, as briefly discussed in the above examples. Nonetheless, I focused only on HT from TE and from polydnaviruses over the course of my PhD, without looking at other possible sources of HT.



3 - Putative mechanisms for horizontal transfers

In multicellular organisms, it is not clear how genetic material can cross all the barriers to achieve a horizontal transfer: exiting its cells and organism, reaching another organism, entering its cells and nucleus, integrating in the recipient genome, and being transmitted to the next generation. We saw above that depending on the genetic source which is transferred, a part of the mechanism can be fulfilled by the source itself. Indeed, we saw that TE and PDV have their own mechanisms to excise from a genome and integrate in another one, and the latter are even equipped to enter cells. Nonetheless, PDV are limited to some parasitoid wasps, so they cannot explain HT involving other taxa. About TE, they do not fulfill all the necessary mechanisms, like exiting its cell, and reaching another organism to enter its cells, except maybe for LTR retrotransposons, some of which are able to produce virus-like particles, such as *gypsy* and *copia*. It was shown that when larvae of *Drosophila* strains, in which *gypsy* is normally inactive, are exposed to these particles, a high level of *gypsy* insertion activity is observed in their progeny (Song *et al.* 1994). Thus, *gypsy* is able to infect cells, similarly to a virus. In the case of non-mobile genetic sources, such as genes, the mechanism involved is really a black box. Hereinafter, I describe putative mechanisms explaining each step required for a HT to occur, but they are of course not mutually exclusive, and could even all co-exist (see overview in figure 3.1).

3.1 . Exiting the donor cell and reaching the recipient cell

Since eukaryotes do not have specialized apparatus for HT, unlike the type IV secretion system of some bacteria for example (figure 1.1), HT between eukaryotes could theoretically take place by acquiring naked genetic material, through feeding, or it could rely on vectors. Naked DNA and RNA circulate in animal fluids (blood, saliva, etc), yet it is unclear how long it takes for such genetic material to be degraded. About feeding, such cases of HT has never been demonstrated. Vectors are thus the best candidate to date to complete a HT. Vectors could be anything transporting genetic material from the donor to the recipient cell. The main putative vectors are viruses, but other vectors are also discussed: intracellular organisms and extracellular vesicles.

3.1.1 . Viruses as vectors of HT

It has been proposed that viruses might act as vectors of HT between multicellular organisms, which means that they could transport non-viral genetic material from one multicellular organism to another. Viruses are very good candidates because they enter host cells, where they replicate, and because they can be transmitted between hosts by infection (Loreto *et al.* 2008; Gilbert and Cordaux 2017). In such a scenario, a first HT would take place from one multicellular organism to a virus (step one), then the virus would infect another species and transfer the recently acquired HT (step two). For such a transfer to take place, we have to find out (i) whether viruses can receive and carry foreign genetic materials, (ii) what are the chances of the newly acquired genetic material to persist long enough in the virus to be transmitted to another species, and (iii) whether a second transfer can take place, from the virus to the second species.

For the first inquiry, several studies showed that retroviruses could encapsidate host RNA, some-

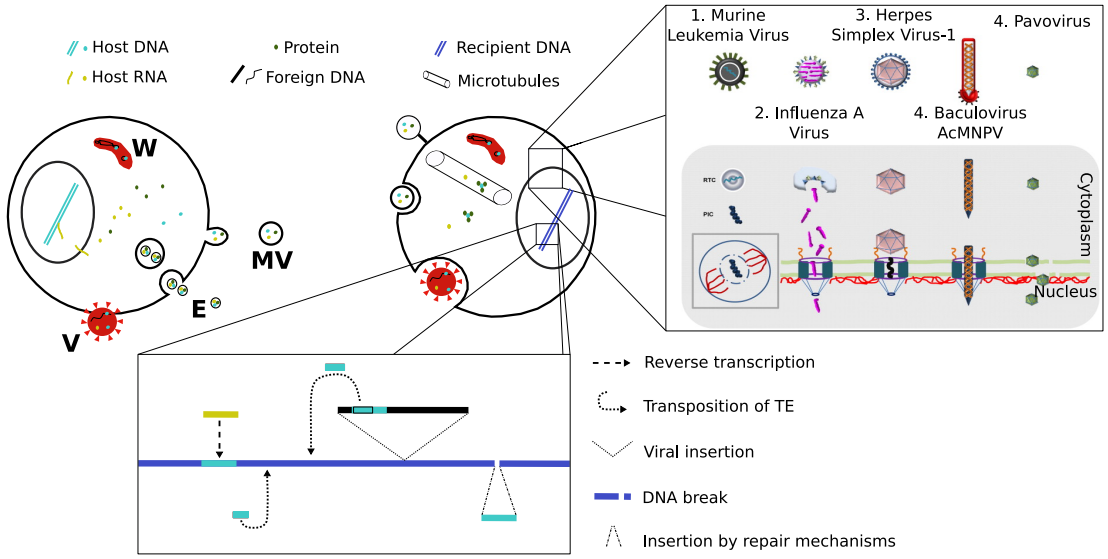


Figure 3.1: **Putative mechanisms for a HT.** The first cell shows the mechanisms to exit a cell: MV = microvesicles, E = exosomes (these two kind of extracellular vesicles transport DNA, RNA, and proteins), V = virus (transporting free DNA, free RNA, and a DNA integration), and W = Wolbachia (transporting a DNA integration in its genome and in its plasmid). The second cell shows the mechanisms to enter a cell and to reach its nucleus: extracellular vesicles can enter cells thanks to a receptor, by fusion, or by phagocytosis, virus have a variety of mechanisms to enter cells, and free DNA is associated to proteins to form a DNA-protein complex, which is transported via microtubules. The zoom on the nuclear membrane is a figure taken from Cohen *et al.* (2011), showing how virues can enter the nucleus: (1) The MLV PIC gains access to the nucleus during mitosis, when the NE is temporarily disassembled. (2) Influenza A virus undergoes extensive disassembly in the cytoplasm. The cytoplasmic released vRNPs contain NLSs and are thereby able to cross the NPC using the host transport machinery. (3) HSV-1 capsids use importins to attach to the cytoplasmic side of the NPC. Interaction with the NPC then triggers the release of the viral genome, which then enters the nucleus through the NPC. (4) Capsids of the baculovirus AcMNPV cross the NPC intact. Genome release presumably occurs inside the nucleus. (5) Parvoviruses transiently disrupt the NE and nuclear lamina, and enter the nucleus through the resulting gaps. The zoom on the recipient DNA shows the mechanisms to integrate in the recipient genome.

times up to 50% of the total encapsidated RNA (Eckwahl *et al.* 2016; Gilbert and Cordaux 2017). The host RNA could then theoretically integrate in the genome of the next host via retroelement-mediated reverse transcription (see section 3.3). For DNA sequences, the ability of some TE to form extrachromosomal circular forms during transposition suggests that they could also be encapsidated in viral particles, although the ability of DNA viruses to encapsidate foreign DNA still needs to be assessed. It was shown in some specific virus-host interactions, such as the sugar beet plants and the Beet curly top Iran virus, that spontaneaous hybrid DNA minicircles (composed of host DNA and of the viral DNA) can be formed (Catoni *et al.* 2018). The authors also showed that these hybrid mini-cirles always carry the regulation regions crucial for the virus life cycle, allowing them to replicate and also to produce RNA transcript in other plant species secondarily infected by this virus. Alternatively, the genetic material could also insert in the genome of the virus, instead of being freely carried in its viral particles. In this sense, numerous genes encoded by large dsDNA viruses originate from their eukaryote hosts (Holzerlandt *et al.* 2002; Thézé *et al.* 2015; Gilbert and Cordaux 2017; Irwin *et al.* 2022). Gilbert *et al.* (2016) were able to estimate the frequency at which such transfers take place in AcMNPV, a large dsDNA virus of the Baculoviridae family. For this, they purified and sequenced viral DNA from moths which had been infected with AcMNPV. They found that about five percent of the viral genomes harbored a *de novo* TE insertion originating from the infected moths. This study demonstrates that HT from host to virus can be quite rampant.

Regardless of whether the genetic material is inserted in the virus or freely carried, it has to persist long enough to be transmitted to another species, which is our second inquiry. Thus, the second step of a HT mediated by a virus (transmitting the newly acquired sequence to another species) has to take place before the free DNA is degraded, or before the integrated one is purged. Viruses have a very high density of genes, so we can expect most insertions in their genomes to be deleterious and to persist at very low frequency and only over a few viral replication cycles. In this sense, Gilbert *et al.* (2016) could not detect any persistence of TE insertion in AcMNPV after ten infection cycles. Yet, the newly acquired genetic material does not need to be present in high frequency in a population of virus to be transmitted, since it was shown that HVT can persist at polymorphic loci (Gilbert and Cordaux 2017). It was also shown that transposable elements were able to transpose from virus to virus, which could help maintaining TE long enough in the viral population (Loiseau *et al.* 2021).

For the third inquiry, *i.e.* a HT from a virus to a multicellular organism, it is actually very common in laboratory conditions. It is simply what happens during transgenesis when a virus is used as vector (Gama Sosa *et al.* 2010). In addition, I gave some examples of HT from viruses (HVT) in chapter 1. Although the frequency at which such endogeneization events take place remain to be assessed, we know that endogeneization is quite frequent and is still ongoing, at least in arthropods. This is suggested by the fact that the majority of arthropods endogenized viral elements (EVE) are specific to some species, and sometimes not even fixed in a species (Thézé *et al.* 2014; Gilbert and Cordaux 2017).

Some type of viruses are probably more prone to vector genetic materials than others, which would depend on the range of species they can infect, their ability to encapsulate foreign genetic materiel and/or insert it in their genome (which seem to differ depending on viruses (Loiseau *et al.* 2021)), their tropism (since only germinal integration will be transmitted), and whether integration in the host genome is part of their life cycle (which is the case for retroviruses only). Yet, the integration of the genetic material of interest can take place independently of the integration of the virus with

various putative mechanisms described in section 3.3.

Another argument in favor of viruses as vectors of HT is that viruses can be transported over long distances via their hosts. To my knowledge no study ever observed a complete HT in real time from a multicellular organism to another shuttled by a virus. Such an event is actually very unlikely to be observed since we would need to catch a recent event and to recover the shuttled genetic material in the three organisms of interest (the donor, the virus, and the recipient). Furthermore, endogenized viruses sometimes belong to unknown viral families, possibly extinct, which impede our ability to assess the origins of some EVE. Is is the case for example for the endogenization event that lead to ichnoviruses in Ichneumonidae wasps (Béliveau *et al.* 2015). The recent development of paleovirology might shed light on the propensity of EVE (Metegnier *et al.* 2015; Legendre *et al.* 2015).

In addition of free viruses, polydnaviruses are also very good candidates as vectors of HT between parasitoid wasps and their hosts, as discussed in the chapter 2. Although I investigated to some extant free viruses as vectors of HT between animals over the course of my PhD (part I), I really emphasized on polydnavriuses (part II).

3.1.2 . Other intracellular parasites as vectors of HT

Intracellular parasites (prokaryotes but also eukaryotes) may also be viewed as vectors of HT. Those that can be horizontally transmitted between animal species and that can reach germinal cells for vertical transmission may facilitate these transfers. In the same way as viruses, genetic elements from the initial host could firstly undergo a HT to the parasite, and then to the new host after horizontal transmission of the parasite.

Intracellular eukaryotes, which are unicellular organisms, have a variety of mechanisms to enter cells, such as phagocytosis, direct penetration, or induced uptake (Sibley 2004). For example, trypanosomes, parasites causing serious diseases in humans and domesticated animals, seem to have acquired a HGT from a vertebrate (Steglich and Schaeffer 2006).

Wolbachia, a diverse group of α-proteobacteria, is a very interesting case because it is found in many species of Arthropods and Nematodes, and some genera of α-proteobacteria even parasitize mammals (Werren *et al.* 2008). More precisely, Hilgenboecker *et al.* (2008) estimated that about 66% of Arthropod species are parasitized by *Wolbachia*, although the frequency of infection within one species can sometimes be very low. *Wolbachia* is vertically transmitted by females, which means that they are present in the female germline, and many *Wolbachia* can manipulate the reproduction of their hosts to increase there transmission thanks to feminization, parthenogenis, male killing or cytoplasmic incompatibility. Although there is a general concordance between the phylogeny of Nematode-associated *Wolbachia* and their hosts, this is not the case for Arthropod-associated *Wolbachia*, suggesting that these *Wolbachia* are also horizontally transmitted between Arthropods (Werren *et al.* 2008). This horizontal transmission can take place between distant species, with some group of *Wolbachia* that seem more prone to this mode of transmission than others (Werren *et al.* 1997). It was also shown that transmission can occur to parasitic insects from their infected hosts, which often belong to different taxonomic orders (Heath *et al.* 1999; Ahmed *et al.* 2015).

3.1.3 . Extracellular vesicles as vectors of HT

Other possible vectors of HT are extracellular vesicles, which share many structural features with viruses. Extracellular vesicles can be defined as any membrane-bound vesicles that are released by cells. Extracellular vesicles have a similar size to viruses, and they transport biological components between

cells (Bongiovanni *et al.* 2021). The components found in these extracellular vesicles consist mostly in lipids, proteins, and RNA, including retrotransposon transcripts (Balaj *et al.* 2011). However, it is still unclear whether DNA is present inside these molecules (Cai *et al.* 2016). Several studies have shown that extracellular vesicles can enter cells of another organism (Coakley *et al.* 2015; Rodrigues *et al.* 2008; Samuel *et al.* 2015; Mu *et al.* 2014). They could be transmitted to recipient species by extracelullar fluids, such as blood, saliva, or interstitial fluids. In this case, the donor and the recipient species would need to be in contact somehow (see the chapter 4). Kawamura *et al.* (2019) detected RNA transcripts of L1 retrotransposons in extracellular vesicles, and they showed that these RNA could be reverse transcribed and insert in the genome of the recipient cells. Furthermore, Ono *et al.* (2019) found a striking experimental example of HT via extracellular vesicles, while assessing the risks of unintentional DNA insertions with CRISPR-Cas9 in cultivated mouse cells. Looking at DNA long insertions at double strand sites in the mouse genome, they found that only 16% of these insertions derived from mouse DNA, the rest coming mainly from *E. coli* or plasmids, which are used in the experiment. Surprisingly, they also identified some insertions from the DNA present in the cell culture medium, *i.e.* either from fetal bovine or goat serum. Repeating the experiment with exosome-free serum abolished most of these insertions, suggesting that HT from serums to cultivated mouse cells were mediated by exosomes.

3.2 . Reaching the nucleus

Once in the cytoplasm, free DNA has a high probability to be degraded by nucleases. Indeed, it was shown that plasmid DNA is degraded in the cytoplasm of HeLa and COS cells with a half-life of only 50–90 min (Lechardeur *et al.* 1999). Nonetheless, studies interested in transfection showed that free DNA is quickly associated with host proteins to form a DNA-protein complex, that would protect the DNA from degradation and make intracellular interactions possible, such as transport to the nucleus via microtubules (Bai *et al.* 2017). Then, DNA still has to enter the nucleus, a substantial barrier for DNA delivery. Indeed, a study estimated that out of the 2000 to 10000 plasmids which are delivered per cell following lipofection, only 20 to 1000 are detected in the nucleus 24–36 hours following DNA addition, *i.e.* 1 to 10% (Bai *et al.* 2017). This DNA can enter the nucleus upon the mitotic disassembly of the nuclear envelope, or to a smaller extent, through nuclear pore complexes (Bai *et al.* 2017). Although these data come from studies on transfection, the same mechanisms could possibly naturally occur during a HT.

Contrary to free DNA, genetic material that reached the cells thanks to a vector could also enter the nucleus thanks to this same vector, if it has the ability to enter the nucleus. In the case of viruses, DNA viruses and retroviruses replicate in the nucleus so they have mechanisms to reach it. Typically, it involves recognition by importins, transport to the nucleus, and binding to nuclear pore complexes. Some viruses enter the nucleus under their intact forms, while others have to disassemble (Whittaker *et al.* 2000). Polydnaviruses can also enter the nucleus since they are known to massively integrate in their host genomes. In the case of *Wolbachia*, several examples of insertions of their genetic material, sometimes even their whole genome, in their host genome suggest that *Wolbachia* can sometimes reach the nucleus (Cordaux and Gilbert 2017; Leclercq *et al.* 2016).

3.3 . Integrating the recipient genome

Once in the nucleus, genetic material still needs a mechanism to integrate the host genome. If this genetic material is inserted and transported by a virus, it could simply be integrated at the same time as the virus. Integration is part of the life cycle of retroviruses only, yet all types of viruses can be accidentally endogenized (Feschotte and Gilbert 2012). In the case of polydnaviruses, DNA can be integrated via HIMs, a mechanism specific to polydnaviruses described in chapter 2. If the genetic material is an autonomous TE, the mechanism of integration is also quite trivial. However, if the TE is non-autonomous, it can be integrated only if the recipient species have the right reverse-transcriptase or transposase. In this case, phylogenetic proximity between the donor and the recipient species probably increases the chance of insertions. In the case of other sources, they could be integrated thanks to the repair mechanisms of the host, such as HR (homologous recombination), MMEJ (microhomology-mediated end-joining, that uses 5-25bp microhomologies), or NHEJ (nonhomologous end joining, that uses 1-4bp microhomologies), depending on the host species. In Ono *et al.* (2019) for example, who found experimental HT mediated by exosomes, they identified that most of the integrations took place via NHEJ. RNA could also be integrated via reverse transcription. It was shown that such integrations are possible thanks to the retrotransposon of the host (Gilbert and Cordaux 2017; Goic *et al.* 2016)

3.4 . Success of HT

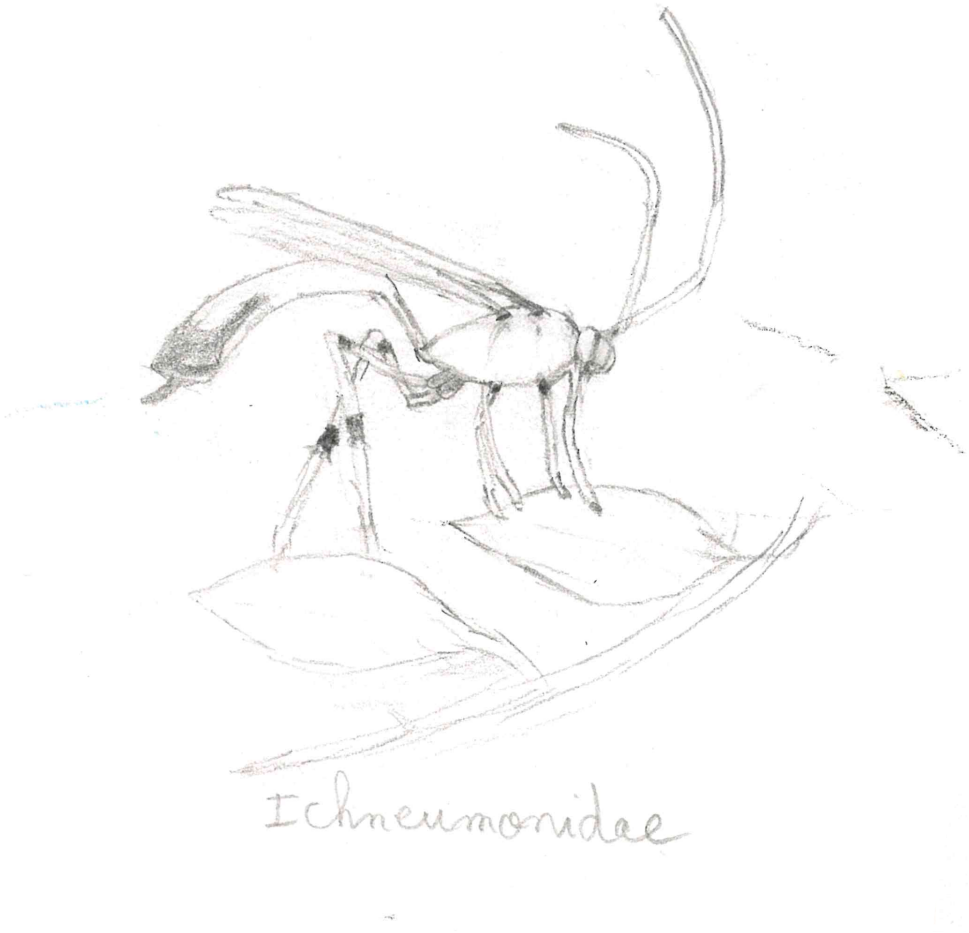
Even if the genetic material successfully passes all these barriers, it is not necessarily transmitted over generations. For this, the integration has to take place in a cell that will be transmitted to offspring. In the case of sexual organisms, these cells are restricted to germinal cells, yet only a small subset are transmitted to offspring. This is why vectors that target germinal cells are better candidates, such as the viruses and parasites that can be vertically transmitted. Instead of integrating in the germline of a parent, genetic material could also be inserted at early stages of embryogenesis, when pluripotent stem cells allow direct access to the germline of the next generation. In this sense, van der Kuyl and Berkhout (2020) enlighten how viruses can access the germline, reviewing studies on humans. They point out that the majority of viral families infecting mammals can be found in semen, and that numerous are able to cross the placenta and infect the developing fetus.

Once transmitted to offspring, all cells (somatic and germinal) harbor the newly acquired genetic material, which will therefore be automatically transmitted to the next generation, except in the absence of reproduction of course. Then, the fate of the newly acquired genetic material in the population is in the hands of classic evolutionary forces, like the rest of the genome, *i.e.* the balance between genetic drift and selective pressure. Since insertions in protein coding regions and regulating regions have a negative impact, we can expect most insertions to have a low probability to be maintained in a population. On the other hand, if the insertion is co-opted and provides an advantage, it would greatly increase its chance of fixation in the population. In the case of HTT, it was shown that despite the deleterious effect of transposition on the host genome, transposing increases their chance of persistence, except of course when the rate of transposition is so high that it leads to the sterility or death of their host (Le Rouzic and Capy 2005; Le Rouzic *et al.* 2007)

In the case of asexual organisms, the Muller's ratchet theory suggests that these organisms ac-

cumulate mutations because of the absence of the recombination that take place through sexual reproduction. For this reason, it is thought that HT are more common in asexual organisms, and would even be important to bring novelty in their genomes (Dunning Hotopp 2011). In this sense, many HT were recovered in bdelloid rotifers, an asexual group, with for example at least 8% of the genes of *Adineta vaga* that would originate from HT from non metazoan species (Gladyshev *et al.* 2008; Flot *et al.* 2013).

We can speculate that specificities of some genomes could have an impact on the probability of fixation. Such specificities could be the genome size (although Li *et al.* (2022) did not find any strong correlation with HGT among the 218 insects they analyzed), the TE number, the TE activity (although Loiseau *et al.* (2021) could not find any HTT from a lepiodpteran species having a high rate of transposition to its infectious virus), the gene density, the level of polyploidy (the deleterious effect of an insertion could be compensated by a higher level of polyploidy), and/or epigenetics factors (genomes which are efficient to silence TE might be less prone to HTT). The effective population size of the population could also play a role, since a population with a low effective population size is more influenced by genetic drift, which could favor the fixation of a HT just by chance.



4 - Putative factors promoting horizontal transfers

As for the mechanisms, the biotic and abiotic factors facilitating HT between multicellular organisms are not clear either. Regardless of the mechanisms implicated in HT, the donor and the recipient organisms need to be in contact somehow, directly or indirectly. Venner *et al.* (2017) suggested to deal with these connections with a network approach, where the intensity and the direction of the ecological links would reflect the potential frequency and direction of HT between species. Therefore, the network topology promoting HT would be represented by nodes corresponding to the reservoir species, each of them linked by edges which correspond to the connectivity between reservoirs (see figure 4.1). In this sense, Peccoud *et al.* (2017) found that geographical proximity may facilitate HTT, in insects at least. However, the design of their dataset did not allow them to clearly understand the strength of sympatry in the network. Other ecological factors could promote HT, such as prey-predator and host-parasite interactions. For the former, Kambayashi *et al.* (2022) found multiple HT events of BovBs, a LINE retrotransposon, in a prey-predator interaction, although the direction is quite surprising since it was transferred from predators (snakes) to their prey (frogs). They also showed that the transfers might have happened by the intermediate of diverse parasites, although they could not decipher whether the DNA sequence was integrated in the parasite genome, or whether the parasite carried bacteria and/or viruses whose genomes contained the DNA sequence. For host-parasite interactions, it was shown that this type of interaction had a role in the HT of four transposons families between invertebrates and vertebrates (Gilbert *et al.* 2010), and that endoparatic lifestyle promotes endogenization and domestication of viruses (Guinet *et al.* 2023).

If intermediates are involved, we can expect the geographical signal to be more or less weaker, depending on the ability of the intermediates to spread. Intermediates could be any living forms (eukaryotes, prokaryotes or viruses) that would receive the DNA from the donor species and then transport it to the recipient one, where the DNA could either be transported in the genome of the intermediate, or simply in a free form in its cells. Viruses are very good candidates, since we also saw in the chapter 3 that they have inherent mechanisms to transfer horizontally between organisms, enter cells, and even to integrate the host genome for some of them. Vectors of viruses could be viewed as indirect vectors of HT. ZOVER, a database of zoonotic viruses, focused the four main vectors of viruses: bats, rodents, mosquitoes, and ticks (Zhou *et al.* 2022). In addition of being a major reservoir for viruses, mosquitoes can bite various species which might increase their ability to spread viruses (Li *et al.* 2015; Goic *et al.* 2016; Whitfield *et al.* 2017).

The intermediate could also be the environment itself. Indeed, it was proposed that marine animals would be more likely to be involved in HT than terrestrial ones, the water acting like an ecological connection (Wang and Liu 2016), like it is the case for HT between bacteria (McDaniel *et al.* 2010). Water could indeed be a vehicle for direct DNA flow, where it would spread with a low UV exposure. In this sense, a large-scale study on 307 vertebrates found that 93.7% of the 975 HTT events they detected imply teleost fishes (Zhang *et al.* 2020). However, the authors could not decipher whether this large number of HTT in teleosts was due to the aquatic habitat or is specific to teleosts only. In addition, the only example of HGT between two metazoa I presented here is between two fishes: from a herring to a smelt (Graham and Davies 2021). The transferred gene encodes an antifreeze protein,

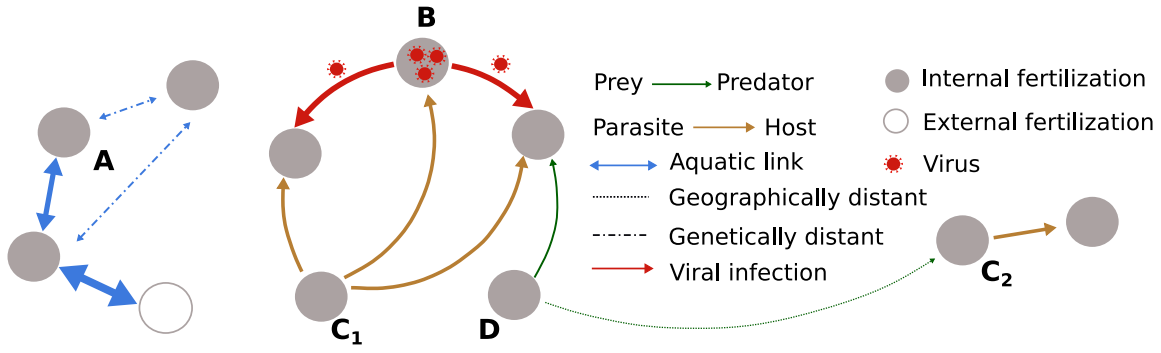


Figure 4.1: **Putative factors promoting HT.** Each circle represents an animals species, the filled ones use internal fertilization while the empty one uses external fertilization. **A.** Network of species linked by their aquatic habitat. **B.** A species acting as a reservoir for viruses (e.g. bats or mosquitoes). **C.** Parasites linked to their hosts: C1 is a generalist parasite, while C2 is a specialized parasite. **D.** A predator linked to its preys. The thickness of the edges is proportional to the the strength of the factor on HT. To date, such relative strengths are unknown, so I drew the thicknesses based on my personal hypotheses.

bringing a substantial benefit in cold waters. The authors suggested that external fertilization, which exists only in aquatic species, might be an additional factor that could favor HT. They argued that water is an environment that contains DNA of all the ecosystem's inhabitants, that could attach to sperm during spawning. Yet, neither the aquatic habitat, nor the mode of fertilization were directly tested. A recent study that scanned 3,325 genomes of various eukaryotic species for Introners, which are introns generated by the transposition of TE in genes, found that aquatic organisms are 6.5 more likely to contain Introners than terrestrial organisms (Gozashti *et al.* 2022). If HT are indeed more numerous in aquatic organisms, it could explain this high amount of Introners in their genomes.

In addition to the ecological factors and this network approach, features about the species genomes themselves, the donor and the recipient ones, could also have an influence on the probability of HT. As discussed hereinabove, teleosts seem more prone to HT than other vertebrates (Zhang *et al.* 2020). Although the authors hypothesized that it might be due to their aquatic habitat, it is not impossible that it is not the explanation, and that teleosts are more prone to HT for other reasons. Such reasons could be specificities of their genomes and/or features of their population (see section 3.4 on the probability of success of HT).

5 - Methods to detect horizontal transfers and limitations

5.1 . The problem of contamination

One of the reasons the possibility of HT in multicellular organisms remained controversial for quite some time is partly due to the lack of reliability of some of the first studies that reported HT based on bioinformatic analyses. For example, in the publication of the human draft genome in 2001 in Nature, the authors reported hundreds of human genes resulting from HT from bacteria (Lander *et al.* 2001). This discovery led to several successive publications claiming in turn that these genes did not come from bacteria, or supporting some of these HGT (Crisp *et al.* 2015; Salzberg 2017). Among the arguments against these HGT are the fact that too little genomes were available at the time to conclude with confidence to a HGT or that some might be due to contamination. For example, a HGT from the initial study of 2001 was firstly rejected, before being proved to be true only recently, in 2023, thanks to the better completeness of current databases (Kalluraya *et al.* 2023). Another study highlighted the problem of contamination in studies on HT: it was shown that one-sixth of tardigrade genes came from HGT, but this number has been greatly over-estimated because of contamination (Bemm *et al.* 2016). When working on HT, high precautions against contamination and a decontamination steps of the genomes are thus primordial. It is also very important to be aware of the possibility of contamination when interpreting the results.

However, we cannot help but sequence some contaminants such as parasites, viruses or bacteria present in the body or even in the cells of the organism of interest. It is possible to validate a HT thanks to a fluorescent *in situ* hybridization or a PCR (Husnik and McCutcheon 2018), yet it is not easily conceivable for large-scale studies (see figure 5.1, panel D). It is also possible to check whether the flanking region of the acquired DNA sequence does correspond to the recipient genome thanks to long reads or paired-end reads (Chuong *et al.* 2017) (see figure 5.1, panel D). In this thesis, we are interested in HT between multicellular organisms only, so we will not look at HT from bacteria, that might be sequenced at the same time as our species of interest. This will greatly decrease the probability of false HT due to contamination in this thesis, although it is not null. Indeed, contaminants and parasites (bacteria but also eukaryotes) can be present in several multicelullar organisms, possibly leading to the wrong interpretation of HT between these multicelullar organisms. In the same way, cross-contamination between the samples of interest can also lead to false HT, although this is less likely to happen when samples are sequenced in different laboratories. One of the studies of this thesis (chapter 14) scans HT between 59 *de novo* genomes, all assembled in our laboratory, this is why we added several steps to decontaminate these assemblies.

The matter of decontamination is actually not that trivial. No decontamination step might lead to the detection of false HT, yet a step of stringent decontamination might delete real portions of the genome, but also biological parasites that would be interesting to investigate. It is thus debatable what version of a genome should be made publicly available online.

5.2 . Methods to identify HT

Methods allowing the detection of HT are usually separated into two groups: parametric and phylogenetic methods. Parametric methods, such as the comparison of the GC content (see figure 5.1, panel B), search for sequences which are different from the genomic average (Ravenhall *et al.* 2015). These methods can only work for HT between species harboring very different genomic structures, such as bacteria and eukaryotes. This is why parametric methods are not suitable in this thesis, in which we investigate HT between animals, so between species with a similar genomic structure. To note when using parametric methods, is that differences between the recipient genome and the acquired sequence tend to disappear with time, since they both undergo the same mutational processes after the transfer (Lawrence and Ochman 1997). Because of this, parametric methods can only detect recent HT, between very distant species.

Phylogenetic methods became possible with the improvement of sequencing methods and the public availability of many assemblies. They rely on sequence alignments in order to figure out whether there is an inconsistency between a DNA sequence and species evolutionary history (Ravenhall *et al.* 2015). This inconsistency can be directly visible when comparing the tree of a DNA sequence with its species tree (see figure 5.1, left of panel A). Some tests are able to decipher whether the DNA sequence tree is statistically different from the species tree. However, the most common methods are probably the implicit ones, which compare sequence similarities. Indeed, we expect the acquired sequence to be less divergent in both species than the divergence between the rest of their genomes (see figure 5.1, middle of panel A). For this, the synonymous distance (dS) of the putative HT between both species is calculated and compared to the distribution of the synonymous distance between the core genes of both species (see figure 5.1, right of panel A) (Schaack *et al.* 2010).

Some automated methods were developed to facilitate the research of HT. VHICA (Vertical and Horizontal Inheritance Consistence Analysis) focuses on HTT and used the method described hereinabove, calculating dS, but with a significant improvement: VHICA takes into account the Codon Usage Bias (Wallau *et al.* 2016). In fact, synonymous substitutions are not totally neutral for a number of genes, as some genes experience a substantial purifying selection at the mRNA and translational level, which generates a Codon Usage Bias (CUB) all along the coding region. A gene with a high CUB will have a lower dS than a gene with a low CUB, even though both genes are evolving vertically from the same common ancestor. VHICA calculate the correlation between CUB and dS among 50 core genes that are assumed to be vertically transmitted. TE are then mapped on this reference CUB-dS relationship, instead of just looking at the dS distribution. CUB is higher in highly expressed genes, this is why TE usually have a low CUB. For this reason, VHICA increases the statistical power, detecting more HTT. VHICA was developed to detect HTT between related species, and was tested among 20 *Drosophila* genomes.

Another automated software, AvP (Alienness vs Predictor), detects HGT in a species of interest, based on the alignment of its proteome against any NCBI protein library (Koutsovoulos *et al.* 2022). It extracts all the information needed to produce input files to perform phylogenetic reconstruction, evaluates HGT from the phylogenetic trees, and can combine multiple other external information for additional support (e.g. gff3 annotation file, transcript quantification file).

5.3 . The case of large-scale studies

In the case of large-scale studies, additional points can be risen. The first one is the problem of heterogeneity of genome qualities in the dataset. For a same number of HGT, the probability to find HGT in a half complete assembly is divided by two. Li *et al.* (2022) did not find any correlation between genome completeness and the number of HGT, yet their result is probably due to a dataset containing mostly genomes of high completeness (an average of 98.2% $\pm$ 3.6 of complete BUSCO genes). It is important to check for this absence of correlation in studies, and to be careful when such a correlation exists. In the case of TE, their high copy number makes them easier to detect, even in incomplete genomes. Large scale studies on HTT are thus probably less impacted by genome completeness than studies on HGT. However, genome fragmentation might be a bigger problem for studies on HTT, since fragmented genomes are often due to the inability to resolve repeat sequences by the assembler (Peccoud *et al.* 2018). The use of long read sequencing was shown to significantly improve genome assemblies, decreasing fragmentation.

The second point to arise in large-scale studies is the count of independent HT events. All genomes of a dataset are not independent, since they share the same evolutionary history prior to their last common ancestor. A HT that took place millions of years ago, will be found in all the descendant species of the dataset. Yet, only a single HT took place. In the same way, a single HTT event will lead to several hits of TE copies resulting from this event. Thus clustering steps are very important when working on such studies.

In any case, it is impossible to retrieve all HT events, this is why current studies give a minimum number of HT events that took place in their dataset. Researchers thus tend to be conservative with the number of events, with the consequence of under-estimating the true number of HT, instead of over-estimating it. For this reason, comparison in terms of absolute frequencies should be avoided. Nonetheless, relative comparison between two large groups (to test a factor for example) can be considered if there is no reason to think that one group contains all the best (or worse) genomes. Thus, we can be quite confident on the interpretations of large-scale studies investigating factors that might favor HT.

5.4 . Dating HT events

The age of the transfer might be estimated by looking at the nucleotide divergence between the sequences of both species. However, such a distance might lead to an overestimation of the age of HT if (i) the two species have not directly exchanged the DNA sequence (but acquired these from a third party) or if (ii) the sampled species diverged from the real donor before the transfer (Peccoud *et al.* 2017). Therefore, the precision of the estimate really depends on our sampling. For the same two reasons, when we observe a HT, we cannot argue that there was a direct transfer between both sampled species, but only that a HT took place at some point in an ancestor of the first species, from an ancestor or a related species of the second species.

Interestingly with HTT, it is possible to estimate the date of the transfer more precisely thanks to the particularity of TE to amplify directly after a HT (Schaack *et al.* 2010). This dating method compares the divergence between TE copies and the founder copy. The founder copy is assimilated to the consensus sequence of all the TE copies. In the case of retrotransposons, the age can also

be estimated based on the divergence between their two long terminal repeats (LTRs) (Wallau *et al.* 2018). Indeed, once a new retrotransposon copy is inserted, its LTR are identical and they then accumulate mutations independently.

Importantly, the age of the detected HT is biased since current methods cannot retrieve very old HT, for which DNA sequences diverged too much to be detected as homologous by algorithms. Most algorithms cannot detect homology beyond 30-40% of divergence (Peccoud *et al.* 2018). For example, Peccoud *et al.* (2017) estimated that the HTT events they recovered between insects took place in the last 10 million years only. For the same reason, the divergence between very diverged species is not calculable, thus we cannot detect HT by comparing the dS of the putative HT to the distribution of dS of core genes. However, if the species are that divergent, we can confidently conclude to a HT for sequences with high similarity. Oppositely, it is not possible to detect HT between species which are too related, because HT DNA sequences will not have had enough time to diverge enough from the rest of the genomes to be detectable by these methods.

5.5 . The case of *de novo* HT

The methods described earlier allow the detection of HT that took place in an ancestor of the sequenced individual, yet it is also possible to look for *de novo* HT, *i.e.* HT that took place in the somatic cells of the sequenced organism. In this case, the similarity of sequences between the insert and the donor will be of 100% and only few cells will harbor the integration event. To detect such HT, one can look for *de novo* junctions between the genomes of two species. Such junctions can be detected either with paired-end reads (one read align on the recipient species while the paired read aligns on the genome of the donor species, see figure 5.1D), or with chimeric reads (see figure 5.1C). A chimeric read is a read for which one extremity aligns on the genome of the recipient species only (so not on the genome of the donor species), and the other extremity aligns on the genome of the donor species only (so not on the genome of the recipient species).

Paired-end reads and chimeric reads were firstly developed to detect *de novo* transposition (Gangadharan *et al.* 2010; Miyao *et al.* 2012; Gilly *et al.* 2014). Relying on the same principles, Gilbert *et al.* (2014) developed a method looking for chimeric reads to detect HT from a lepidopteran to a virus after infection. I adapted this method in part II of this thesis in order to detect *de novo* HT from parasitoid wasps to their lepidopteran hosts during parasitism.

5.6 . The direction of the transfer

It is often difficult, if not impossible, to decipher the direction of the transfer (which is the donor species and which is the recipient), although some fortuitous 'tagging' may exist. For example, Graham and Davies (2021) were able to determine the direction of a transfer with confidence (a gene transferred from herring to smelt that confers a better resistance to cold) thanks to some accompanying transposable elements. In the specific case of polydnaviruses, one can guess that it was transferred from wasps to its host, thanks to the specific pattern left by its mechanism of insertion: the DRJ motif in the sequence (as used by Gasmi *et al.* (2015)), and the J1 and J2 motifs at the extremities (see figure 2.4). Otherwise, one can sometimes guess the direction by parsimony with the topology of the transferred DNA tree, like in figure 5.1A where it seems that the gene went

from a bacteria to a beetle. Yet, depending on the topology of the tree and the representativeness of the dataset, it is not always that obvious. For example in figure 12 of the article n°5, the wasp sequences are nested in the Lepitoperan sequences, yet this is simply due to do the fact that the dataset is composed of many Lepidopteran genomes but very few wasp genomes. The direction of the transfer is from wasp to Lepidoptera despite this misleading topology. Another possibility with TE is to estimate the age of the founder copy in both genomes, as discussed above, the genome with the younger one being the recipient species (Wallau *et al.* 2018).

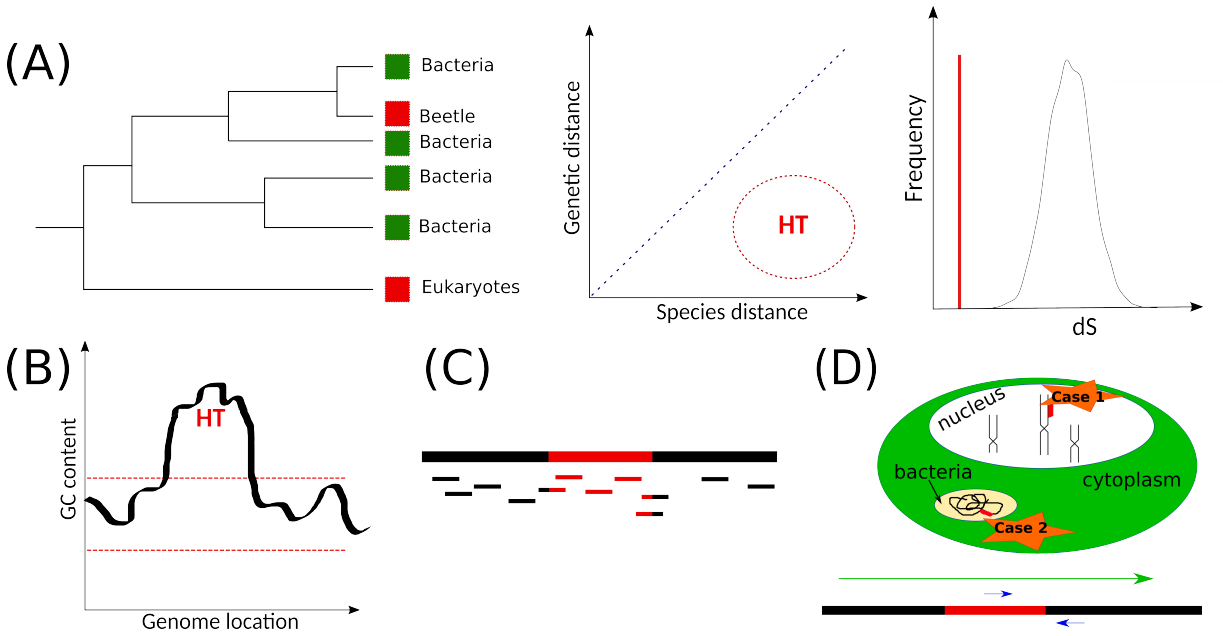


Figure 5.1: **Overview of some methods used to detect HT. A.** Phylogenetic methods. On the left, a phyogenetic incongruence between species and genes. Here, a beetle gene clusters with bacterial genes, which suggest HT. In the middle, an unexpected low divergence of a sequence in comparison with the species distance suggest a HT. The threshold to conclude to a HT can be calculated thanks to synonymous distance, as shown on the right. There, the dS of a sequence (in red) falls outside the distribution of dS of the core genes (in black). **B.** An example of parametric methods: a GC content in a window of the genome drastically different from the rest of the genome suggests HT from an organism harboring a highly different GC content profile than the recipient species. **C.** Detection of *de novo* HT thanks to chimeric reads: the inserted sequence (in red) is not present in the reference genome of the recipient species (in black), so the reads mapping, entirely or partially, on the red part do not map on the reference genome. A chimeric read is a read for which one extremity maps *only* on the reference genome and the other extremity maps *only* on the donor genome. Such reads indicate the presence of a *de novo* insertion. **D.** Methods to decipher between HT and contamination. On top, fluorescence *in situ* hybridization uses a fluorescent probe (in red) complementary to the putative HT to directly observe whether the probe hybridize on the chromosome of the recipient (case 1, HT confirmed), or of the donor (case 2, contamination). On the bottom, long reads (green arrow), or paired-end reads (two blue arrows) or PCR (also blue arrows) can validate whether the candidate HT (in red) is flanked by the recipient genome (in black).

6 - Goals and context of the thesis

Although there is now a consensus that horizontal transfers do take place between multicellular organisms, their abundance and mechanisms are still at the stage of speculations. During this thesis, I investigated these two aspects, focusing on HT between animals only, with an emphasis on insects. Insects are a very good model since they are very diversified, they are numerous, they are relatively easy to capture in the wild, it is possible to rear them in a laboratory, and they have a short generation time. My work is organized around three parts. Part I is entitled "Free viruses as vectors of horizontal transfers". Here I shortly investigated whether free viruses can act as vector of HTT in laboratory conditions, *i.e.* transport DNA from one organism to another. Part II is entitled "Domesticated viruses as vectors of horizontal transfers from parasitoid wasps", in which I investigated whether polydnaviruses can act as vectors of HT. This part also allowed me to study a parasitoid/host interaction, which could promote horizontal transfers. Finally, part III is entitled "Factors influencing horizontal transfers". In this part, I am working on two large-scale studies, yet both were still in progress at the time of the writing of this manuscript. In the first one, I was evaluating the strength of the aquatic habitat and the external fertilization as factors shaping global trends of HTT in animals. Because of the scarcity of fully aquatic insects and the non-existence, to my knowledge, of external fertilization in this group, I chose to work more generally with animals for this study, sampling genomes on NCBI that cover a maximum number of transitions of habitats and modes of fertilization. In the second study, I was working with insects in order to evaluate the strength of geographical proximity and phylogeny proximity on HTT thanks to a dataset that we produced specifically for this study. In both studies, I am focusing only on horizontal transfers of transposable elements (HTT) because based on previous studies we expect to find numerous such events, enabling us to perform formal quantitative analyses of their distribution across species.

Part I is part of the TransVir project (ANR-15-CE32-0011-01), in collaboration with the university of Poitiers, the goal of which is to assess whether free viruses can act as vectors of HT between animals.

Part II is part of the project CoteBIO (ANR17-CE32-0015-02), the goal of which is to investigate whether *Cotesia typhae* (Hymenoptera: Braconidae) could be used as a bio-control agent in France against *Sesamia nonagrioides* (Lepidoptera: Noctuidae), the Mediterranean corn borer, which is a major pest in Mediterranean regions and Sub-Saharan Africa. Pictures of both species can be observed in figure 2.3). CoteBio aims to (i) evaluate the potential risks of the introduction of *C. typhae* in France, which I contributed to during this PhD, (ii) analyze the behavior and the variability of the reproductive success of *C. typhae*, (iii) test the efficiency of *C. typhae* in semi-natural conditions, and (iv) develop methods for massive rearing. *S. nonagrioides* is structured into four populations: West Africa, Center-Africa, East-Africa, and paleartic along the Mediterranean sea from Spain to the Middle East (MOYAL *et al.* 2011). Larvae dig tunnels in the corn stems during their whole larval stage, which greatly impede corn productions. Current methods to protect the fields rely on chemical pesticides and transgenic plants such as Bt maize that expresses insecticidal proteins (Farinós *et al.* 2018). However, a resistance to the toxin was identified (Camargo *et al.* 2018) and some countries, like France, do not authorize transgenic organisms anyway. Using a bio-control agent could be a

reliable alternative, which is currently under study by the team "Ecology and Evolution" of EGCE, using the parasitoid wasp *C. typhae* as agent. This species was recently described in East-Africa (Kaiser *et al.* 2017). Contrary to its closely related species *Cotesia sesamiae*, *C. typhae* is a specialist species, both at the plant (*Typhae domingensis*) and the host (*S. nonagrioides*) levels. In East-Africa, *C. typhae* enter the *Typhae domingensis* stems in order to inject dozens of eggs in *S. nonagrioides* larvae (step (1) in figure 2.3). These eggs hatch in the caterpillar, before feeding from its hemolymph. About two weeks later, wasp larvae leave the caterpillar (step (2) in figure 2.3), and form cocoons just after. Because of the high mortality rate caused to infected caterpillars, and the specificity of the host, *C. typhae* is a really good candidate for a bio-control agent. However, possible non-target effects have to be meticulously investigated first, one of which being the possibility of HT between the imported populations of *C. typhae* to french lepidoptera, like *S. nonagrioides* but also possibly to non-target species. Since *C. typhae* belongs to the microgastroid complex of braconid wasps, its genome contains a bracovirus (a polydnavirus). Thus, this project was a very good opportunity to study how polydnaviruses may facilitate HT among insects.

Part III is composed of two independent large-scale studies. The first one was decided during my PhD and is not part of any larger project, while the second one is part of the TranspHorizon project (ANR-18-CE02-0021-01). The goal of this ANR project is to evaluate the impact of geographical and phylogenetic proximities on the success of TE invasion in genomes following a HTT, and to identify parameters influencing this success (population size, dynamics of the defensive system, and age). This project uses insects as a model, and mostly *Drosophila* for the parts on experimental evolution.

Part I

Free viruses as vectors of horizontal transfers

7 - Article n°1: Assessing the Impact of a Viral Infection on the Expression of Transposable Elements in the Cabbage Looper Moth (*Trichoplusia ni*)

The first version of this article, to which I did not contribute, was submitted to GBE at the end of the PhD of Vincent Loiseau (at the end of 2020). His PhD being over when he received the comments from the reviewers, I entirely took care of the revision, which entailed major modifications. This is why I became co-first author. Only the parts "TE Identification and Database" and "Expression of AcMNPV-Borne TE Copies" did not change. In the other parts, I had to redo the analyses, rewrite the article and design entirely new figures. In this article, we used two published datasets of RNAseq to assess the impact of a viral infection on the expression of TE in *Trichoplusia ni*, the cabbage looper moth. The two goals of this study were to determine whether a viral infection could increase TE activity, which would increase the chance of insertion in the virus, and to assess whether viral-borne TE can be expressed, which would increase their chance to be inserted again in another insect genome during a second step of infection.

We found a moderate impact of AcMNPV infection on TE expression in *T. ni*, although potentially sufficient to affect TE activity and genome architecture. Interestingly, we found a host-derived TE integrated into AcMNPV genomes, which is highly expressed in infected cells. This suggests that virus-borne TE may be able to transpose. This result supports the hypothesis according to which free viruses may act as vectors of horizontal transfer of TE in insects.

Assessing the Impact of a Viral Infection on the Expression of Transposable Elements in the Cabbage Looper Moth (*Trichoplusia ni*)

Héloïse Muller[1,†], Vincent Loiseau[1,†], Sandra Guillier[1], Richard Cordaux[2], and Clément Gilbert [1,*]

[1]Universite Paris Saclay, CNRS, IRD, UMR Evolution, Genomes, Comportement et Ecologie, Gif-sur-Yvette, France

[2]Laboratoire Ecologie et Biologie des Interactions, Equipe Ecologie Evolution Symbiose, Universite de Poitiers, CNRS, France

[†]These authors contributed equally to this work.

*Corresponding author: E-mail: clement.gilbert@egce.cnrs-gif.fr.

Accepted: 4 October 2021

Abstract

Most studies of stress-induced transposable element (TE) expression have so far focused on abiotic sources of stress. Here, we analyzed the impact of an infection by the AcMNPV baculovirus on TE expression in a cell line (Tnms42) and midgut tissues of the cabbage looper moth (*Trichoplusia ni*). We find that a large fraction of TE families (576/636 in Tnms42 cells and 503/612 in midgut) is lowly expressed or not expressed at all [$\leq$ 4 transcripts per million (TPM)] in the uninfected condition (median TPM of 0.37 in Tnms42 and 0.46 in midgut cells). In the infected condition, a total of 62 and 187 TE families were differentially expressed (DE) in midgut and Tnms42 cells, respectively, with more up- (46) than downregulated (16) TE families in the former and as many up- (91) as down-regulated (96) TE families in the latter. Expression log2 fold changes of DE TE families varied from −4.95 to 9.11 in Tnms42 cells and from −4.28 to 7.66 in midgut. Large variations in expression profiles of DE TEs were observed depending on the type of cells and on time after infection. Overall, the impact of AcMNPV on TE expression in *T. ni* is moderate but potentially sufficient to affect TE activity and genome architecture. Interestingly, one host-derived TE integrated into AcMNPV genomes is highly expressed in infected Tnms42 cells. This result shows that virus-borne TEs can be expressed, further suggesting that they may be able to transpose and that viruses may act as vectors of horizontal transfer of TEs in insects.

Key words: mobile elements, dsDNA virus, AcMNPV, horizontal transfer, Lepidoptera.

Significance

How an infection by a double-stranded DNA virus affects the expression of transposable elements (TEs) has never been studied. Here, we show that most TE families annotated in the genome of the cabbage looper moth are not expressed in the uninfected condition and that 62 and 187 TE families are differentially expressed (DE) in infected conditions in midgut tissues and a cell line, respectively. We demonstrate that moth TEs integrated into AcMNPV genomes can be co-expressed with their neighboring gene. Our results show that the impact of an infection by AcMNPV on moth TE expression is moderate but potentially sufficient to influence the TE activity and genome architecture. They also reveal that virus-borne TEs are likely able to transpose to another DNA target.

Introduction

TEs are selfish genetic elements able to move in the genome of their hosts and that 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 (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 et al. 2017). However, like many other mutation types, most transposition events are neutral or harmful and are thought to negatively impact host fitness (Brookfield and Badge 1997; Barrón et al. 2014; Mita and Boeke 2016). In response to the deleterious effects of TEs, several TE-repressing mechanisms have evolved in host genomes, such as DNA methylation, histone modifications, or posttranscriptional repression through the PIWI-interacting RNA pathway (Slotkin et al. 2007; Deniz et al. 2019). Thus, host–TE interactions are often referred to as an arms race and they often result in the complete extinction of TE families and degradation of TE copies that are eliminated from the genome with time, mainly due to the neutral evolution of most TE sequences (Le Rouzic et al. 2007; Blumenstiel 2019).

Typically, few TE families are expected to be transpositionally active in a genome, most of them being repressed and thus not expressed (Yoder et al. 1997; Zilberman et al. 2007). However, perturbations of genome stability, environmental changes, or infection can lead to a stress-mediated modulation of TE expression (Miousse et al. 2015). Several examples of TE activation 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; Zovoilis et al. 2016; Huang et al. 2017; Hummel et al. 2017; Ryan et al. 2017; Dubin et al. 2018). Such activation is often thought to be caused by epigenetic modifications or activation of transcription factors (Capy et al. 2000; Horváth et al. 2017). Interestingly, some TEs even bear a stress response element, that is, a regulatory sequence activated in response to a stress, enabling TEs to be upregulated in stressful conditions (Bucher et al. 2012; Casacuberta and González 2013). However, the impact of stress on TE expression appears to be 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 et al. 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 in constitutive heterochromatin, which is generally associated with gene-poor, TE-rich regions (Trojer and Reinberg 2007; Saksouk et al. 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 single species (Barah et al. 2013; Niederhuth et al. 2016; Fouché et al. 2020).

In the study of eukaryotic TE response to stress, most efforts focused on plants. To our knowledge, few studies have investigated the impact of a biotic stress like a viral infection on 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 et al. 2020). This pattern was observed particularly near antiviral response genes and was common to all analyzed data sets, whatever the virus type, the host species 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 the single-stranded RNA Sindbis virus (SINV) on *Drosophila simulans* and *Drosophila melanogaster* flies (Roy et al. 2020). It was found that viral infection can modulate the piRNA and siRNA pathways known to be involved in TE expression control. In turn, a global decrease in TE transcript amounts was observed in *D. simulans* and *D. melanogaster* flies during the exponential phase of SINV replication. Overall, these studies suggest that viral infection can affect TE activity in animals.

Interestingly, several other studies reporting host TEs integrated in baculovirus genomes provide direct evidence that some TEs can be active during a viral 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 the 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 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 indicated that many class I and class II 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, noninfected 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, copy number, and position in the host genome on the level of host TE expression remains unclear. Finally, whether TE copies integrated into viral genomes are expressed during infection has never been assessed.

Here, we addressed these questions by reanalyzing two published time course RNA-seq data sets that were initially produced to measure variation in gene expression levels of the moth *Trichoplusia ni* in response to an infection by the AcMNPV baculovirus (Chen et al. 2013; Shrestha et al. 2018). These experiments were carried out in the Tnms42 cell line and in the midgut of *T. ni* fifth instar larvae. We found three times more DE TE families in Tnms42 cells than in *T. ni* midguts. One of the *T. ni* TE families previously found inserted in AcMNPV genomes (TFP3) was particularly overexpressed in infected Tnms42 cells compared to all other TE families, likely due to the expression of copies inserted into AcMNPV genomes. Overall, our study shows that infection by AcMNPV affects the expression of a moderate number of TE families in Tnms42 cells and *T. ni* midgut. It further reveals that TEs inserted in viral genomes can be transcribed, in agreement with the possible role viruses may play as vectors of horizontal transfer of TEs between insects.

Results

TE Landscape in *T. ni* Genomes

We first annotated TE copies de novo in the two *T. ni* genomes used in this study [HighFive (Hi5) germ cell line and larva]. The TE landscape of the *T. ni* larva genome was not characterized in the original publication (Chen et al. 2019). Using RepeatMasker and a library of 847 TE consensus sequences obtained through various searches (see methods), we masked 231,670 copies (or TE fragments) that make up 8.66% of the *T. ni* larva genome. These copies were masked by 702 out of the 847 TE consensus. DNA TEs were the most abundant with 126,660 copies (54.7%), including 11,511 Helitron copies, followed by LINEs (85,814 copies, 37%) and LTRs (19,196 copies, 8,3%). The most abundant superfamilies were the class II DNA/PIF-Harbinger (48,702 copies), the class I LINE/L2 (46,890 copies), and the class II DNA/mariner (22,017 copies), which collectively accounted for half of all TE copies (fig. 1*a*). On the contrary, some superfamilies were less abundant, like DNA/MuLE (835 copies), DNA/PiggyBac (1,111 copies), or LINE/Dong (1,181 copies). The overall nucleotide divergence between copies and consensus sequences ranged from 0% to 41.3% (median 15%). A total of 3,103 copies (1.34%) were identical to their consensus (0% divergence).

To map RNA-seq reads on TE copies less fragmented than those produced by our automatic TE annotation procedure, we ran the tool "One Code to Find them All" (Bailly-Bechet et al. 2014). We used the options –unknown and –strict, filtering copies greater than 80 bp in length and with more than 80% identity with the consensus. The filtered TE landscape contains 66,683 copies masked by 612 TE consensus and making 8.25% of the *T. ni* larva genome. These copies contain 55.7% of DNA TEs, 39.9% of LINEs and 4.5% of LTRs (fig. 1). This filtering and aggregation step decreased the number of TE copies by a factor of 3.5 but the overall size of the resulting TE copies is relatively similar to the total size of the nonfiltered TE copies (28.8 vs. 27.5 Mb).

De novo TE annotation of the *T. ni* Hi5 germ cell line genome was previously done (Fu et al. 2018). However, to facilitate comparison between expression profiles of the *T. ni* cell line and midgut, we performed our own TE annotation of this genome using the same pipeline as the one used for the *T. ni* larva genome. We annotated 11.5% and 9.98% of the cell line genome as TEs before and after aggregating and filtering copies, respectively. This is similar to the figure obtained by Fu et al. when excluding SINEs (9.41%). We retained copies masked by 636 consensus sequences after filtering. The TE landscape was overall very similar to that of the larva genome, with 53.9% of DNA TEs, 37.1% of LINEs, and 9.0% of LTRs before filtering, versus 55.0% of DNA, 40% of LINE, and 5.0% of LTR after filtering.

Genome-Wide TE Differential Expression during AcMNPV Infection of Tnms42 Cells

Reads produced by Chen et al. (2013) were mapped on the 75,680 TE copies annotated in the *T. ni* Hi5 genome and the differential expression was computed by TE consensus, which we consider here as each representing a separate family (636 TE families included in this analysis). Among the 636 TE families, 576 are considered as not or very lowly expressed in the mock condition [transcripts per million (TPM) < 4]. The median for TE expression in the mock is 0.37 TPM and the maximum is 65.6 TPM for Tni_Contig_13_Harbinger. Thus, few TE families are expressed in the mock condition in Tnms42 cells, with only two of them being highly expressed, that is, their average TPM is higher than 50.

The number of DE TE families varied from 53 (8.3% of all TE families) to 94 (14.8%) depending on the time point during the course of the AcMNPV infection in *T. ni* cells (fig. 2 and supplementary table S1). When considering all time points together, a total of 187 TE families (29% of all families) were found to be DE during at least one time point. Overall, the strength of differential expression went from 30-fold decrease to 553-fold increase (log2 fold change $=$ -4.95 to 9.11), with a median over all DE TE families and all time points at -2.01 log2 fold change and an average at 0.07. Among all 187 DE TE families, 91 and 96 were up- and downregulated during at least one time point, respectively, and one was alternatively down- and upregulated during the

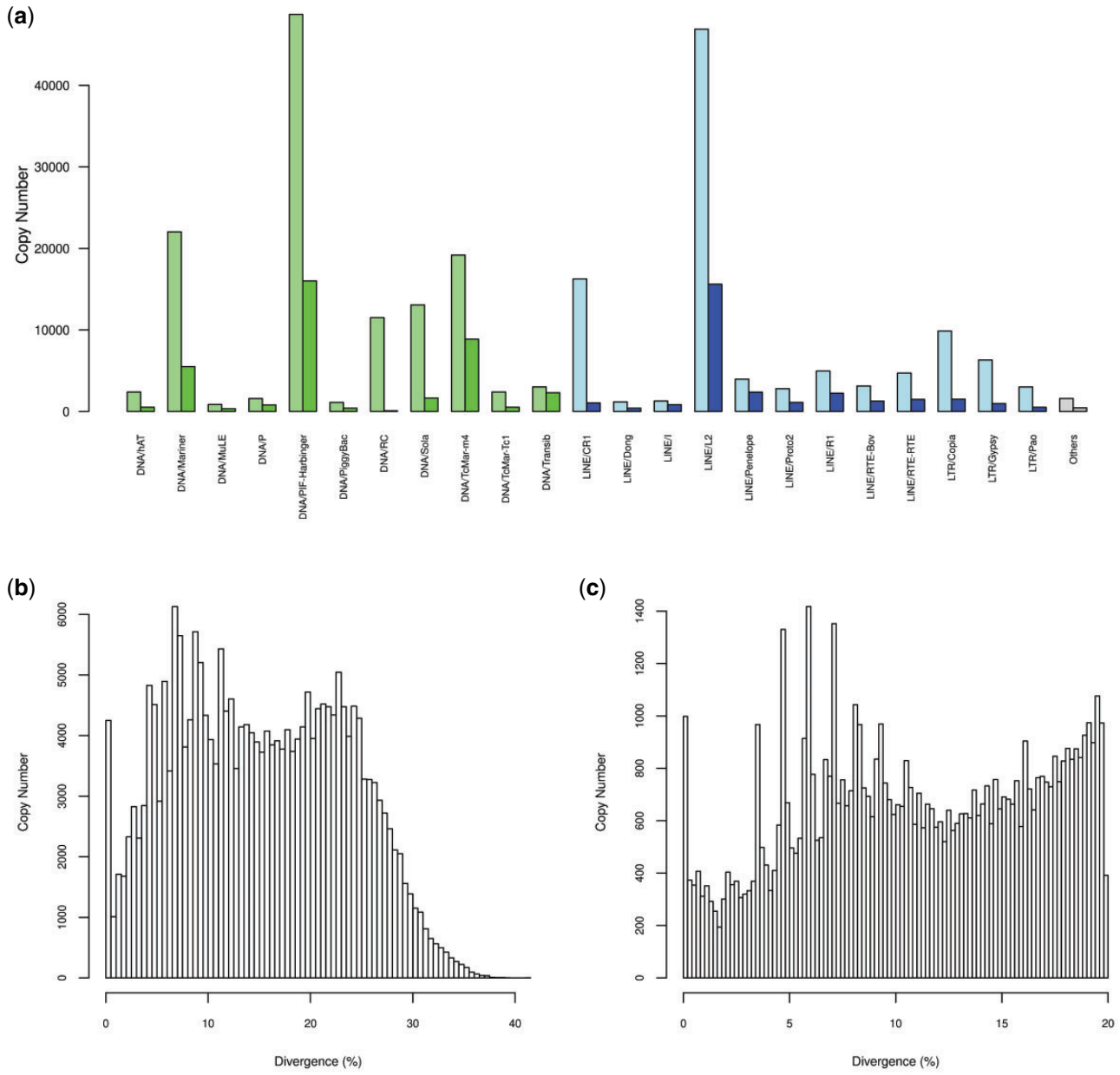


Fig. 1.—Transposable element landscape of the *T ni* larva genome. (*a*) Copy number of the different TE superfamilies detected before filtering (light colors on left) and after filtering (bright colors on right). Class I TE superfamilies are in green, class II TE superfamilies are in blue, and superfamilies with low copy number (<115 copies) are in gray. (*b*) Histogram of observed TE copy nucleotide divergence to consensus for the nonfiltered 702 TE families. (c) Histogram of observed TE copy nucleotide divergence to consensus for the 614 TE families included in the study, after filtering.

course of the experiment (table 1 and fig. 2). Among the 91 upregulated TE families, 8 were induced, that is, they showed no or very low expression in the mock (TPM < 4) and were expressed in at least one time point in the infected condition (TPM > 4), including two TE families (LINE/R1_8 and LINE/Proto_3) that became highly expressed (TPM ≥ 50). Among the 96 downregulated TE families, 7 can be considered repressed in at least one time point, that is, their TPM was higher than 4 in the mock and less than 4 in the infected condition. No repressed TE families were highly expressed in the mock (TPM ≥ 50). Altogether, these observations show that infection by AcMNPV moderately affects the expression of a substantial proportion of TE families in the *T. ni* Tnms42 cell line genome, with a similar number of up- and downregulated TE families.

The time course RNA-seq data produced by Chen et al. (2013) shows that a number of TEs (8% of all TE families) are DE as early as 0 hpi, which in fact corresponds to the first

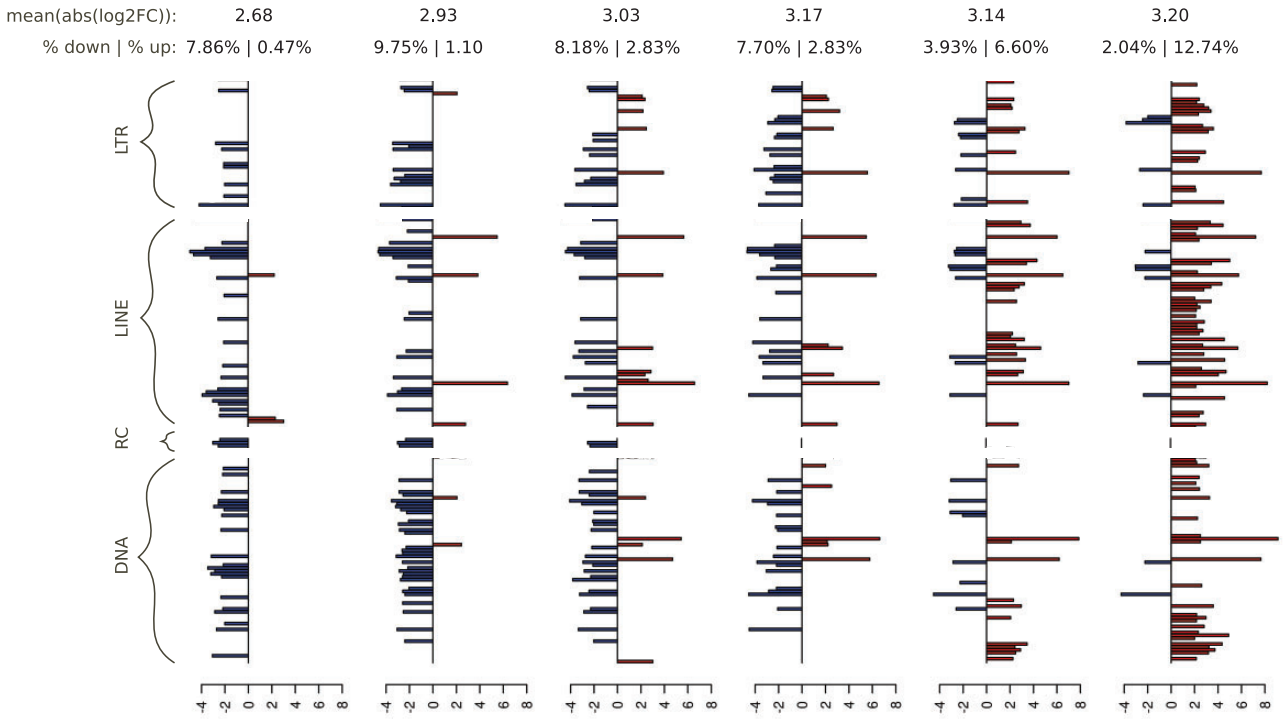


Fig. 2.—DE TE families in the Tnms42 cell line data set. Only the 187 significant DE TE families with and absolute log2 fold change superior to 2 are considered. On top, the first line shows the average absolute log2FC at each time point, while the second line shows the percentage of TE families downregulated and upregulated at each time point. For each time point, log2FC is indicated in red for upregulated TE families or in blue for downregulated ones. The 187 DE TE family names can be found in supplementary table S1.

Table 1
DE TE families in the Cell Line and Midgut Data Set for Each Class

Data Set	TE Families		DE TEs	LTRs	LINEs	DNA	Helitron
Cell line	638	Upregulated	91	20	41	30	0
		Downregulated	95	23	29	40	3
		Other[a]	1	0	1	0	0
Midgut	614	Upregulated	46	23	7	16	0
		Downregulated	16	4	3	9	0

[a]This TE is alternatively down- and upregulated during the course of the experiment.

1-h incubation period of the cell line with the virus. As observed for host genes (Chen et al. 2014), the impact of the viral infection on TE expression is thus rapidly measurable. The number of TE families impacted by the viral infection is then quite stable through time until 36 hpi, followed by a slight increase afterwards (around 10.5–11% of TE families are DE from 6 to 36 hpi, and 14.8% at 48 hpi; fig. 2). However, the direction of the impact varies with time after infection, with the majority of DE TE families being downregulated early after infection (from 0 to 18 hpi) and upregulated at later time points (fig. 2). Contrary to TEs which are relatively stably affected by the infection throughout the experiment, the impact of AcMNPV infection on the expression of *T. ni* genes continuously increases with time, with about 20% of *T. ni* unigenes being DE at 0 hpi and 40% at 48 hpi (Chen et al. 2014). Moreover, the direction of gene differential expression is inverted compared to that of TEs, with most DE *T. ni* genes being upregulated early after infection (0 and 6 hpi) and downregulated at later time points (Chen et al. 2014). The processes underlying how genes and TE expression is affected by AcMNPV infection in Tnms42 cells are thus different.

The proportion of DE TE families was relatively similar for both retrotransposons (113 out of 354 or 31.9%) and DNA transposons (70 out of 278 or 25.2%). Furthermore, among DE TE families, the proportion of upregulated ones differed only moderately between the two types of TEs (43% and 54% for DNA transposons and retrotransposons, respectively) (table 1). Thus, overall, our results do not reveal any important difference in the way expression of the two TE classes is affected by AcMNPV infection in Tnms42 cells. Among all

downregulated TE families (95) 11 showed a log2FC lower than −4, with an extremum at −4.95 (i.e., 31-fold less expressed). Three of these were DNA transposons (DNA/TcMar-Tc_7, DNA/PIF-Harbinger_3, and DNA/PiggyBac_10) and the remaining eight were retrotransposons (LTR/Gypsy_12, LTR/Copia1_4, LINE/R1_2, LINE/R1_3, LINE/R1_4, LINE/L2_10, LINE/L2_21, LINE/Dong-R4_1). Among the most upregulated TE families, six had log2FC higher than six, with a maximum at 9.11 for DNA/TFP3 (i.e., 553-fold more expressed). Four of them were retrotransposons (LINE/Proto_3, LINE/I, LINE/R1_8, and LTR/Gypsy_11) and the two others were DNA transposons (DNA/TFP3 and DNA/PiggyBac_9). DNA/TFP3 particularly stood out among these upregulated TE families because of its remarkably high expression level. From an expression of 7.8 TPM in the mock, it reaches 52 TPM at 6 hpi and 802 TPM at 48 hpi in the infected condition (fig. 3*A*). For comparison, the next highest expression level after TFP3 in the infected condition is 104 TPM. TFP3 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).

Out of 94 TE families found inserted in AcMNPV genomes in previous studies (Fraser et al. 1983; Jehle et al. 1998; Gilbert et al. 2014, 2016), 41 passed our filters to be included in our library (i.e., best match to a TE protein over at least 50% of the length of this protein). Among these 41 TE families, 28 were found in the genome of Tnms42 cells (marked with an asterisk in supplementary file 1). Only eight of them were found to be DE (DNA/TFP3, DNA/PIF-Harbinger_4, DNA/Sola_6, DNA/hAT_1, and LTR/Gypsy_15 were upregulated, whereas DNA/Sola_2, DNA/Sola_3, and DNA/PiggyBac_2 were downregulated). Among the 20 remaining TE families, only one was highly expressed (TPM $\geq$ 50) in the mock condition and 17 were not expressed (TPM $<$ 4). Thus, there is no link between a specific TE expression pattern in infected Tnms42 cells and integration of TEs into AcMNPV in previous studies.

Genome-Wide TE Differential Expression during AcMNPV Infection of *T. ni* Larvae Midguts

Reads produced by Shrestha et al. (2018) were mapped on the 66,683 TE copies annotated in the *T. ni* larva genome and the differential expression was computed by TE consensus, which we consider here as each representing a separate family (612 TE families included here). Among these 612 TE families, 148 are expressed in at least one time point in mocks (TPM $\geq$ 4). More precisely, 464 are considered as not or very lowly expressed in mocks (TPM $<$ 4 in all time points), whereas 81 can be considered as always expressed with confidence (TPM $\geq$ 4 in all time points). Considering all time points, the median for TE expression in mocks is 0.41 TPM (compared to 0.37 TPM in the Tnms42 cell line) and the maximum is 1,847 TPM. Thus, as for the cell line data set, most TE families are not expressed at one or more time points in mock conditions, although the strength of expression is overall slightly higher in the midgut data set.

The number of DE TE families varied from 0 to 59 (9.64% of all TE families) depending on the time point during the course of the AcMNPV infection in *T. ni* larvae midgut (fig. 4 and supplementary table S2). When considering all time points together, a total of 62 TE families (10.13% of all families) were found to be DE during at least one time point. Overall, the strength of differential expression went from 19-fold decrease to 202-fold increase (i.e., log2FC = −4.28 to 7.66), with a median at 2.26 and an average at 1.31 log2 fold change. Among all 59 DE TE families, 46 were upregulated and 16 were downregulated during at least one time point (fig. 4 and table 1). Among the 46 upregulated TE families, 10 were induced, that is, they showed no or very low expression in the mock (TPM $<$ 4) and were expressed in at least one time point in the infected condition (TPM $\geq$ 4), including one TE family (LINE/Proto_1) that became highly expressed (TPM $>$ 50). Nine TE families were repressed in *T. ni* larval midgut (TPM $\geq$ 4 in mocks and TPM $<$ 4 in the infected condition). Overall, these data show that as in the *T. ni* Tnms42 cell line, AcMNPV infection moderately affects the expression of several TE families, the majority of which are upregulated in the infected condition. The impact of AcMNPV infection on TE expression is lower than in the cell line as only 10.13% of TE families are affected (compared to 29% in the cell line) and both positive and negative maximum log2 fold changes (log2FCs) are slightly lower than in the cell line.

The time course RNA-seq data produced by Shrestha et al. (2018) reveals that contrary to the Tnms42 cell line, the number of TE families affected by AcMNPV and the strength of differential expression increase with time, with no DE TE family from 0 to 6 hpi and only one (upregulated) from 12 to 18 hpi and two at 24 hpi (fig. 4 and supplementary table S2). The number of DE TE families really began to increase at 36 hpi with seven DE TE families, followed by 10 DE TE families at 48 hpi and finally 59 at 72 hpi. Regarding the strength of differential expression, 94% of DE TE families reached their extremum of differential expression at 72 hpi (vs. 41.7% in the cell line data set). In contrast to the cell line, in which TE and gene expression seemingly responded differently to AcMNPV infection, the pattern observed in larval midguts is very much similar to that of *T. ni* genes. Indeed, only very few DE genes were detected at early time points (67 in total at 0, 6, and 12 hpi), followed by medium numbers at intermediate time points (82–475 genes) and a sharp increase at 72 hpi (1,910 genes) (Shrestha et al. 2019). Thus, in midgut cells, TE and gene regulation seems to be affected in a more similar way than in Tnms42 cells.

As in Tnms42 cells, the proportion of DE TE families was similar for both retrotransposons (37/338 or 10.9%) and DNA transposons (25/269 or 9.3%). However, among DE TE

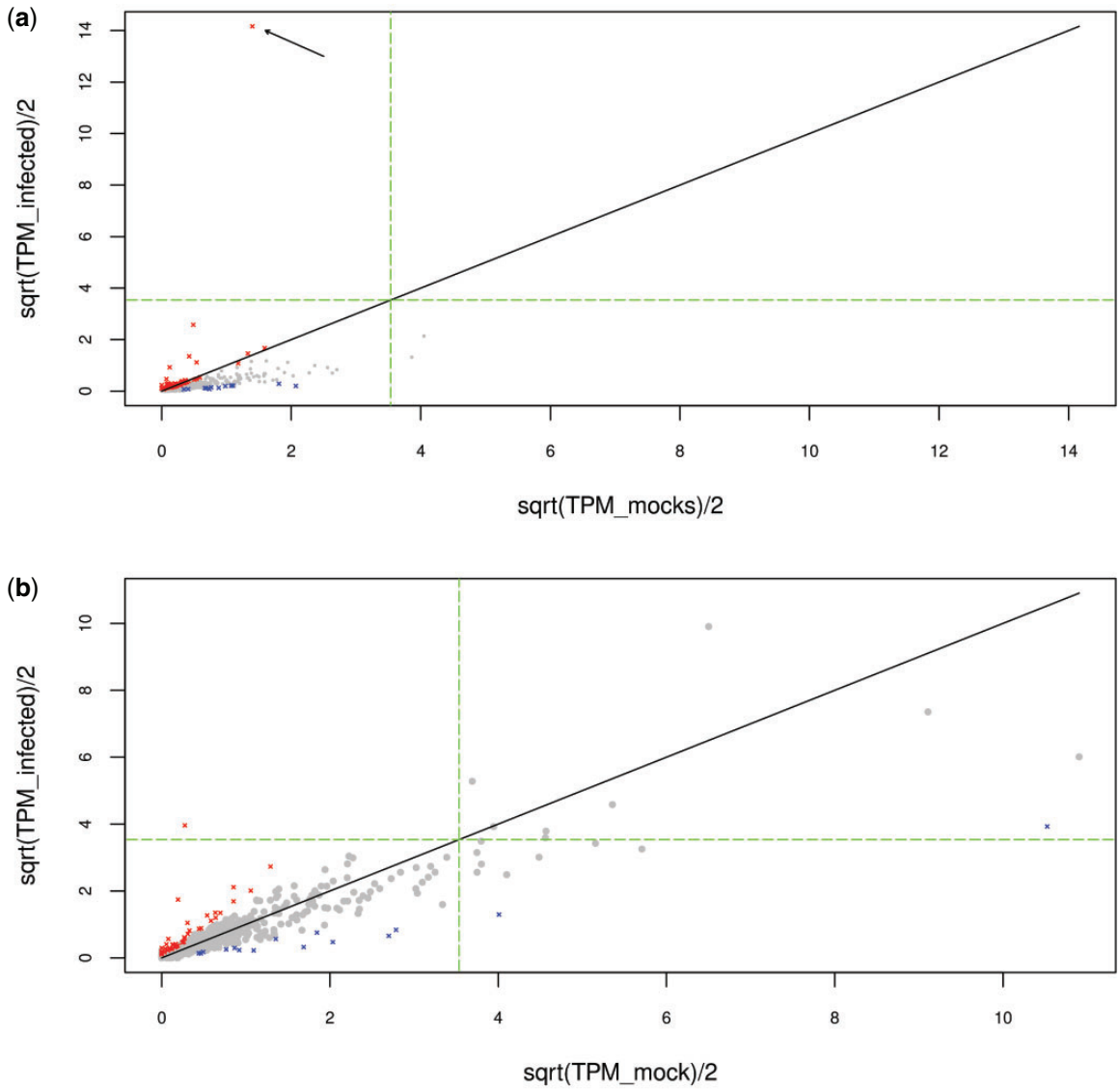


Fig. 3.—TE family expression in mock condition as a function of TE expression in infected condition. TE families that are significantly DE are colored in red (upregulated) and blue (downregulated). The green dash line represents the cutoff for high expression, at 50 TPM. (*a*) TE expression in the Tnms42 cell line data set at 48 hpi. The arrow points to TFP3. (*b*) TE expression in the midgut data set at 72 hpi. The same analysis has been performed for all time points of the two time course RNA-seq experiments and results are shown in supplementary figs. S1 and S2.

families, the proportion of upregulated ones is higher for retrotransposons (81%) than for DNA transposons (64%) (table 1). One of these retrotransposons (LINE/Proto_1) is the most upregulated TE family in *T. ni* larvae midgut, being 202-fold more expressed in the infected condition than in mock at 72 hpi (i.e., log2FC = 7.66) (fig. 4). This TE family is activated at 72 hpi, reaching 63 TPM. The two most downregulated TEs were two TcMar families, being repressed at 72 hpi (i.e., log2FC = −4.3 and −4.1 and TPM < 4), whereas in mocks they were expressed at 11 and 5 TPM at 72 hpi.

Moreover, out of 41 TE families found inserted in AcMNPV genomes in previous studies and that passed our filters to be included here (Fraser et al. 1983; Jehle et al. 1998; Gilbert et al. 2014, 2016), 23 were found in the *T. ni* larva genome (indicated with "◦" in supplementary file 1). Only one of them (DNA/Sola_3) is DE, being upregulated at 72 hpi (log2FC = 3.7). This DNA/Sola was also DE (log2FC = −2.6 at 6 hpi) in the cell line. Among the 22 remaining TE families, two were highly expressed in at least one time point (TPM ≥ 50), with a median expression of 48 and 196 TPM in mocks and a maximum of 63.3 and 331.8 TPM, respectively. As observed in the Tnms42 cell line, there is no link between a specific TE expression pattern in *T. ni* midgut and the propensity of TEs to integrate into AcMNPV genomes.

Investigation of Possible Factors Affecting TE Expression in *T. ni* Somatic Tissues

We assessed whether the expression level of TE families in the *T. ni* larvae mock condition could be associated with factors such as copy number, average proximity with genes, or age (as approximated by the average percent similarity between TE copies and the TE family consensus sequence) (supplementary fig. S3). Overall, we did not find any strong correlations with these three factors. We found a weak but significant positive correlation between TE family expression level and copy number ($r = 0.09$ at 0 hpi and $r = 0.38$ at 72 hpi in mock, $P < 0.05$ and $P < 0.001$, respectively). This correlation could be expected as it is difficult to see how increasing copy number could lead to a decrease in TE family expression level. Similarly, we found a very weak but significant positive correlation between TE family expression level and TE family age ($r \sim 0.12$ at 0 and 72 hpi, $P < 0.05$). This correlation may appear counterintuitive as one could expect that younger copies may be more likely to be functional and more strongly expressed. However, it may be partly explained by the fact that TE copy number is also weakly but significantly correlated to TE family age (supplementary fig. S3). TE family expression levels were not correlated to average proximity of TE copy with genes. Interestingly, we also noticed that TE family expression levels were correlated in the mock and infected conditions (supplementary fig. S3), reinforcing the idea that, though AcMNPV infection impacts expression of a number of TE families in *T. ni* midgut, this impact is overall moderate.

Expression of AcMNPV-Borne TE Copies

Our search for TE-virus chimeric reads revealed no such read in the RNA-seq data set 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/or 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 data set (the most deeply sequenced data set) were at a frequency lower than 0.1% and the highest insertion frequency in this data set was 1.4% (Gilbert et al. 2016). Furthermore, only a subset of these TE insertions may be cotranscribed with their neighboring gene. Thus, the absence of TE-virus chimeras in these data might not necessarily reflect absence of AcMNPV-borne TEs but such TEs might simply be expressed at levels 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 options we used does not allow finding alignments shorter than 28 bp. In addition, the average sequencing depth did not exceed 2,550× in this study. Though sufficient to detect TE insertions in principle (Gilbert et al. 2016), 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 data set from the AcMNPV-infected *T. ni* cell line (Chen et al. 2014). Considering the seven time points (six plus the 24 hpi not included in the DEseq analysis, see Materails and Methods) and the various biological replicates at each time point, 11,914 chimeric reads were identified. Among the fourteen TE families involved in these chimeras (supplementary file S2), six were found integrated into AcMNPV genomes in previous studies (Bauser et al. 1996; Fraser et al. 1996; Gilbert et al. 2016). Three class II piggybac and one Harbinger TE families were found in different replicates at different time points (table 2). The eight other TE families (seven class II and one class I) were found in a single or a just a few replicates or time points. The various TE copies found here integrated into AcMNPV genomes might result from de novo transposition from the Tnms42 cell genome or might have been present in the AcMNPV isolate used to infect these cells.

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_S.frugiperda), 22 reads aligned at the 3′ end of piggybac (22360_S.mediterranea), and insertions of Harbinger Hitchhiker 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 (supplementary fig. S4), 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 48 hpi. 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 target site duplication (TSD) motifs flanking these insertions (TTAA and ATAA) correspond to those typically generated upon transposition of piggybac elements (supplementary fig. S4; 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

Table 2

Number of TE/Virus Chimeric Reads in the Chen et al. (2013) AcMNPV RNA-Seq Data Sets

		0 h		6 h		12 h		18 h		24 h		36 h		48 h	
		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	1,158	1,300	1,017	1,060	1,017	1,076	1,055	1,023	1,380	1,488
		3		107		2,458		2,077		2,093		2,078		2,868	

NOTE.—TE families for which less than 10 chimeric reads were found in all data sets were lumped in the "Others" category. This table includes only chimeric reads mapping at the 5′ or 3′ extremity of TE families ($N = 11{,}684$).

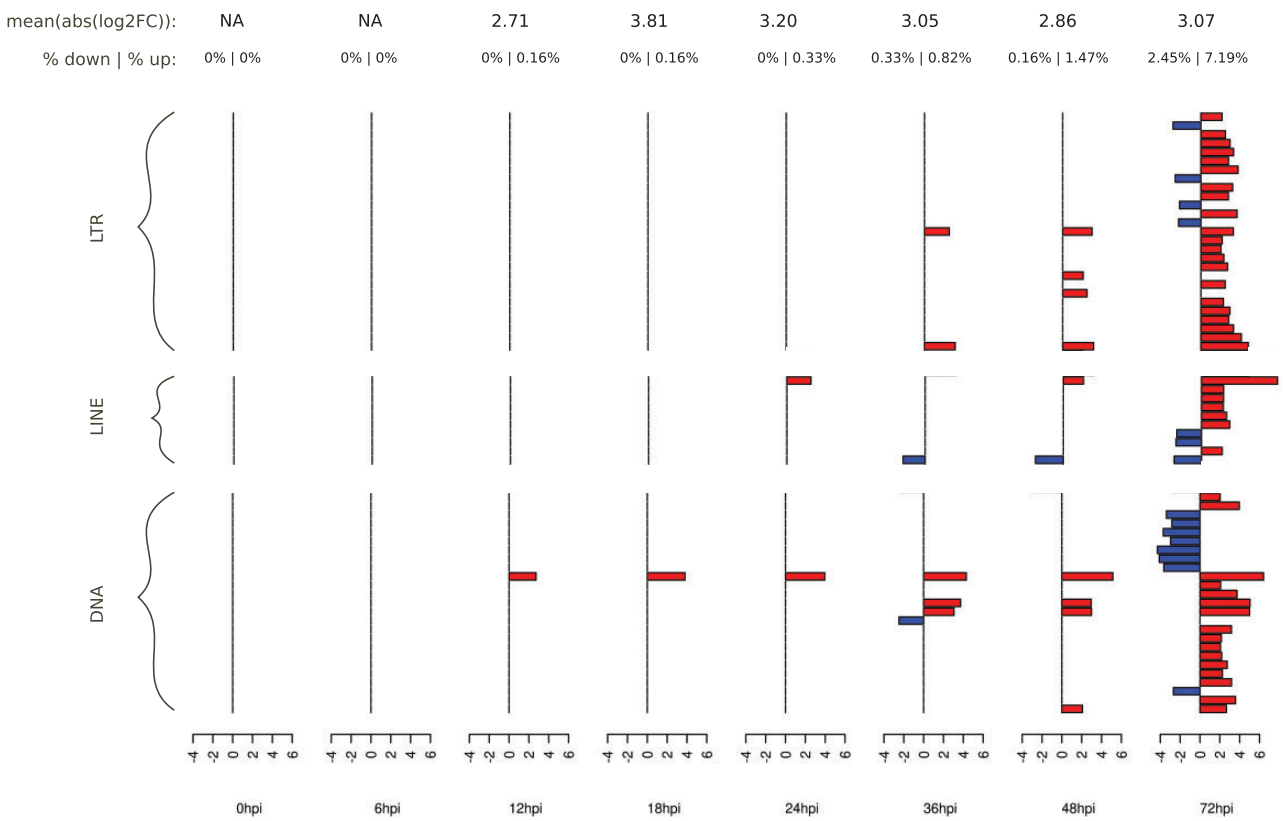


FIG. 4.—DE TE families in the midgut data set. Only the 62 significant DE TE families with an absolute log2 fold change above 2 are shown. On top, the first line shows the average absolute log2FC at each time point, while the second line shows the percentage of TE families downregulated and upregulated at each time point. For each time point, log2 fold change is indicated in red for upregulated TE families or in blue for downregulated ones. TE family names can be found in table S2.

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 at 24 hpi, 2,078 at 36 hpi, and 2,868 at 48 hpi (table 2). The peak of TE-virus chimeras detected at 12 hpi was in line with the results of Chen et al. (2014), who showed that the expression of AcMNPV genes reaches its highest levels at this time of the infection.

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 5 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* (fig. 5). To obtain further insight into the expression level of TFP3 copies inserted in these genes, we compared their expression to the overall expression of *PH*, *FP*, and *Ac-Orf78* as reported by Chen et al. (2013). Because Chen et al. (2013) measured AcMNPV gene expression levels in reads per kilobase per million reads mapped (RPKM), we also calculated the average RPKM over the three replicates for the three TFP3 copies inserted in each gene, for each time point (supplementary table S3). Importantly, TFP3 RPKM were calculated only taking the reads mapping on the virus-TFP3 junctions. Measuring the expression of viral-borne TFP3 copies over their entire length is impossible because it is impossible to assess which reads mapping internally to the TFP3 sequence come from TFP3 copies located in the *T. ni* genomes and which ones correspond to viral-borne copies. Thus, we measured the coexpression of TFP3 copies and neighboring viral genes. This is reflected by the fact that RPKM calculated for viral gene–TFP3 junctions are strongly correlated to those calculated by Chen et al. (2013) for *PH* (Pearson's rho = 0.94, $P < 0.001$) and *FP* (Pearson's rho = 0.75, $P < 0.05$) genes. In other words, the more *PH* and *FP* are expressed, the higher the expression of TPF3 copies inserted in those genes. Because only the TFP3–virus junction is considered for measuring TFP3 expression, it is likely that the true expression level of viral-borne TFP3 is much higher. Yet, it is noteworthy that even underestimated, the overall expression of the three viral-borne TFP3 copies at 12 hpi (935 RPKM) is higher than the maximum expression level reached by about 35% of AcMNPV genes during the entire duration of the experiment (see supplementary fig. 3 in Chen et al. 2013).

These results are in line with the high upregulation of TFP3 during the course of the infection we observed in our analysis of DE TEs in the cell line data set. Indeed, in the cell line data set, this TE was found to be the most upregulated TE and the most expressed in late infected conditions (fig. 3 and supplementary fig. S1). Interestingly, our results also suggest that the upregulation of TFP3 upon viral infection may be due in large part to expression of viral-borne TFP3 copies rather than to enhanced expression of TFP3 copies located in the *T. ni* genome, which would explain why TFP3 was such an outlier. If the absence of TFP3 chimeric reads in the midgut data set is biological, the absence of insertion of this TE in AcMNPV might explain why TFP3 was not DE in the midgut.

Discussion

Impact of a Baculovirus Infection on TE Expression in *T. ni*

We characterized expression patterns of TE families in midgut of larvae as well as in Tnms42 cells facing a biotic stress in the form of an AcMNPV infection. We found that the genome of *T. ni* larvae (Chen et al. 2019) and Tnms42 cells (Fu et al. 2018) has a similar percentage of DNA, LTR, and LINE families, with 636 TE families in the cell line genome and 612 TE families in the *T. ni* larvae genome, including 587 families in common. We further showed that in the mock condition at 0 hpi, a larger number of TE families are confidently considered expressed in the midgut data set (60 in Tnms42 cells vs. 101 in midgut), with 30 shared expressed TE families. Taking all time points postinfection together, our analyses reveal that a moderate number of TE families are DE in AcMNPV-infected *T. ni* Tnms42 cells (187/636 TE families) and midgut (62/612 TE families), with widely overlapping ranges of fold changes in expression in the two data sets (from 30-fold decrease to 553-fold increase in cells and from 19-fold decrease to 202-fold increase in midgut). Of note, these fold changes overlap with those calculated for non-TE *T. ni* gene expression in Tnms42 cells (from 134-fold decrease to 20.8-fold increase; Chen et al. 2014) and midgut (from 640-fold decrease to 163-fold increase, Shrestha et al. 2019) but they tend to be shifted toward stronger upregulation. Altogether, our results are in line with earlier studies and further suggest that a viral infection can affect TE activity and thus influence genome architecture in animals (Macchietto et al. 2020; Roy et al. 2020). The overall magnitude of the changes in TE family expression level is, however, moderate here, with TE expression levels in mocks being overall strongly correlated to those in infected conditions, no sign of strong global TE unleashing, and relatively small fold changes for most TE families in both data sets.

Differences in TE Expression in Response to AcMNPV Infection in *T. ni* Midgut Tissues and Tnms42 Cells

Besides a marked difference in the proportion of TE families affected by AcMNPV infection in Tnms42 cells (29.4% of all TE families) and *T. ni* midgut (10.1%), the response to AcMNPV also differs in terms of the proportion of up- versus downregulated TE families between the two data sets. While there are in total three times more upregulated than downregulated TEs in the midgut (fig. 4 and table 1), there are about as many up- as downregulated TE families inTnms42 cells (fig. 2 and table 1). The impact of AcMNPV infection on

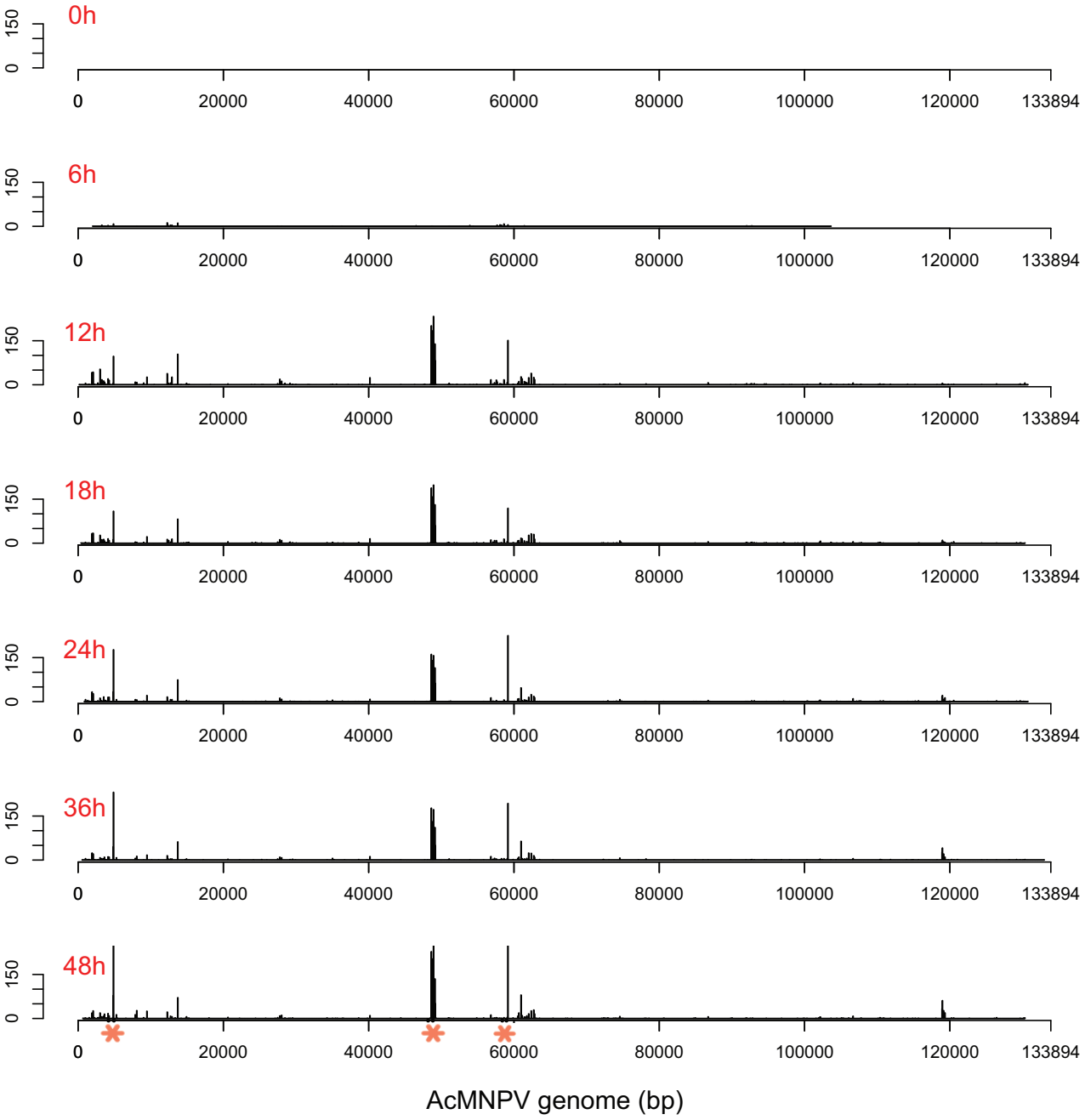


Fig. 5.—Distribution of expressed TE insertions along the AcMNPV genome for each time point in the Chen et al (2013) data sets. The *Y*-axis corresponds to the number of reads. Insertions are binned into 50-bp windows. The three major insertion hotspots, shown by orange asterisks on 48-hpi graph, correspond from left to right to the *PH*, *FP*, and *Orf78* genes.

TE expression tends to be unidirectional in the midgut, with a relatively steady increase in both the number of DE TE families and strength of differential expression with time (fig. 4). By contrast, in Tnms42 cells, the pattern of differential expression is more erratic, with many TE families that are DE at early time points postinfection not remaining DE in the next time points, and extrema of differential expression at any time point depending on TE families (fig. 2 and supplementary table S1). Altogether, these findings are rather consistent with what is known about cell lines, which generally grow under fewer constraints than tissues and often undergo important chromatin remodeling and chromosomal rearrangements, as described for the *T. ni* Hi5 cell line (Fu et al. 2018). Such modifications could lead to higher TE activity in cell lines

during stress conditions, as TE-restriction pathways may be 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 al. 1986; Granados et al. 1994). The larger proportion of upregulated TE families in the cell line versus midgut (91 vs. 46) is also in accordance with Macchietto et al. (2020) who found a global upregulation of TE expression after viral infections of various cell lines. However, while Macchietto et al. (2020) observed an early wave of TE upregulation, here, we find that TE upregulation occurs mainly at later time points in Tnms42 cells.

When looking at the impact of AcMNPV infection on TE expression by TE types, there are clear differences between Tnms42 cells and *T. ni* midgut tissues. These differences are perhaps best illustrated by the fact that only 26 shared TE families were DE in both data sets (supplementary table S1) and that for a given TE family, the direction of differential expression was not necessarily the same in Tnms42 cells and *T. ni* midgut. In fact, only 14 of these shared DE TEs were DE in the same direction (10 upregulation and 4 downregulation) in both data sets. Furthermore, the TE families with the highest or lowest log2FCs in one data set were not DE in the other data set. More globally, the differential impact of AcMNPV infection in the two data sets is also illustrated by the fact that the proportion of TE families that are DE among each TE type significantly differs between Tnms42 cells and *T. ni* midgut ($P < 0.01$, chi2 test). For example, the proportion of DE DNA transposons is twice higher in *T. ni* midgut than in Tnms42 cells (9% of DNA TEs in midgut and 25% of DNA TEs in cells). Furthermore, while a relatively similar proportion of LINE and LTR families is DE in the cell line (35% of LINEs and 27% of LTR), there are 3.5 times more LTR than LINE families that are DE in the midgut (18% of LTR vs. 5% of LINE). These differences of regulation in both data sets are possibly due to differences in the regulatory landscapes of the two cell types. Another explanation can be the expression of copies specific to each genome (i.e., present in one but not in the other genome) located in different genomic regions. The low number of DE TEs shared between both data sets is not unexpected given what is known about the impact of stress on TE expression in eukaryotes, as no unique, clear trend emerges (Horváth et al. 2017). 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 this respect, it is noteworthy that *T. ni* Hi5 cells (from which the Tnms42 cell line derives) differ widely from ovary (the tissue from which the cell line is derived) and other *T. ni* tissues in their piRNA response. For example, only 71 piRNA clusters were annotated in Hi5 cells compared to 348 in *T. ni* ovaries and many piRNA clusters that are active in ovaries only produce few piRNA in Hi5 cells (Fu et al. 2018). In addition to the piRNA pathway that actively represses TEs in lepidopterans (Lewis et al. 2018), epigenetic marks, such as 5-methylcytosine, 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.

Transposable Elements Previously Found Integrated into AcMNPV Genomes Show No Specific Expression Pattern

Interestingly, the *T. ni* TE families previously found to be inserted in AcMNPV genomes (Gilbert et al. 2014, 2016) are not more represented in DE TEs, and even the DE ones are as much down- as upregulated. This suggests 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. Looking at the absolute TE expression, we observed that even after upregulation, most DE TE families are overall not the most expressed TEs (fig. 3 and supplementary figs. S1 and S2). Thus, it might be possible that DE TEs were not found inserted in AcMNPV genome because their upregulation was not strong enough to reach sufficient expression. Moreover, some TEs found inserted in AcMNPV were weakly expressed in our study, suggesting that a weak expression might be enough for insertion in AcMNPV. Alternatively, transposition into viral genomes may occur in tissues other than those constituting the midgut or this cell line, in which TE expression might be different. In this regard, it is noteworthy that the tissue tropism of AcMNPV includes most cell types of lepidopteran larvae (Engelhard et al. 1994; Barrett et al. 1998; Rahman et al. 2004). It would thus be interesting to repeat this analysis on several other tissues and/or on whole larvae.

Transposable Elements Integrated into Viral Genomes Can Be Expressed

Our results show that at least 11 TE families from a *T. ni* cell line can be inserted in and transcribed from AcMNPV genomes. Our approach only allows us to detect TE copies that are cotranscribed with the upstream or downstream viral gene. Yet we predict that viral-borne TFP3 copies may be expressed from their own promoter, as piggybac elements are known to carry such a promoter, located in their 5′ repeated sequence (Cadiñanos and Bradley 2007). Chimeric reads are not expected if AcMNPV-borne TEs are transcribed from their own promoters. Viral-borne TFP3 copies are identical to TFP3 copies located in the genome of *T. ni* and it is thus not possible to assess which proportion of the RNA-seq reads mapping to the internal part of the element correspond to viral-borne or host-borne TFP3 copies. For the same reason, it was not possible to assess whether some TE transcripts were coencapsidated into virions but not inserted into the viral genome in our data sets, as found in several RNA viruses (e.g., Routh et al. 2012). Thus, although the expression level of virus-borne TFP3 copies is here equivalent to that of some AcMNPV genes, the expression of these virus-borne TFP3 copies is likely underestimated.

The three genes (*FP*, *PH*, and *Ac-Orf78*) bearing highly expressed TFP3 copies are known to be involved in the formation of occlusion bodies (OBs). Inactivation of *FP* or *PH* leads to a drop of AcMNPV OB formation (Hink and Vail 1973; Fraser et al. 1983) 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 likely explains 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). It is thus likely that 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 is 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, as shown by the presence of many TE-virus transcripts (chimeric RNAseq reads), increasing our ability to detect TE-virus chimeras in this data set. Importantly, the longer read length (101 bp) produced by Chen et al. (2014) also probably contributed to more efficiently detect TE-virus chimeras than in the Shrestha et al. (2018) data set (read length 51 bp).

In conclusion, we found that TEs integrated into AcMNPV genomes can be expressed at substantial levels, a prerequisite for such TEs to be able to further transpose from viral genomes to other viral genomes or to the genome of another host. Thus, our results further contribute to support viruses as potential vectors of TEs between animals. Importantly, they also suggest 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.

Materials and 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 Hi5 cells, were infected with the wild-type AcMNPV strain E2 (Chen et al. 2013). For infections, 3×10^6 Tnms42 cells were infected at a multiplicity of infection (MOI) of 10. After a 1-h incubation, the inoculum was removed and the cells were rinsed and further cultured with new medium. 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. Polyadenylated RNA isolated from 20 μg total RNA was used for sequencing. The sequencing library was constructed with the TruSeq protocol and sequenced on an Illumina platform. Single-end reads of 101 bp were produced for the infected condition and for the 0-hpi mock condition, whereas 51-bp reads were produced for the mock condition of the other time points and for one replicate at each time point of the infected condition. To avoid any bias potentially introduced by different read lengths, we only used replicates produced from 101-bp reads. Further information can be found in Chen et al. (2013). Please note that reads corresponding to the mock condition at 24 hpi cannot be retrieved from the SRA.

RNA-Seq Data of Midgut *T. ni* Larvae Infected by AcMNPV

RNA-seq data were 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 0–5 h without diet, and newly molted 5th instar larvae were used for oral infections. Larvae were orally inoculated with 5 μl of a 10% sucrose solution containing a total of 7×10^4 OBs 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. 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. Poly(A) mRNAs isolated from 3 μg of total RNA were used to construct a library with the TruSeq protocol and sequenced on an Illumina 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: 1) one derived from a single male *T. ni* larva (accession number PPHH01000000; Chen et al. 2019) and 2) one derived from the *T. ni* Hi5 germ cell line (accession number NKQN00000000; Fu et al. 2018). Importantly, the larvae used to sequence the genome in Chen et al. (2019) and to produce the midgut RNA-seq reads in Shrestha et al. (2018) arise from the same strain (Cornell strain). Midgut RNA-seq reads were thus mapped onto TE copies retrieved from the genome assembled by Chen et al. (2019). Similarly, the Tnms42 cell line, used to produce RNA-seq reads in Chen et al. (2013) is an alphanodavirus-free derivative from the Hi5 cell line, for which a genome is available (Chen et al. 2019). The cell line RNA-seq reads were thus mapped onto TE copies retrieved from the Hi5 cell line genome.

TE Identification and Database

The TE library used to annotate TE copies 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 TEs previously found inserted in viral genomes (Wang and Fraser 1993; Fraser et al. 1995; Gilbert et al. 2016). Finally, we annotated TEs in the RNA-seq data. The 48 data sets produced by Shrestha et al. (2018) were pooled and 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 nonviral 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 data sets from Chen et al. (2013). After the Trinity assembly, we found 103,650 nonviral contigs 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 in the genome and transcriptome assemblies. 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 847 TE families is provided in supplementary file S1 and was used to annotate TE copies in the two *T. ni* genomes using RepeatMasker version 4.0.7 (http://www.repeatmasker.org). We then grouped RepeatMasker hits into more complete TE copies with the tool "One code to find them all" with the option –strict and –unknown (keeps only the copies greater than 80 bp in length and with more than 80% identity with the consensus) (Bailly-Bechet et al. 2014).

Mapping of TE Copies

The RNA-seq data were trimmed using Trimmomatic version 0.38 (Bolger et al. 2014) to remove adapters 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). The trimmed RNA-seq data were mapped to the TE copies of their corresponding *T. ni* genomes with 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). The minimum criteria for a read to align on a TE copy with the "very-sensitive" option is at least an alignment of 20-bp substring without any mismatch, with a 6-bp interval. It corresponds to 6 and 14 20-bp substrings for a read of 51 or 101 bp, respectively.

Count Tables of TEs and Genes

We produced count tables for the two time course RNA-seq data sets (Chen et al. 2013; Shrestha et al. 2018), independently for each data set and time point. Gene count tables were generated with Kallisto (Bray et al. 2016) for normalization purpose. The entry files provided to Kallisto were fastq files containing trimmed RNA-seq reads and a file containing the host transcriptome (tni_transcript_v1.fa at tnibase.org for the midgut data set and GBKU01.1.fsa_nt on NCBI for the Tnms42 cells data set) and the 156 CDS of AcMNPV reference genome (NC_001623). TE count tables were generated with the module TEcount of TEtools version 1.0.0 (Lerat et al. 2016). Entry files provided to TEtools were the sam files containing mapping information on the TE copies for all replicates of each time point, both for mock and infected conditions, the fasta file of the TE copies and a rosette file giving the name of the TE family for each TE copy. Gene counts and TE counts were then concatenated.

Differential Expression Analysis

Differential expression analysis was computed on the concatenated count tables with the R Bioconductor package DESeq2 (Love 2014 Genome Biology), using an FDR level of 0.05 (Benjamini and Hochberg 1995). For this, the DESeqDataSet object was built with DESeqDataSetFromTximport for the gene counts and with DESeqDataSetFromMatrix for the TEs counts, both with the design ~ condition. DESeq was then run on the concatenated DESeqDataSet object, and the results were generated with the contrast c("condition," "infected," "mock"). For the midgut data set, we followed the procedure used by Shrestha et al. (2018) to study *T. ni* gene expression and compared normalized read counts between infected larvae and mocks for each time point post-infection. For the cell line data set, we also followed Chen et al. (2013). They reasoned that contrary to AcMNPV-infected *T. ni* cells, which stop dividing, uninfected cultured cells undergo important stresses as they grow due to space constraints, which may induce variation in gene expression in the mock condition after some time. For this reason, Chen et al. (2013) calculated differential expression of Tnms42

genes by comparing normalized read counts for each time point postinfection in the infected condition to read counts obtained in the mock at the first time point postinfection [0 hpi, corresponding to an hour of incubation by the virus, as mentioned in Chen et al. (2013)]. We were interested only in DE TE families; thus, we discarded the differential expressed genes that we initially included only for normalization purpose. TE families were considered as differentially expressed if their adjusted *P*-value was <0.05 and their absolute log2 fold change was ≥ 2. All analyses were performed using R version 4.0.4 (R Core Team 2019, https://www.R-project.org/).

TPM Computation

Based on the concatenated gene and TE count tables, TE family counts were normalized to TPM. For each data set, we first calculated the number of reads per kilobase (RPK) by dividing the count of each gene or TE family in each replicate by the length of the gene or the length of the corresponding TE consensus in kilobases. RPK of all genes and TE families were then summed up by replicate and divided by one million. We finally used this million-factor for each replicate to divide all RPK values previously calculated. Since Wagner et al. (2013) suggested to use either 2 or 4 TPM as cutoff for nonexpressed genes, we chose 4 TPM as a cutoff. The cutoff for a highly expressed TE family being quite arbitrary, we chose 50 TPM because it corresponds to about 100 RPKM in our data set, which was used as a cutoff for highly expressed genes in Chen et al. (2013). For visual purpose, we chose to plot the TPM with a square-root transformation. Indeed, as explained by Wagner et al. (2013) and also as observed with our data, the standard log transformation leads to an over-dispersion at low TPMs, which might let one think that TE families show a relatively large and continuous range of expression levels, whereas most of them are actually lowly expressed.

Correlations with TE Expression

We investigated possible correlations between the expression level of TE families in *T. ni* larvae mock condition and the following factors: TE copy number, TE age, and TE proximity with genes. For TE copy number, we counted the number of copies included in our study for each TE family (the copies which passed the "One code to find them all" filters). TE age was estimated by the percentage of divergence to the consensus for each copy. We used the average divergence of copies to estimate the age of a TE family. About TE proximity with genes, we indicated a distance of 0 for TE copies inside genes, otherwise, we counted the distance to the closest gene, in base pairs (genome annotation tni_gene_v1.gff3 at tnibase.org). The closest gene could be downstream or upstream; in any case, we calculated only positive values. Then, we calculated the average distance to nearest gene for each family. We investigated correlations between factors with the R package "corrplot," using the Pearson method.

Detection of TE/Virus Junctions in Transcriptomic Data

In addition to the DE analysis, we measured the expression of host 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 in a viral gene and continue in a TE sequence integrated in the viral genome. This approach allowed us to ensure that only TE-containing transcripts initiating in the viral (not host) genome were included. To identify chimeric reads, all reads were aligned to the AcMNPV WP10 genome (GenBank accession number KM609482) and to a library of TEs including all TEs annotated in this study (supplementary file S1 and see above) and many other TEs found in various databases. Analyses to identify chimeric reads were performed on R (R Core team 2019). The pipeline we used 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. All TEs found integrated into and expressed from AcMNPV genomes are provided in supplementary file S2.

Identification of Target Site Duplications

To confirm host TE insertions in viral genomes, we searched for TSD that are signatures 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.

Data Availability

The data used in this article are available in the GenBank Sequence Read Archive [SRA] database under accession numbers SRA057390 (Chen et al. 2013) and PRJNA484772 (Shrestha et al. 2018). Supplementary file S1 contains the consensus sequence of the 847 TE families used to annotate TE copies in *T. ni* genomes. Supplementary file S2 contains all TEs found integrated into and expressed from AcMNPV genomes.

Supplementary Material

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

Acknowledgments

C.G. acknowledges funding from ANR grants ANR-15-CE32-0011-01 TransVir and ANR-18-CE02-0021-01TranspHorizon.

Literature Cited

Arkhipova IR. 2018. Neutral theory, transposable elements, and eukaryotic genome evolution. Mol Biol Evol. 35(6):1332–1337.
Barah P, Jayavelu ND, Mundy J, Bones AM. 2013. Genome scale transcriptional response diversity among ten ecotypes of Arabidopsis thaliana during heat stress. Front Plant Sci. 4:532.
Barrett JW, Brownwright AJ, Primavera MJ, Palli SR. 1998. Studies of the nucleopolyhedrovirus infection process in insects by using the green fluorescence protein as a reporter. J Virol. 72 (4):3377–3382.
Barrón MG, Fiston-Lavier A-S, Petrov DA, González J. 2014. Population genomics of transposable elements in *Drosophila*. Annu Rev Genet. 48 (1):561–581.
Bailly-Bechet M, Haudry A, Lerat E. 2014. "One code to find them all": a perl tool to conveniently parse RepeatMasker output files. Mobile DNA. 5(1):13.p
Bauser CA, Elick TA, Fraser MJ. 1996. Characterization of hitchhiker, a transposon insertion frequently associated with baculovirus FP mutants derived upon passage in the TN-368 cell line. Virology. 216 (1):235–237.
Benjamini Y, Hochberg Y. 1995. Controlling the false discoveryrate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B. 85:289–300.
Blumenstiel JP. 2019. Birth, school, work, death, and resurrection: the life stages and dynamics of transposable element proliferation. Genes. 10 (5):336.
Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for illumina sequence data. Bioinformatics. 30 (15):2114–2120.
Bouallègue M, Rouault J-D, Hua-Van A, Makni M, Capy P. 2017. Molecular evolution of *PiggyBac* superfamily: from selfishness to domestication. Genome Biol Evol. 9 (2):323–339.
Bourque G, et al. 2018. Ten things you should know about transposable elements. Genome Biol. 19 (1):199.
Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near optimal probabilistic RNA-seq quantification. Nat Biotechnol. 34(5):525–527.
Brookfield JFY, Badge RM. 1997. Population genetics models of transposable elements. In: Capy P, editor. Evolution and impact of transposable elements. Vol. 6. Dordrecht: Springer Netherlands. p. 281–294. https://doi.org/10.1007/978-94-011-4898-6_28.
Bucher E, Reinders J, Mirouze M. 2012. Epigenetic control of transposon transcription and mobility in Arabidopsis. Curr Opin Plant Biol. 15 (5):503–510.
Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 12 (1):59–60.
Cadiñanos J, Bradley A. 2007. Generation of an inducible and optimized PiggyBac transposon system. Nucleic Acids Res. 35 (12):e87.
Capy P, Gasperi G, Biémont C, Bazin C. 2000. Stress and transposable elements: co-evolution or useful parasites?. Heredity. 85 (2):101–106.
Casacuberta E, González J. 2013. The impact of Transposabletni_gene_v1_noDetails.gff3 elements in environmental adaptation. Mol Ecol. 22 (6):1503–1517.
Chen W, et al. 2019. A high-quality chromosome-level genome assembly of a generalist herbivore, *Trichoplusia ni*. Mol Ecol Resour. 19 (2):485–496.
Chen Y-R, et al. 2013. The transcriptome of the baculovirus Autographa californica multiple nucleopolyhedrovirus in Trichoplusia ni cells. J Virol. 87 (11):6391–6405.
Chen Y-R, et al. 2014. Transcriptome responses of the host Trichoplusia ni to infection by the baculovirus Autographa californica multiple nucleopolyhedrovirus. J Virol. 88 (23):13781–13797.
Chuong EB, Elde NC, Feschotte C. 2017. Regulatory activities of transposable elements: from conflicts to benefits. Nat Rev Genet. 18 (2):71–86.
Cowley M, Oakey RJ. 2013. Transposable elements re-wire and fine-tune the transcriptome. PLoS Genet. 9(1):e1003234.Edited by Elizabeth M. C. Fisher.
Deniz Ö, Frost JM, Branco MR. 2019. Regulation of transposable elements by DNA modifications. Nat Rev Genet. 20:417–431.
Dubin MJ, Mittelsten Scheid O, Becker C. 2018. Transposons: a blessing curse. Curr Opin Plant Biol. 42 (April):23–29.
Engelhard EK, Kam-Morgan LN, Washburn JO, Volkman LE. 1994. The insect tracheal system: a conduit for the systemic spread of Autographa californica M nuclear polyhedrosis virus. Proc Natl Acad Sci U S A. 91 (8):3224–3227.
Fouché S, et al. 2020. Stress-driven transposable element de-repression dynamics and virulence evolution in a fungal pathogen. Mol Biol Evol. 37(1):221–239.
Fraser MJ, Brusca JS, Smith GE, Summers MD. 1985. Transposon-mediated mutagenesis of a baculovirus. Virology. 145 (2):356–361.
Fraser MJ, Cary L, Boonvisudhi K, Wang HG. 1995. Assay for movement of lepidopteran transposon IFP2 in insect cells using a baculovirus genome as a target DNA. Virology. 211(2):397–407.
Fraser MJ, Clszczon T, Elick T, Bauser C. 1996. Precise excision of TTAA-specific lepidopteran transposons *piggyBac* (IFP2) and *tagalong* (TFP3) from the baculovirus genome in cell lines from two species of Lepidoptera. Insect Mol Biol. 5(2):141–151.
Fraser MJ, Smith, Gale E and Summers Max D. 1983. Acquisition of host cell DNA sequences by baculoviruses: relationship between host DNA insertions and FP mutants of Autographa californica and Galleria mellonella nuclear polyhedrosis viruses. J Virol. 47 (2):287–300.
Fu Y, et al. 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.
Galindo-González L, Mhiri C, Deyholos MK, Grandbastien M-A. 2017. LTR-retrotransposons in plants: engines of evolution. Gene. 626 (August):14–25.
Gilbert C, et al. 2014. Population genomics supports baculoviruses as vectors of horizontal transfer of insect transposons. Nat Commun. 5 (1):3348.
Gilbert C, et al. 2016. Continuous influx of genetic material from host to virus populations. PLoS Genet. 12(2):e1005838.
Grabherr MG, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 29 (7):644–652.
Granados RR, Derksen ACG, Dwyer KG. 1986. Replication of the Trichoplusia ni granulosis and nuclear polyhedrosis viruses in cell cultures. Virology. 152 (2):472–476.
Granados RR, Guoxun L, Derksen ACG, McKenna KA. 1994. A new insect cell line from Trichoplusia ni (BTI-Tn-5B1-4) susceptible to Trichoplusia ni single enveloped nuclear polyhedrosis virus. J Invertebr Pathol. 64 (3):260–266.
Hink WF, Vail PV. 1973. A plaque assay for titration of alfalfa looper nuclear polyhedrosis virus in a cabbage looper (TN-368) cell line. J Invertebr Pathol. 22(2):168–174.
Horváth V, Merenciano M, González J. 2017. Revisiting the relationship between transposable elements and the eukaryotic stress response. Trends Genet. 33 (11):832–841.
Huang J, et al. 2017. EARE-1, a transcriptionally active Ty1/Copia-like retrotransposon has colonized the genome of Excoecaria agallocha through horizontal transfer. Front Plant Sci. 8 (January):45.

Hummel B, et al. 2017. The evolutionary capacitor HSP90 buffers the regulatory effects of mammalian endogenous retroviruses. Nat Struct Mol Biol. 24 (3):234–242.
Jehle JA, Nickel A, Vlak JM, Backhaus H. 1998. Horizontal escape of the novel Tc1-like lepidopteran transposon TCp3.2 into Cydia pomonella granulovirus. J Mol Evol. 46 (2):215–224.
Lanciano S, Mirouze M. 2018. Transposable elements: all mobile, all different, some stress responsive, some adaptive? Curr Opin Genet Dev. 49 (April):106–114.
Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods. 9 (4):357–359.
Le Rouzic A, Boutin TS, Capy P. 2007. Long-term evolution of transposable elements. Proc Natl Acad Sci U S A. 104 (49):19375–19380.
Lerat E, Fablet M, Modolo L, Lopez-Maestre H, Vieira C. 2016. TEtools facilitates big data expression analysis of transposable elements and reveals an antagonism between their activity and that of PiRNA genes. Nucleic Acids Res. 45(4):e17.
Lewis SH, et al. 2018. Pan-arthropod analysis reveals somatic piRNAs as an ancestral defence against transposable elements. Nat Ecol Evol. 2(1):174–181.
Loiseau V, et al. 2020. Wide spectrum and high frequency of genomic structural variation, including transposable elements, in large double-stranded DNA viruses. Virus Evol. 6 (1):vez060.
Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. Genome Biol. 15 (12):550.
Macchietto MG, Langlois RA, Shen SS. 2020. Virus-induced transposable element expression upregulation in human and mouse host cells. Life Sci Alliance. 3 (2):e201900536.
Menees TM, Sandmeyer SB. 1996. Cellular stress inhibits transposition of the yeast retrovirus-like element Ty3 by a ubiquitin-dependent block of virus-like particle formation. Proc Natl Acad Sci U S A. 93 (11):5629–5634.
Miousse IR, et al. 2015. Response of transposable elements to environmental stressors. Mutat Res Rev Mutat Res. 765 (July):19–39.
Mita P, Boeke JD. 2016. How retrotransposons shape genome regulation. Curr Opin Genet Dev. 37 (April):90–100.
Niederhuth CE, et al. 2016. Widespread natural variation of DNA methylation within angiosperms. Genome Biol. 17 (1):194.
R Core Team. 2019. R: a language and environment for statistical computing. Vienna, Austria: R Foundation or Statistical Computing. Available from: https://www.R-project.org/.
Rahman M, Masmudur KP. Gopinathan 2004. Systemic and in vitro infection process of Bombyx mori nucleopolyhedrovirus. Virus Res. 101 (2):109–118.
Rognes T, Flouri T, Nichols B, Quince C, Mahé F. 2016. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 4 (October):e2584.
Romero-Soriano V, Garcia Guerreiro MP. 2016. Expression of the retrotransposon Helena reveals a complex pattern of TE deregulation in Drosophila hybrids. PLoS One. 11(1):e0147903.
Routh A, Domitrovic T, Johnson JE. 2012. Host RNAs, including transposons, are encapsidated by a eukaryotic single-stranded RNA virus. Proc Natl Acad Sci U S A. 109 (6):1907–1912.
Roy M, et al. 2020. Viral infection impacts transposable element transcript amounts in *Drosophila*. Proc Natl Acad Sci U S A. 117 (22):12249–12257.
Ryan CP, Brownlie JC, Whyard S. 2017. Hsp90 and physiological stress are linked to autonomous transposon mobility and heritable genetic change in nematodes. Genome Biol Evol. 8 (12):3794–3805.
Saksouk N, Simboeck E, Déjardin J. 2015. Constitutive heterochromatin formation and transcription in mammals. Epigenetics Chromatin. 8 (1):3.
Schnable PS, et al. 2009. The B73 maize genome: complexity, diversity, and dynamics. Science. 326 (5956):1112–1115.
Shrestha A, et al. 2018. Global analysis of baculovirus Autographa californica multiple nucleopolyhedrovirus gene expression in the midgut of the lepidopteran host *Trichoplusia ni*. J Virol. 92 (23):e01277–18.
Shrestha A, et al. 2019. Transcriptional responses of the *Trichoplusia ni* midgut to oral infection by the baculovirus Autographa californica multiple nucleopolyhedrovirus. J Virol. 93 (14):e00353–19.
Sinzelle L, et al. 2008. Transposition of a reconstructed harbinger element in human cells and functional homology with two transposon-derived cellular genes. Proc Natl Acad Sci U S A. 105 (12):4715–4720.
Slotkin R, Keith R. Martienssen 2007. Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet. 8 (4):272–285.
Sotero-Caio CG, Platt RN, Suh A, Ray DA. 2017. Evolution and diversity of transposable elements in vertebrate genomes. Genome Biol Evol. 9 (1):161–177.
Tao XY, et al. 2013. The Autographa californica multiple nucleopolyhedrovirus ORF78 is essential for budded virus production and general occlusion body formation. J Virol. 87 (15):8441–8450.
Trojer P, Reinberg D. 2007. Facultative heterochromatin: is there a distinctive molecular signature?. Mol Cell. 28 (1):1–13.
Van Meter M, et al. 2014. SIRT6 represses LINE1 retrotransposons by ribosylating KAP1 but this repression fails with stress and age. Nat Commun. 5 (1):5011.
Volff J-N. 2006. Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. Bioessays. 28 (9):913–922.
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(4):937–951.
Wagner GP, Kin K, Lynch VJ. 2013. A model based criterion for gene expression calls using RNA-seq data. Theory Biosci. 132(3):159–164.
Wang HG, Fraser MJ. 1993. TTAA serves as the target site for TFP3 lepidopteran transposon insertions in both nuclear polyhedrosis virus and *Trichoplusia Ni* genomes. Insect Mol Biol. 1 (3):109–116.
Wicker T, et al. 2007. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 8 (12):973–982.
Wood HA. 1980. Isolation and replication of an occlusion body-deficient mutant of the Autographa californica nuclear polyhedrosis virus. Virology. 105 (2):338–344.
Yoder JA, Walsh CP, Bestor TH. 1997. Cytosine methylation and the ecology of intragenomic parasites. Trends Genet. 13(8):335–340.
Zilberman D, Gehring M, Tran RK, Ballinger T, Henikoff S. 2007. Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. Nat Genet. 39 (1):61–69.
Zovoilis A, Cifuentes-Rojas C, Chu H-P, Hernandez AJ, Lee JT. 2016. Destabilization of B2 RNA by EZH2 activates the stress response. Cell. 167(7):1788–1802.e13.

Associate editor: Esther Betran

Part II

Domesticated viruses as vectors of horizontal transfers from parasitoid wasps

8 - Article n°2: Draft nuclear genome and complete mitogenome of the Mediterranean corn borer, *Sesamia nonagrioides*, a major pest of maize

In order to investigate HT from the parasitoid wasp *Cotesia typhae* to its natural host *Sesamia nonagrioides*, we needed to assemble and annotate both reference genomes. The former was included in the article n°3, whereas the latter was the object of a dedicated article, in which we reported an in-depth annotation of genes involved in sex-determination and amylase production. DNA extraction, DNA sequencing, the first version of the nuclear genome assembly, and the annotation of amylase genes were done before the beginning of my PhD by co-authors.

I began this work during my M2 internship, during which I improved the nuclear assembly by purging haplotigs and heterozygous overlaps and by removing some contamination. I also used an automatic pipeline to annotate the genome, and annotated by hand the sex-determination genes which could be useful in the context of biocontrol. I also assembled a complete mitochondrial genome that I annotated. I also found a cluster of nuclear mitochondrial DNA (NUMTs) that results from the recent nuclear integration of two copies of the mitochondrial genome, one of which is rearranged in three pieces.

Finally, as a preliminary analysis in the context of HT from domesticated viruses of *C. typhae* to *S. nonagrioides*, I used the bracovirus sequences of *C. sesamiae* (CsBV), which was the closest available genome, to look for bracoviral integration. However, I did not find any sign of recent HT from CsBV to *S. nonagrioides*. Nonetheless, I found an helitron (a TE) that was horizontally transferred between the wasp and the moth. Whether this transfer was facilitated by the integration of wasp DNA circles in germline genomes of lepidopteran larvae during parasitism is an interesting possibility that deserves further investigation.

INVESTIGATIONS

Draft nuclear genome and complete mitogenome of the Mediterranean corn borer, *Sesamia nonagrioides*, a major pest of maize

Héloïse Muller*†, David Ogereau*, Jean-Luc Da-Lage*, Claire Capdevielle*, Nicolas Pollet*, Taiadjana Fortuna*, Rémi Jeannette*, Laure Kaiser* and Clément Gilbert*[1]

*Université Paris-Saclay, CNRS, IRD, UMR Évolution, Génomes, Comportement et Écologie, 91198, Gif-sur-Yvette, France, †Master de Biologie, École Normale Supérieure de Lyon, Université Claude Bernard Lyon I, Université de Lyon, 69342 Lyon Cedex 07, France

ABSTRACT The Mediterranean corn borer (*Sesamia nonagrioides*, Noctuidae, Lepidoptera) is a major pest of maize in Europe and Africa. Here, we report an assembly of the nuclear and mitochondrial genome of a pool of inbred males and females third instar larvae, based on short- and long-read sequencing. The complete mitochondrial genome is 15,330 bp and contains all expected 13 and 24 protein-coding and RNA genes, respectively. The nuclear assembly is 1,021 Mbp, composed of 2,553 scaffolds and it has an N50 of 1,105 kbp. It is more than twice larger than that of all Noctuidae species sequenced to date, mainly due to a higher repeat content. A total of 17,230 protein-coding genes were predicted, including 15,776 with InterPro domains. We provide detailed annotation of genes involved in sex determination (*dsx, IMP, PSI*) and of alpha-amylase genes possibly involved in interaction with parasitoid wasps. We found no evidence of recent horizontal transfer of bracovirus genes from parasitoid wasps. These genome assemblies provide a solid molecular basis to study insect genome evolution and to further develop biocontrol strategies against *S. nonagrioides*

KEYWORDS

Genome Assembly
Lepidoptera
Crop pest
Sex determination
Alpha-amylase
Bracoviruses

INTRODUCTION

The Mediterranean corn borer (*Sesamia nonagrioides*, Noctuidae) is a major pest of maize in Mediterranean regions and in Sub-Saharan Africa (Bosque-Perez *et al.* 1998; Moyal *et al.* 2011; Kergoat *et al.* 2015; Kankonda *et al.* 2018). The damage it causes to maize is due to the moth's larval feeding behaviour, which involves digging tunnels in the stem of the plants. Strategies to control *S. nonagrioides* mainly rely on chemical pesticides and transgenic plants such as Bt maize that expresses insecticidal proteins (Farinós *et al.* 2018). However, as observed in other species, an allele conferring resistance to Bt-toxin has been recently identified in *S. nonagrioides* (Camargo *et al.* 2018). Furthermore, most EU countries take positions against genetically modified crops (Farinós *et al.* 2018). Alternative methods implementing various biological agents such as viruses, pheromones, sterile insects or RNA interference have been developed to control other pests (Beevor *et al.* 1990; Moscardi 1999; Cork *et al.* 2003; Tian *et al.* 13 juil. 2009; Jin *et al.* 2013; Alamalakala *et al.* 2018). In addition, several biological control programs targeting lepidopteran stemborers rely on the use of parasitoid wasps belonging to the genus *Cotesia* (Kfir *et al.* 2002; Muirhead *et al.* 2012; Midingoyi *et al.* 2016). One species of *Cotesia*, *C. typhae*, belonging to the *C. flavipes* species complex, has recently been described as parasitizing exclusively *S. nonagrioides*. The potential of *C. typhae* as a biological control agent against this pest is being currently studied (Kaiser *et al.* 2017). In this context, and because knowing the genetics and genomics of pest species is essential to develop biocontrol programs (Leung *et al.* 2020), we assembled the nuclear and mitochondrial genomes of *S. nonagrioides* using short and long sequencing reads. We provide detailed annotations of genes encoding *alpha-amylases*, which are likely involved in host recognition, and of genes involved in sex determination, which may be useful in a strategy relying on the release of sterile males. We also report the results of a search for polydnaviral genes that would have been horizontally transferred from *Cotesia* wasps to *S. nonagrioides*.

Manuscript compiled: Thursday 29th April, 2021
[1]Corresponding author: Université Paris-Saclay, CNRS, IRD, UMR Évolution, Génomes, Comportement et Écologie, 91198, Gif-sur-Yvette, France. E-mail: clement.gilbert@egce.cnrs-gif.fr

MATERIALS AND METHODS

DNA extraction

We extracted large amounts of high quality DNA from whole bodies of 10 third instar larvae of *S. nonagrioides*, males and females, sampled in our laboratory population. We initiated this population in 2010 with individuals sampled in several localities of the French region Haute Garonne (Longages N43.37; E1.19 and vicinity). Since then, we mixed the population at least every two years with individuals collected in several localities and regions of south-west France (Pyrénées Atlantiques, Haute Garonne, Tarn et Garonne, Lot et Garonne, Landes, Gironde). An analysis of *S. nonagrioides* population genetics in France revealed weak genetic differentiation over France (Naino-Jika *et al.* 2020). The laboratory population is reared on a diet adapted from Overholt *et al.* (1994). Mating and oviposition occur in a cage where we introduce 30 pupae of each sex weekly. The pupae can be sexed by comparing their abdominal characters (Giacometti 1995). The 10 larvae used to extract DNA result from two successive crossings of siblings that we implemented to further reduce heterozygosity. We ground the pool of 10 larvae in liquid nitrogen, amounting to 100 mg of fine dry powder. We then extracted DNA using Nucleobond AXG100 columns and the Buffer Set IV from Macherey Nagel, following the manufacturer's protocol. We obtained 60 µg of DNA, quantified with QuBit (ThermoFisher Scientific). We checked the integrity of DNA on an agarose gel (Figure S1) and we did a spectrophotometer measure (Nanodrop 2000) to check the absence of proteins and other contaminants.

Sequencing and genome assembly

We sub-contracted Genotoul (genotoul.fr) to build a paired end library (2x150 pb; insert size = 350 bp) for sequencing on an Illumina platform. We performed long-read sequencing using the Oxford Nanopore Technology (ONT) in our lab on six flowcells (R9.4). Sequencing was performed over the course when ONT upgraded ligation kits. Thus, while our three first libraries were prepared with the SQK-LSK108 kit, the three last were prepared with the SQK-LSK109 kit, including one with an additional Bluepippin size selection step (15 kb cut-off). We assembled the genome with the MaSuRCA hybrid assembler v3.3.1 (Zimin *et al.* 2017). We set all parameters to default, except those related to the location of the data, number of threads (64) and Jellyfish hash size (JF_SIZE = 12000000000). We used all 278,683,802 untrimmed Illumina reads (41,8 Gb) produced by Genotoul, as recommended by Zimin *et al.* (2017). We filtered Nanopore reads using Nanofilt (De Coster *et al.* 2018) to only keep reads longer than 7 kb (3,085,942 reads amounting to 45,6 Gb with an N50 of 17 kb). We then purged haplotigs and heterozygous overlaps from the assembly using the purge_dups pipeline described by Guan *et al.* (2020). We used all the default parameters, except for minimap2, for which we specified that we have ONT reads (xamp-ont), and for get_seqs, where we used the option -e to remove duplications at the ends of the contigs only. We checked for contamination in the assembly using blobtools v1.1 (Laetsch and Blaxter 2017), with default parameters. Blobtools requires three inputs: (i) the assembly, (ii) a hit file that we generated using our assembly as a query to perform a blastn search (-task megablast, -max_target_seqs 1, -max_hsps 1, -evalue 1e-25) against the NCBI database NT (downloaded in March 2019) and (iii) an indexed BAM file that we generated by mapping the trimmed Illumina reads (Trimmomatic v0.38 (Bolger *et al.* 2014)) against the assembly with Bowtie2 v2.3.4.1 (Langmead and Salzberg 2012). We also ran the module "all" of MitoZ v2.3 in order to assemble the mitogenome, annotate it and visualize it (Meng *et al.* 2019). We used the raw Illumina reads as input as recommended by Meng *et al.* (2019), we set all parameters to default and we set the genetic code and clade to invertebrate and Arthropoda respectively. Once assembled, we used the mitogenome as a query to perform a blastn search against the assembly to identify possible nuclear mitochondrial DNA (NUMTs). We validated the largest of these NUMTs by PCR, using primers covering three nuclear-mitochondrial junctions (junction 1 F: CAACACCGATGACATATTGGGT; junction 1 R: CGCACACATAAACAATAACGCC; junction 2 F: TGAGGGAGAAGGTAAGTCGA; junction 2 R: TGAGGAGGCGTATTGAGGTT; junction 4 F: GCGGCTCCTCCTAGATTAAATC; junction 4 R: ACTCTCCACGACCAAACCTC).

Genome size estimation

We estimated the genome size of *S. nonagrioides* using the R packages findGSE and GenomeScope that rely on k-mer frequencies (Vurture *et al.* 2017; Sun *et al.* 2018). We counted the number of k-mer on the Illumina reads using Jellyfish, with k equals 17, 21, 25 and 29 (Marçais and Kingsford 2011).

Genome annotation

We annotated genes and repeated elements of *S. nonagrioides* using Maker v2.31.10 (Holt and Yandell 2011; Campbell *et al.* 2014). First, we identified repeated elements *de novo* with RepeatModeler v2.0.1 (https://github.com/Dfam-consortium/RepeatModeler). We then ran a first round of Maker to (i) mask repeated elements and (ii) perform a preliminary gene annotation using the transcriptome of *S. nonagrioides* (Glaser *et al.* 2015) and the proteomes of three related species: *Busseola fusca* (Hardwick *et al.* 2019), *Spodoptera litura* (Zhu *et al.* 2018) and *Trichoplusia ni* (Chen *et al.* 2019). We merged the outputs of this first round into a GFF3 file, which we used to train SNAP, a gene predictor. We then ran a second round of Maker using this first GFF3 file and SNAP. We then trained Augustus, another gene predictor, with the second GFF3 file, generated by the second round of Maker. Finally, we ran a third and last round of Maker with the second GFF3 file and Augustus. This pipeline led to the final GFF3 file, containing the annotation of *S. nonagrioides*.

Functional annotation

We identified putative protein functions by blastp search (-evalue 1e-6 -max_hsps 1 -max_target_seqs 1) using the predicted proteins of *S. nonagrioides* against the non-redundant database UniProtKB/Swiss-Prot that contains unique proteins. In addition, we identified the GO terms and the conserved domains with InterProScan v5.46-81.0. To do this, we ran the 16 analyses proposed by InterProScan, including Pfam.

Comparison with other Noctuidae

We assessed the quality of our *S. nonagrioides* assembly by comparing its statistics to six other Noctuidae genomes for which all characteristics used in our comparison are available: *T. ni* (Talsania *et al.* 2019), *S. litura* (Cheng *et al.* 2017), *Spodoptera exigua* (Zhang *et al.* 2019), *Spodoptera frugiperda* (Kakumani *et al.* 2014), *Helicoverpa armigera* (Pearce *et al.* 2017) and *Helicoverpa zea* (Pearce *et al.* 2017).

RESULTS AND DISCUSSION

Nuclear genome assembly

The MaSuRCA assembler yielded a preliminary assembly of the *S. nonagrioides* genome composed of 4,300 scaffolds, with a total size of 1,162 Mb and an N50 of 955 kb. The completeness of this assembly was good as the BUSCO pipeline (v5.0.0) revealed that

it contained 98.7% of the Lepidoptera core genes (n=5,286) (Waterhouse *et al.* 2018). However, given the relatively high amount of duplicated BUSCO genes (7.8%), we deemed that it likely contained haplotigs, heterozygous overlaps and other assembly artefacts. In agreement with this hypothesis, a run of the purge_dup pipeline decreased the amount of duplicated BUSCO genes to 2.7% and removed a large amount of scaffolds (n = 1,748) with only minor effects on assembly size and N50. Our purged assembly totals 2,552 scaffolds that are 3,386 to 17,305,627-bp long (median length = 66,541 bp). Its N50 is 1,105 kbp and its size is 1,021 Mbp, which falls within the range of genome size estimates based on flow cytometry (C-value = 0.97 pg or 951 Mbp) (Calatayud *et al.* 2016) and k-mer frequency (971 Mbp [FindGSE] to 1,406 Mbp [GenomeScope]) (Table S1). The average Nanopore and Illumina sequencing depths are 46.3X and 38.9X, respectively, with 95.3% of the Illumina reads mapping to the purged assembly. The level of completness as assessed by the KAT pipeline was also good as 96.0% of the k-mer identified in the input illumina reads were included in our purged assembly (Mapleson *et al.* 2017). The missing 4% k-mer mostly corresponds to usual sequencing errors (Figure S2). KAT also estimated a very low level of heterozigosity (0.03%), leading to the absence of a heterozygous peak in the plots of k-mer frequencies (Figure S2). It is noteworthy that the genome size inferred by KAT was lower than the ones given by FindGSE and GenomeScope (560-730 Mb versus 960-1,600 Mb; Table S1), which may be due to the lower ability of KAT to properly estimate the size of genomes containing large amounts of repeated sequences. Related to this, the genome of *S. nonagrioides* is more than twice bigger than the other Noctuidae genomes sequenced to date (337-438 Mbp) (Table 1). This difference can be explained by a higher amount of repeated elements (661.6 versus 49.2-to-147.7 Mbp), which make up 64.78% of the *S. nonagrioides* genome, versus only 14 to 33.12% in the other Noctuidae (Figure 1). In fact, as seen in other groups of taxa (Sessegolo *et al.* 2016; Lower *et al.* 2017), genome size is correlated to the amount of repeated sequences in Lepidoptera (Talla *et al.* 2017), a trend that clearly holds among sequenced noctuid genomes included in our comparison (r=0.98 without *S. nonagrioides* and 0.99 when it is included). The quality of our *S. nonagrioides* purged assembly, as measured by its N50 and percent of core Lepidoptera genes, is close to that of the *Helicoverpa armigera* genome, the third best assembly of Noctuidae to date (Table 1).

Our search for contamination using Blobtools revealed that the amount of contaminating DNA present in our purged assembly is likely low. Among the 2,552 scaffolds of our purged assembly, we assigned 2,507 scaffolds to arthropods, representing 95.127% of the assembly size. Among the remaining 45 scaffolds, we retrieved no-hit for 25 of them and we assigned the rest to Chordates (2), undefined viruses (15), undefined (2) and Proteobacteria (1). Upon submission of the purged assembly to Genbank, the Proteobacteria scaffold was the only one identified by the NCBI staff as contaminated. It contains an internal 3,395-bp fragment showing 95% identity to the genome of *Escherichia coli* (K-12 strain C3026). This fragment is not covered by any Illumina reads so we removed it from our assembly. We manually placed each of the genome sequences lying upstream and downstream of this contaminant in two new scaffolds, leading to a total of 2,553 scaffolds in our final assembly. The sequencing depth and GC content of the remaining 44 scaffolds not assigned to arthropods fall in the range of the arthropod scaffolds, suggesting they may well correspond to *S. nonagrioides* DNA (Figure S3). Thus, we decided not to remove these scaffolds from our final assembly. Instead we listed them in Table S2 so that they can be easily retrieved and further studied or removed if needed.

Mitochondrial genome assembly

We assembled a complete circular mitogenome of 15,330 bp, which is 79.6% AT rich, and contains all expected 13 coding protein genes, 22 tRNA genes and two rRNA genes (Figure S4). We then used this sequence as a query to perform a sequence similarity search against our assembly to identify possible nuclear mitochondrial DNA (NUMTs) (Richly and Leister 2004). This search retrieved five significant alignments scattered on two scaffolds, for a total of 31.10 kb, a quantity falling within the range of what has been previously described in arthropods (Hazkani-Covo *et al.* 2010). One of the alignments is 735-bp long, it shows 96.19% identity to the mitogenome and it is located on scf7180000016552_1. The four remaining hits are all on the same scaffold (scf7180000018078_1). They are 15,328, 8,188, 4,637 and 2,216-bp long and all show more than 99.8% identity to the mitogenome (Figure 2). The assembly of the cluster, including two mitochondrial breakpoints and four nuclear-mitochondrial junctions, is supported by both Nanopore and Illumina reads (Figure 2). The sequencing depths at the nuclear-mitochondrial junctions (21X to 35X for trimmed Illumina reads and 46X to 55X for Nanopore reads longer than 7 kb) fall in the distribution of sequencing depths for the whole genome (average = 38.9X, SD = 27.3 for trimmed Illumina reads and average = 46.3X for Nanopore reads longer than 7 kb). We also validated the nuclear-mitochondrial junctions by PCR followed by Sanger sequencing (see methods). Thus, we conclude that this cluster results from the recent nuclear integration of two copies of the mitochondrial genome, one of which is rearranged in three pieces.

Genome annotation

Our automatic annotation of the *S. nonagrioides* genome yielded 17,230 protein-coding genes (average length = 10,570 bp) corresponding to 17.83% of the genome and including 85,919 exons (2.44% of the genome) (Table 2). We assigned 33.88% of all repeated sequences to a known superfamily of transposable elements (TEs) and classified another 1.03% of them as simple repeats (Figure 1B). The percentage of unclassified repeats (62.94%) is in the range of the other Noctuidae (17.78 to 89.79%). Among the classified TEs, *S. nonagrioides* has mostly LINE elements (70.66%), a similar percentage of LTR and DNA elements (17.13% and 12.21% respectively), and no SINE. This landscape, which will have to be refined using manual curation, is very similar to what was found in *T. ni* (Figure 1C). The two *Helicoverpa* species display the most different TE landscapes, where almost half of the classified TEs are DNA elements. We assessed the completeness of our annotation based on two metrics, the Annotation Edit Distance (AED) and the percentage of proteins with a Pfam domain, as recommended (Holt and Yandell 2011; Yandell and Ence 2012). The AED varies from 0 to 1, where 0 means a perfect congruence between gene annotation and its supporting evidence (Holt and Yandell 2011; Yandell and Ence 2012). A genome annotation with 90% of its gene models with an AED of 0.5 or better is considered as well annotated (Campbell *et al.* 2014). Here, we obtained an AED of 0.5 or better for 94.1% of our gene models. Regarding the second metric, it has been shown that the proportion of proteins with a Pfam domain is relatively stable between species, varying between 57% and 75% in eukaryotes (Yandell and Ence 2012). We found that 62.4% of *S. nonagrioides* proteins have a Pfam domain. Thus, both the AED and Pfam domain metrics indicate a relatively well-supported genome annotation. When compared to the other Noctuidae species, the

Table 1 Genome assembly statistics

Species	Number of fragments	Total size of the assembly (Mb)	N50 (kb)	Ns (%)	Complete BUSCO (duplicated) [a]
Sesamia nonagrioides	2,553	1,021	1,105	0.001	98.2% (2.7%)
Trichoplusia ni	601	339	894	0	94.3% (1.5%)
Spodoptera litura	2,974	438	13,592	2.488	99.1% (0.5%)
Spodoptera exigua	301	446	14,363	0.075	98.1% (1.2%)
Spodoptera frugiperda	37,235	358	54	7.732	86.3% (1.2%)
Helicoverpa armigera	997	337	1,000	11.009	98.3% (0.3%)
Helicoverpa zea	2,975	341	201	10.184	96.6% (0.8%)

[a] Lepidoptera core genes (n=5,286)

number of predicted genes in *S. nonagrioides* is in the range of the other species, although in the upper border (17,230 versus 11,595 – 17,707) (Table 2). We found that 91.56% of these predicted genes have an InterPro domain (71.47% - 93.2% in other Noctuidae).

Sex-determination genes

A good knowledge of sex determination in a pest species could be useful in the context of the sterile insect technique. It could help developing genetic sexing strains, in turn facilitating the mass production and release of sterile males (Marec and Vreysen 2019). We set out to provide a detailed annotation of genes likely involved in sex determination in *S. nonagrioides*. Sex is chromosomally-determined in lepidopterans, all species studied so far displaying a form of female-heterogamety (i.e. Z0/ZZ or a ZW/ZZ) (Traut *et al.* 2007). At the gene level, sex determination is best understood in *Bombyx mori*, which females carry a W dominant gene called *Feminizer (Fem)*. *Fem* is the precursor of a piwi-interacting RNA (piRNA) that downregulates the expression of a Z-linked gene: *Masculinizer (Masc)* (Kiuchi *et al.* 2014; Katsuma *et al.* 2014). In males, Masc splices doublesex (dsx) into its male isoform (dsxM). In females, fem piRNA inhibits Masc, leaving dsx in its default form, the female isoform (dsxF) (Nagaraju *et al.* 2014; Xu *et al.* 2017; Wang *et al.* 2019). In addition, the product of *IMP* (Insulin-like growth factor 2 mRNA-binding protein), a gene located on the Z chromosome, binds to PSI (P-element somatic inhibitor) in males. This interaction increases the binding activity of PSI to dsx, allowing PSI to participate with Masc in dsx mRNA splicing to its male isoform (Suzuki *et al.* 2010; Xu *et al.* 2017). Our automatic annotation coupled to alignments using *B. mori* genes as queries retrieved *bona fide* orthologs of *dsx*, *IMP* and *PSI* in our assembly of *S. nonagrioides*, the structure and genomic coordinates of which are given in Figure S5-7. The exons of *S. nonagrioides dsx (Sndsx)* align over the entire length of the female and male isoforms of *Bmdsx* (NP_001036871.1 and NP_001104815). The automatic annotation of *Sndsx* is incomplete as both the 5′ and 3′ UTRs of the gene are missing. Our similarity search for *SnPSI* retrieved all 14 coding exons of *BmPSI*. Its automatic annotation also includes predicted 5′ and 3′ UTRs. For *IMP*, we also found a complete ortholog gene, with a predicted 3′ UTR. Finally, our annotation of the *S. nonagrioides* ortholog of *Masc* is less complete, in agreement with the fact that this gene is less conserved among lepidopterans (Harvey-Samuel *et al.* 2020). The *BmMasc* gene encodes a 588 aa protein (NP_001296506). Using this protein as a query to perform a similarity search against the *Plutella xylostella* genome, Harvey-Samuel *et al.* (2020) identified two sequences encompassing a 7-aa long highly conserved motif of *Masc* which includes a cysteine-cysteine domain necessary for promoting male-specific splicing of dsx. One sequence was annotated as a zing finger CCCH domain-containing protein 10-like and the other as a cytokinesis protein SepA-like. An RNAi experiment allowed them to identify the second one as *PxyMasc*. Here, our similarity search returned 11 hits between 60 and 143 aa long, all on different scaffolds. Only one hit (position 210,793 to 211,113 of scaffold scf7180000016834_1) overlaps with the highly conserved cysteine-cysteine domain of *Masc*. This hit is 113 aa long and has 31.86% identity with the BmMasc protein.

Amylases

Obonyo *et al.* (2010) found that soluble materials deposited on the host caterpillar cuticle were important chemical cues for the proper recognition of the host by the female wasp in the host-parasitoid system *Chilo partellus* (Lepidoptera: Crambidae)/ *Cotesia flavipes* (Hymenoptera: Braconidae). Bichang'a *et al.* (2018) identified that the protein alpha-amylase from the oral secretions of the host caterpillar played an important role in antennation and oviposition behaviors prior to egg-laying. Therefore, we investigated alpha-amylase genes in more details in the *S. nonagrioides* genome. Our similarity search using the *Helicoverpa armigera* amylase protein sequence XP_021188243 as a query returned three different gene copies, hereafter named *SnAmy1* to *SnAmy3*, located on two scaffolds: scf7180000017447_1 (*SnAmy1* and *SnAmy2*) and scf7180000016148_1 (*SnAmy3*) (Figure S8). *SnAmy1* and *SnAmy2* are tandemly arranged in inverted orientation, 55 kbp apart. *SnAmy1* is 5,882-bp long; *SnAmy2* is 8753-bp long. Both encode exactly 500 amino acid long proteins. They share 97.6% nucleotide identity. *SnAmy3* is 7,198-bp long and diverges by 25% from the two other copies. The three genes have seven introns each. We found a subterminal intron located before the last three codons, as noticed in other Lepidopteran amylase genes and in some Hymenopteran amylase genes (Da Lage *et al.* 2011). For example, in *SnAmy2* we found the last three codons downstream of ca. 4 kb of intronic sequence. In *SnAmy3*, we showed by RT-PCR that two isoforms are transcribed through alternative splicing, with one isoform leading to the presence of a 42 amino acid long C-terminal tail to the protein through reading in-frame codons in the last intron up to the first stop found. Indeed, two isoforms are also found in the orthologous gene in *T. ni*. To date it is not known whether the longer isoform is translated. We also found *SnAmy1* and *SnAmy3* transcripts in salivary glands and in the midgut (not shown). Amylase genes often form multigene families in insects, with varying levels of divergence among copies (Da Lage 2018). We identified three amylase types in Lepidoptera, named type A, B, and C. Upon inspection of the phylogenetic tree (Figure S9),

Table 2 Genome annotation statistics

Species	Predicted genes	InterPro domains (% of predicted genes)	GO terms (% of predicted genes)	Pfam domain (% of predicted genes)	Number of exons in predicted genes / count per predicted gene
Sesamia nonagrioides	17,230	15,776 (91.56)	8,472 (49.17)	10,751 (62.40)	85,919 / 4.99
Trichoplusia ni	14,101	13,143 (93.2)	8,680 (61.56)	10,846 (76.91)	105,550 / 7.48
Spodoptera litura	15,317	13,637 (89.03)	11,440 (74.69)	NA	NA / 6.64
Spodoptera exigua	17,707	13,234 (74.74)	8,814 (49.78)	NA	NA / 5.88
Spodoptera frugiperda	11,595	NA	7,743 (66.79)	NA	64,725 / 5.58
Helicoverpa armigera	17,086	12,212 (71.47)	11,324 (66.28)	10,700 (62.62)	NA
Helicoverpa zea	15,200	11,061 (72.77)	10,221 (67.24)	9,795 (64.44)	NA

SnAmy1 and *SnAmy2* belong to type A and may result from a recent duplication since there is only one copy in *H. armigera*, whereas *SnAmy3* belongs to type B. The type C copy, which is ancestral to butterflies and moths, was lost in *S. nonagrioides*. Synteny comparison with *H. armigera* indicates that this type C copy was neighbor to the type A copies (not shown).

Investigation of horizontal transfer of bracoviruses

In its native range in Eastern Africa, *S. nonagrioides* is naturally parasitized by the braconid wasp *C. typhae* which is sister to *C. sesamiae* within the *C. flavipes* species complex (Kaiser *et al.* 2017). During oviposition, braconid wasps inject their eggs in host caterpillars together with bracoviruses. These bracoviruses contain circular DNA molecules (DNA circles) many of which typically become integrated into somatic host genomes. Integration of DNA circles will ensure proper persistence and expression of wasp genes during the development of wasp embryos (Beck *et al.* 2011; Chevignon *et al.* 2018). In addition, ancient events of horizontal transfer of bracoviral genes from wasps to various lepidopteran species have been reported, suggesting that integration of these genes has also occurred in the germline of lepidopterans (Gasmi *et al.* 2015; Di Lelio *et al.* 2019). Here, we investigated whether the *S. nonagrioides* genome contains traces of wasp DNA circles resulting from recent events of HT from wasp to moth. Given that the circles of *C. typhae* have not been sequenced, we used the 26 DNA circles of the sister species *C. sesamiae* (Jancek *et al.* 2013) (NCBI BioProject PRJEB1050) as queries to perform similarity searches on our assembly. Our results revealed no evidence for recent events of HT of DNA circles from *Cotesia* wasps to *S. nonagrioides*. Specifically, we retrieved significant alignments only for three circles (2, 28 and 32,) and they all covered less than 2% of the circle length. Interestingly however, a region of circle 32 (HF562927.1, position 18,762 to 19,959) yielded 46 hits longer than 500 bp (up to 678 bp) showing 95.4 to 99.4% nucleotide identity. We used this 1197-bp sequence as a query to perform a similarity search against GenBank non-redundant proteins and against a custom TE protein database, which yielded no significant alignment. However, this region yielded a 209-bp significant alignment showing 88.7% identity to a *B. mori* helitron (Helitron-N1_BM, 266-bp long). Given the high nucleotide identity between the wasp and moth sequences (95.4 to 99.4%) and the deep divergence time between hymenopterans and lepidopterans (>300 million years (Misof *et al.* 2014)), we infer that this helitron-like sequence has been recently transferred between *S. nonagrioides* and *C. sesamiae*. This event adds up to the list of helitrons reported to have undergone HT between parasitoid wasps and lepidopterans (Thomas *et al.* 2010; Guo *et al.* 2014; Coates 2015; Heringer *et al.* 2017; Han *et al.* 2019). Whether these transfers were facilitated by the integration of wasp DNA circles in germline genomes of lepidopterans larvae during parasitism is an interesting possibility that deserves further investigation.

CONCLUSIONS AND PERSPECTIVES

We have assembled the complete mitochondrial genome and a draft nuclear genome of *S. nonagrioides*. The nuclear genome is remarkable in that it is the largest noctuid genome sequenced by far, being two to three times larger than the 10 other noctuid genomes available in GenBank as of January 2021. This difference merely stems from a higher repeat content in *S. nonagrioides*, in line with the known correlation between genome size and the amount of repeated sequences. It will be interesting to decipher the causes of this higher repeat content, by comparing population sizes, mutation rates and the dynamics of TE activity between the various noctuid species. We found no sign of recent HT from the bracovirus circles of *C. sesamiae*, which is sister to *C. typhae*, to *S. nonagrioides*. However, it will be necessary to repeat this analysis using the bracovirus circles from *C. typhae*, the very species that parasitizes *S. nonagrioides*. Finally, given the N50 of the nuclear genome assembly and the high percent of core Lepidoptera genes it contains, we predicted that the vast majority of *S. nonagrioides* genes are present in one scaffold and can be easily retrieved. This genome thus provides a solid tool to further study the evolutionary history of Noctuidae and it represents an interesting new asset to develop biocontrol strategies against *S. nonagrioides*.

DATA AVAILABILITY STATEMENT

The data associated to this paper is available on NCBI under the BioProject ID PRJNA680928 and GenBank accession number JADWQK000000000. The BioProject includes the annotated nuclear and mitochondrial assemblies and the raw short and long reads. The data is also available in the DRYAD database at the following address: https://doi.org/10.5061/dryad.dfn2z3515. Supplemental Material available at figshare: https://doi.org/10.25387/g3.14185070. Figure S1 shows the electropherogram and its corresponding gel generated by a fragment analyzer. Figure S2 shows the plots generated by GenomeScope, FindGSE and KAT. Figure S3 shows the Blobplot of *S. nonagrioides* scaffolds. Figure S4 is a map of the annotated *S. nonagrioides* mitogenome generated with mitoZ. Figures S5 to S7 show the structure of the genes involved in sex determination. Figure S8 shows the structure of the alpha-amylase gene copies. Figure S9 shows the Maximum Likelihood tree of lepidopteran

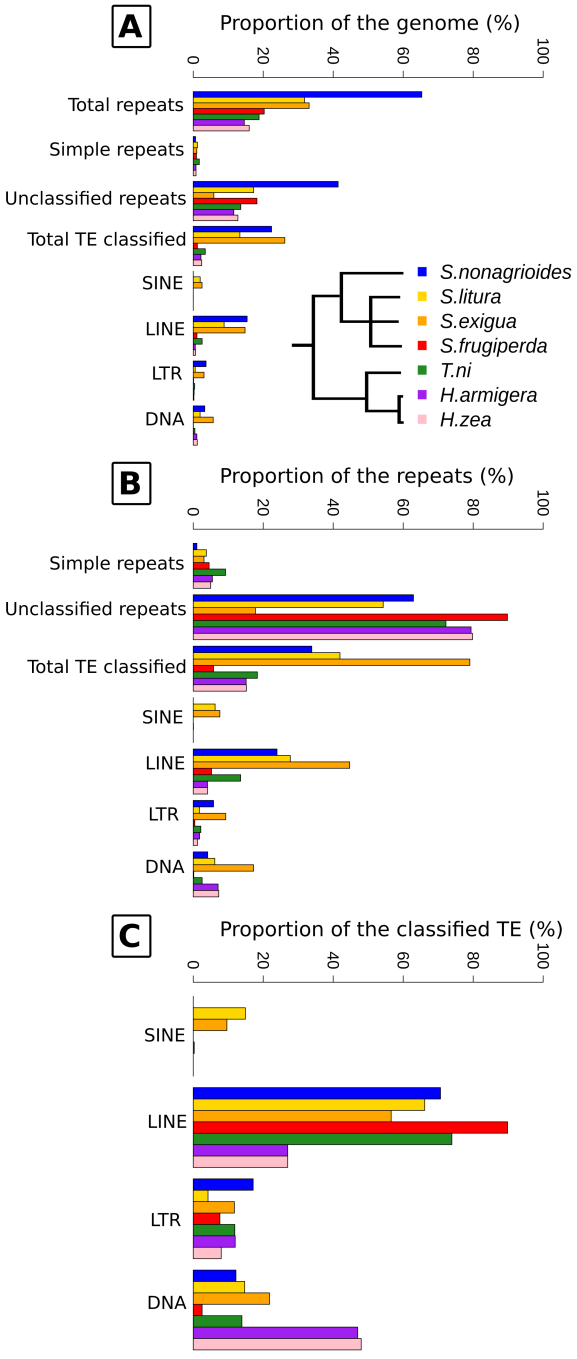


Figure 1 Comparison of repeat contents in sequenced Noctuidae genomes. **(A)** Genome proportion of each type of repeats as annotated by RepeatModeler v2.0 and RepeatMasker v4.1. **(B)** Proportion of each type of repeats among all repeated sequences **(C)**. Proportion of each type of transposable elements (TEs) among all TEs classified by RepeatModeler V2.0. Phylogenetic relationships among Noctuidae are taken from (Toussaint *et al.* 2012; Talsania *et al.* 2019).

alpha-amylases. Table S1 lists size estimates for the *S. nonagrioides* genome. Table S2 lists the name of all 44 scaffolds not assigned to arthropods.

FUNDING

This study was funded by Agence Nationale de la Recherche (ANR CoteBio ANR-17-CE32-0015).

LITERATURE CITED

Alamalakala, L., S. Parimi, N. Patel, and B. Char, 2018 Insect RNAi: Integrating a New Tool in the Crop Protection Toolkit. In *Trends in Insect Molecular Biology and Biotechnology*, edited by D. Kumar and C. Gong, pp. 193–232, Springer International Publishing, Cham.

Beck, M. H., S. Zhang, K. Bitra, G. R. Burke, and M. R. Strand, 2011 The Encapsidated Genome of Microplitis demolitor Bracovirus Integrates into the Host Pseudoplusia includens. Journal of Virology **85**: 11685–11696.

Beevor, P. S., H. David, and O. T. Jones, 1990 Female sex Pheromones of Chilo spp. (Lepidoptera: Pyralidae) and their Development in Pest Control Applications. International Journal of Tropical Insect Science **11**: 787–794.

Bichang'a, G., J.-L. Da Lage, C. Capdevielle-Dulac, M. Zivy, T. Balliau, *et al.*, 2018 α-Amylase Mediates Host Acceptance in the Braconid Parasitoid Cotesia flavipes. Journal of Chemical Ecology **44**: 1030–1039.

Bolger, A. M., M. Lohse, and B. Usadel, 2014 Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics **30**: 2114–2120.

Bosque-Perez, Schulthess, and Polaszek, 1998 African cereal stem borers: Economic importance, taxonomy, natural enemies and control. https://www.cabi.org/cpc/abstract/19981108334.

Calatayud, P.-A., C. Petit, N. Burlet, S. Dupas, N. Glaser, *et al.*, 2016 Is genome size of Lepidoptera linked to host plant range? Entomologia Experimentalis et Applicata **159**: 354–361.

Camargo, A. M., D. A. Andow, P. Castañera, and G. P. Farinós, 2018 First detection of a Sesamia nonagrioides resistance allele to Bt maize in Europe. Scientific Reports **8**: 3977.

Campbell, M. S., C. Holt, B. Moore, and M. Yandell, 2014 Genome Annotation and Curation Using MAKER and MAKER-P. Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.] **48**: 4.11.1–4.11.39.

Chen, W., X. Yang, G. Tetreau, X. Song, C. Coutu, *et al.*, 2019 A high-quality chromosome-level genome assembly of a generalist herbivore, Trichoplusia ni. Molecular Ecology Resources **19**: 485–496.

Cheng, T., J. Wu, Y. Wu, R. V. Chilukuri, L. Huang, *et al.*, 2017 Genomic adaptation to polyphagy and insecticides in a major East Asian noctuid pest. Nature Ecology & Evolution **1**: 1747–1756.

Chevignon, G., G. Periquet, G. Gyapay, N. Vega-Czarny, K. Musset, *et al.*, 2018 Cotesia congregata Bracovirus Circles Encoding PTP and Ankyrin Genes Integrate into the DNA of Parasitized Manduca sexta Hemocytes. Journal of Virology **92**.

Coates, B. S., 2015 Horizontal transfer of a non-autonomous Helitron among insect and viral genomes. BMC Genomics **16**: 137.

Cork, A., N. Q. Kamal, S. N. Alam, J. C. S. Choudhury, and N. S. Talekar, 2003 PHEROMONES AND THEIR APPLICATIONS TO INSECT PEST CONTROL. Journal of Entomology p. 15.

Da Lage, J.-L., 2018 The Amylases of Insects. International Journal of Insect Science **10**: 1179543318804783.

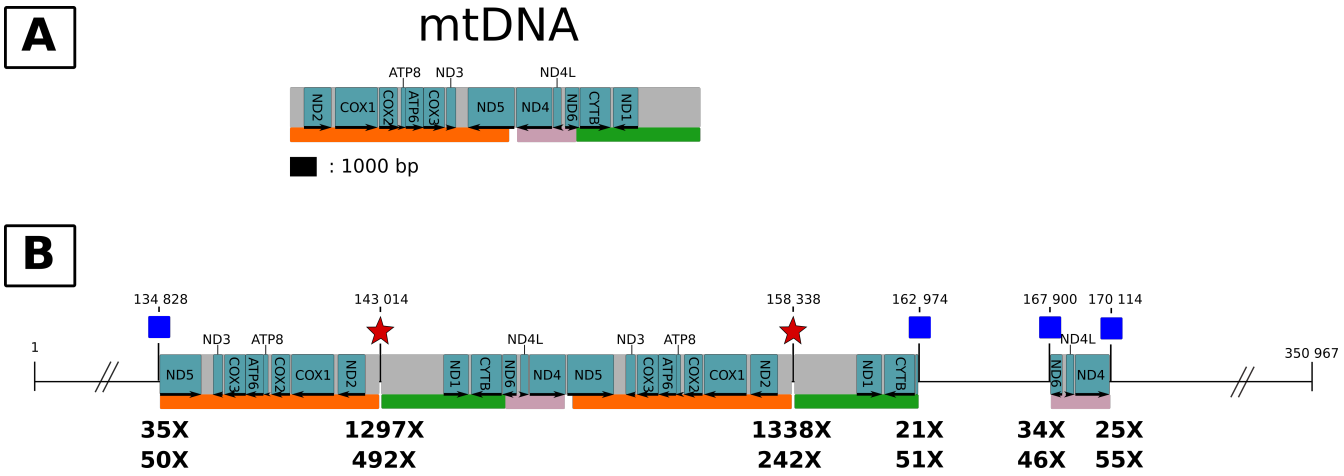


Figure 2 A large nuclear mitochondrial sequence (NUMT) in the *S. nonagrioides* genome. **A**. Map of the mitogenome assembly. Protein coding genes are represented by blue rectangles, in which arrows show their directions. The names of the genes are indicated in the blue rectangles or on top of them. The scale at the bottom of the mtDNA applies also for the panel B. **B**. Map of scaffold scf7180000018078_1 showing the coordinates of the four blastn hits obtained using the mitogenome as a query on the *Sesamia nonagrioides* nuclear assembly. Only the NUMT sequence is drawn to scale, the rest of the scaffold is not, as indicated by broken lines. The scale of the NUMT is the same as in A. The four hits are located between each coordinate and the orange, green and pink rectangles help to visualize which parts of the mtDNA have been integrated and possibly rearranged in the nuclear DNA. The nuclear-mitochondrial junctions (blue squares) are supported by short reads (depth of the trimmed reads indicated by the top numbers under the junctions). Long reads support both nuclear-mitochondrial junctions (depth of the reads longer than 7kb indicated by the bottom numbers under the junctions) and mitochondrial breakpoints (red stars). Indeed, we visually checked that some long reads straddle on the mitochondrial breakpoints and the nuclear genome. The coverage at the mitochondrial breakpoints is much higher because these breakpoints are also mapped by all reads originating from the cytoplasmic mitogenomes. The first, second and fourth nuclear-mitochondrial junctions are also supported by PCR.

Da Lage, J.-L., F. Maczkowiak, and M.-L. Cariou, 2011 Phylogenetic Distribution of Intron Positions in Alpha-Amylase Genes of Bilateria Suggests Numerous Gains and Losses. PLOS ONE **6**: e19673.

De Coster, W., S. D'Hert, D. T. Schultz, M. Cruts, and C. Van Broeckhoven, 2018 NanoPack: Visualizing and processing long-read sequencing data. Bioinformatics **34**: 2666–2669.

Di Lelio, I., A. Illiano, F. Astarita, L. Gianfranceschi, D. Horner, *et al.*, 2019 Evolution of an insect immune barrier through horizontal gene transfer mediated by a parasitic wasp. PLOS Genetics **15**: e1007998.

Farinós, G. P., P. Hernández-Crespo, F. Ortego, and P. Castañera, 2018 Monitoring of Sesamia nonagrioides resistance to MON 810 maize in the European Union: Lessons from a long-term harmonized plan. Pest Management Science **74**: 557–568.

Gasmi, L., H. Boulain, J. Gauthier, A. Hua-Van, K. Musset, *et al.*, 2015 Recurrent Domestication by Lepidoptera of Genes from Their Parasites Mediated by Bracoviruses. PLOS Genetics **11**: e1005470.

Giacometti, R., 1995 Rearing of Sesamia nonagrioides Lefevre on a meridic diet (Lepidoptera, Noctuidae). Redia **78**: 19–27.

Glaser, N., A. Gallot, F. Legeai, M. Harry, L. Kaiser, *et al.*, 2015 Differential expression of the chemosensory transcriptome in two populations of the stemborer Sesamia nonagrioides. Insect Biochemistry and Molecular Biology **65**: 28–34.

Guan, D., S. A. McCarthy, J. Wood, K. Howe, Y. Wang, *et al.*, 2020 Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics **36**: 2896–2898.

Guo, X., J. Gao, F. Li, and J. Wang, 2014 Evidence of horizontal transfer of non-autonomous Lep 1 Helitrons facilitated by host-parasite interactions. Scientific Reports **4**: 5119.

Han, G., N. Zhang, J. Xu, H. Jiang, C. Ji, *et al.*, 2019 Characterization of a novel Helitron family in insect genomes: Insights into classification, evolution and horizontal transfer. Mobile DNA **10**: 25.

Hardwick, K. M., A. M. E. Ojwang', F. Stomeo, S. Maina, G. Bichang'a, *et al.*, 2019 Draft Genome of Busseola fusca, the Maize Stalk Borer, a Major Crop Pest in Sub-Saharan Africa. Genome Biology and Evolution **11**: 2203–2207.

Harvey-Samuel, T., V. C. Norman, R. Carter, E. Lovett, and L. Alphey, 2020 Identification and characterization of a Masculinizer homologue in the diamondback moth, Plutella xylostella. Insect Molecular Biology **29**: 231–240.

Hazkani-Covo, E., R. M. Zeller, and W. Martin, 2010 Molecular Poltergeists: Mitochondrial DNA Copies (numts) in Sequenced Nuclear Genomes. PLOS Genetics **6**: e1000834.

Heringer, P., G. B. Dias, and G. C. S. Kuhn, 2017 A Horizontally Transferred Autonomous Helitron Became a Full Polydnavirus Segment in Cotesia vestalis. G3: Genes, Genomes, Genetics **7**: 3925–3935.

Holt, C. and M. Yandell, 2011 MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics **12**: 491.

Jancek, S., A. Bézier, P. Gayral, C. Paillusson, L. Kaiser, *et al.*, 2013 Adaptive Selection on Bracovirus Genomes Drives the Specialization of Cotesia Parasitoid Wasps. PLoS ONE **8**.

Jin, L., A. S. Walker, G. Fu, T. Harvey-Samuel, T. Dafa'alla, *et al.*, 2013 Engineered Female-Specific Lethality for Control of Pest Lepidoptera. ACS Synthetic Biology **2**: 160–166.

Kaiser, L., J. Fernandez-Triana, C. Capdevielle-Dulac, C. Chantre,

M. Bodet, *et al.*, 2017 Systematics and biology of Cotesia typhae sp. n. (Hymenoptera, Braconidae, Microgastrinae), a potential biological control agent against the noctuid Mediterranean corn borer, Sesamia nonagrioides. ZooKeys pp. 105–136.

Kakumani, P. K., P. Malhotra, S. K. Mukherjee, and R. K. Bhatnagar, 2014 A draft genome assembly of the army worm, Spodoptera frugiperda. Genomics **104**: 134–143.

Kankonda, O. M., B. D. Akaibe, N. M. Sylvain, and B.-P. L. Ru, 2018 Response of maize stemborers and associated parasitoids to the spread of grasses in the rainforest zone of Kisangani, DR Congo: Effect on stemborers biological control. Agricultural and Forest Entomology **20**: 150–161.

Katsuma, S., M. Kawamoto, and T. Kiuchi, 2014 Guardian small RNAs and sex determination. RNA Biology **11**: 1238–1242.

Kergoat, G. J., E. F. A. Toussaint, C. Capdevielle-Dulac, A.-L. Clamens, G. Ong'amo, *et al.*, 2015 Integrative taxonomy reveals six new species related to the Mediterranean corn stalk borer Sesamia nonagrioides (Lefèbvre) (Lepidoptera, Noctuidae, Sesamiina). Zoological Journal of the Linnean Society **175**: 244–270.

Kfir, R., W. A. Overholt, Z. R. Khan, and A. Polaszek, 2002 Biology and Management of Economically Important Lepidopteran Cereal Stem Borers in Africa. Annual Review of Entomology **47**: 701–731.

Kiuchi, T., H. Koga, M. Kawamoto, K. Shoji, H. Sakai, *et al.*, 2014 A single female-specific piRNA is the primary determiner of sex in the silkworm. Nature **509**: 633–636.

Laetsch, D. R. and M. L. Blaxter, 2017 BlobTools: Interrogation of genome assemblies. F1000Research **6**: 1287.

Langmead, B. and S. L. Salzberg, 2012 Fast gapped-read alignment with Bowtie 2. Nature Methods **9**: 357–359.

Leung, K., E. Ras, K. B. Ferguson, S. Ariëns, D. Babendreier, *et al.*, 2020 Next-generation biological control: The need for integrating genetics and genomics. Biological Reviews **n/a**.

Lower, S. S., J. S. Johnston, K. F. Stanger-Hall, C. E. Hjelmen, S. J. Hanrahan, *et al.*, 2017 Genome Size in North American Fireflies: Substantial Variation Likely Driven by Neutral Processes. Genome Biology and Evolution **9**: 1499–1512.

Mapleson, D., G. Garcia Accinelli, G. Kettleborough, J. Wright, and B. J. Clavijo, 2017 KAT: A K-mer analysis toolkit to quality control NGS datasets and genome assemblies. Bioinformatics **33**: 574–576.

Marçais, G. and C. Kingsford, 2011 A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics **27**: 764–770.

Marec, F. and M. J. B. Vreysen, 2019 Advances and Challenges of Using the Sterile Insect Technique for the Management of Pest Lepidoptera. Insects **10**: 371.

Meng, G., Y. Li, C. Yang, and S. Liu, 2019 MitoZ: A toolkit for animal mitochondrial genome assembly, annotation and visualization. Nucleic Acids Research **47**: e63–e63.

Midingoyi, S.-k. G., H. D. Affognon, I. Macharia, G. Ong'amo, E. Abonyo, *et al.*, 2016 Assessing the long-term welfare effects of the biological control of cereal stemborer pests in East and Southern Africa: Evidence from Kenya, Mozambique and Zambia. Agriculture, Ecosystems & Environment **230**: 10–23.

Misof, B., S. Liu, K. Meusemann, R. S. Peters, A. Donath, *et al.*, 2014 Phylogenomics resolves the timing and pattern of insect evolution. Science (New York, N.Y.) **346**: 763–767.

Moscardi, F., 1999 Assessment of the application of baculoviruses for control of Lepidoptera. Annual Review of Entomology **44**: 257–289.

Moyal, P., P. Tokro, A. Bayram, M. Savopoulou-Soultani, E. Conti, *et al.*, 2011 Origin and taxonomic status of the Palearctic population of the stem borer Sesamia nonagrioides (Lefèbvre) (Lepidoptera: Noctuidae). Biological Journal of the Linnean Society **103**: 904–922.

Muirhead, K. A., N. P. Murphy, N. Sallam, S. C. Donnellan, and A. D. Austin, 2012 Phylogenetics and genetic diversity of the Cotesia flavipes complex of parasitoid wasps (Hymenoptera: Braconidae), biological control agents of lepidopteran stemborers. Molecular Phylogenetics and Evolution **63**: 904–914.

Nagaraju, J., G. Gopinath, V. Sharma, and J. N. Shukla, 2014 Lepidopteran Sex Determination: A Cascade of Surprises. Sexual Development **8**: 104–112.

Naino-Jika, A. K., B. L. Ru, C. Capdevielle-Dulac, F. Chardonnet, J. F. Silvain, *et al.*, 2020 Population genetics of the Mediterranean corn borer (Sesamia nonagrioides) differs between wild and cultivated plants. PLOS ONE **15**: e0230434.

Obonyo, M., F. Schulthess, B. Le Ru, J. van den Berg, J.-F. Silvain, *et al.*, 2010 Importance of contact chemical cues in host recognition and acceptance by the braconid larval endoparasitoids Cotesia sesamiae and Cotesia flavipes. Biological Control **54**: 270–275.

Overholt, W. A., J. O. Ochieng, P. Lammers, and K. Ogedah, 1994 Rearing and Field Release Methods for Cotesia Flavipes Cameron (Hymenoptera: Braconidae), a Parasitoid of Tropical Gramineous Stem Borers. International Journal of Tropical Insect Science **15**: 253–259.

Pearce, S. L., D. F. Clarke, P. D. East, S. Elfekih, K. H. J. Gordon, *et al.*, 2017 Genomic innovations, transcriptional plasticity and gene loss underlying the evolution and divergence of two highly polyphagous and invasive Helicoverpa pest species. BMC Biology **15**: 63.

Richly, E. and D. Leister, 2004 NUMTs in Sequenced Eukaryotic Genomes. Molecular Biology and Evolution **21**: 1081–1084.

Sessegolo, C., N. Burlet, and A. Haudry, 2016 Strong phylogenetic inertia on genome size and transposable element content among 26 species of flies. Biology Letters **12**: 20160407.

Sun, H., J. Ding, M. Piednoël, and K. Schneeberger, 2018 findGSE: Estimating genome size variation within human and Arabidopsis using k-mer frequencies. Bioinformatics **34**: 550–557.

Suzuki, M. G., S. Imanishi, N. Dohmae, M. Asanuma, and S. Matsumoto, 2010 Identification of a Male-Specific RNA Binding Protein That Regulates Sex-Specific Splicing of Bmdsx by Increasing RNA Binding Activity of BmPSI. Molecular and Cellular Biology **30**: 5776–5786.

Talla, V., A. Suh, F. Kalsoom, V. Dincă, R. Vila, *et al.*, 2017 Rapid Increase in Genome Size as a Consequence of Transposable Element Hyperactivity in Wood-White (Leptidea) Butterflies. Genome Biology and Evolution **9**: 2491–2505.

Talsania, K., M. Mehta, C. Raley, Y. Kriga, S. Gowda, *et al.*, 2019 Genome Assembly and Annotation of the Trichoplusia ni TniFNL Insect Cell Line Enabled by Long-Read Technologies. Genes **10**.

Thomas, J., S. Schaack, and E. J. Pritham, 2010 Pervasive horizontal transfer of rolling-circle transposons among animals. Genome Biology and Evolution **2**: 656–664.

Tian, H., H. Peng, Q. Yao, H. Chen, Q. Xie, *et al.*, 13 juil. 2009 Developmental Control of a Lepidopteran Pest Spodoptera exigua by Ingestion of Bacteria Expressing dsRNA of a Non-Midgut Gene. PLOS ONE **4**: e6225.

Toussaint, E. F. A., F. L. Condamine, G. J. Kergoat, C. Capdevielle-Dulac, J. Barbut, *et al.*, 2012 Palaeoenvironmental Shifts Drove

the Adaptive Radiation of a Noctuid Stemborer Tribe (Lepidoptera, Noctuidae, Apameini) in the Miocene. PLOS ONE **7**: e41377.

Traut, W., K. Sahara, and F. Marec, 2007 Sex Chromosomes and Sex Determination in Lepidoptera. Sexual Development **1**: 332–346.

Vurture, G. W., F. J. Sedlazeck, M. Nattestad, C. J. Underwood, H. Fang, *et al.*, 2017 GenomeScope: Fast reference-free genome profiling from short reads. Bioinformatics **33**: 2202–2204.

Wang, Y.-H., X.-E. Chen, Y. Yang, J. Xu, G.-Q. Fang, *et al.*, 2019 The Masc gene product controls masculinization in the black cutworm, Agrotis ipsilon. Insect Science **26**: 1037–1044.

Waterhouse, R. M., M. Seppey, F. A. Simão, M. Manni, P. Ioannidis, *et al.*, 2018 BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. Molecular Biology and Evolution **35**: 543–548.

Xu, J., S. Chen, B. Zeng, A. A. James, A. Tan, *et al.*, 2017 Bombyx mori P-element Somatic Inhibitor (BmPSI) Is a Key Auxiliary Factor for Silkworm Male Sex Determination. PLOS Genetics **13**: e1006576.

Yandell, M. and D. Ence, 2012 A beginner's guide to eukaryotic genome annotation. Nature Reviews Genetics **13**: 329–342.

Zhang, F., J. Zhang, Y. Yang, and Y. Wu, 2019 A chromosome-level genome assembly for the beet armyworm (Spodoptera exigua) using PacBio and Hi-C sequencing. bioRxiv p. 2019.12.26.889121.

Zhu, J.-Y., Z.-W. Xu, X.-M. Zhang, and N.-Y. Liu, 2018 Genome-based identification and analysis of ionotropic receptors in Spodoptera litura. The Science of Nature **105**: 38.

Zimin, A. V., D. Puiu, M.-C. Luo, T. Zhu, S. Koren, *et al.*, 2017 Hybrid assembly of the large and highly repetitive genome of Aegilops tauschii, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. Genome Research **27**: 787–792.

9 - Article n°3: Genome-Wide Patterns of Bracovirus Chromosomal Integration into Multiple Host Tissues during Parasitism

In this article, we report a high-quality assembly for the genome of *C. typhae* and we investigate whether any of its bracovirus (CtBV) DNA circles could integrate in the genome of several tissues of *S. nonagrioides* during parasitism. We also provide more insights on the mechanisms by which bracoviral DNA circles integrate in the host genome. All the experiments, the assembly, and the automatic annotation of *C. typhae* genome were done before my PhD by co-authors. I began this work during my M2 internship, and altogether, the study took most of the first year of my PhD to complete. I manually annotated 27 proviral segments in the genome of *C. typhae*, and I achieved the analyses on CtBV integration. We found that the 16 segments harboring the conserved Host Integration Motif (HIM) massively integrate in *S. nonagrioides* genome in all its tissues/structures we studied. Interestingly, CtBV also undergoes chromosomal integration in a caterpillar in which parasitism failed, raising the question of the possibility of vertical transmission of CtBV by surviving caterpillars.

AMERICAN SOCIETY FOR MICROBIOLOGY | Journal of Virology®

GENETIC DIVERSITY AND EVOLUTION

Genome-Wide Patterns of Bracovirus Chromosomal Integration into Multiple Host Tissues during Parasitism

Héloïse Muller,[a] Mohamed Amine Chebbi,[b,c] Clémence Bouzar,[a] George Périquet,[b] Taiadjana Fortuna,[a] Paul-André Calatayud,[a,d] Bruno Le Ru,[a,d] Julius Obonyo,[d] Laure Kaiser,[a] Jean-Michel Drezen,[b] Elisabeth Huguet,[b] Clément Gilbert[a]

[a]Université Paris-Saclay, CNRS, IRD, UMR Évolution, Génomes, Comportement, et Écologie, Gif-sur-Yvette, France
[b]UMR 7261 CNRS, Institut de Recherche sur la Biologie de l'Insecte, Faculté des Sciences et Techniques, Université de Tours, Tours, France
[c]ViroScan3D SAS, Lyon, France
[d]International Centre of Insect Physiology and Ecology, Institut de Recherche pour le Développement Team, Nairobi, Kenya

Elisabeth Huguet and Clément Gilbert are equal senior authors.

ABSTRACT Bracoviruses are domesticated viruses found in parasitic wasp genomes. They are composed of genes of nudiviral origin that are involved in particle production and proviral segments containing virulence genes that are necessary for parasitism success. During particle production, proviral segments are amplified and individually packaged as DNA circles in nucleocapsids. These particles are injected by parasitic wasps into host larvae together with their eggs. Bracovirus circles of two wasp species were reported to undergo chromosomal integration in parasitized host hemocytes, through a conserved sequence named the host integration motif (HIM). Here, we used bulk Illumina sequencing to survey integrations of *Cotesia typhae* bracovirus circles in the DNA of its host, the maize corn borer (*Sesamia nonagrioides*), 7 days after parasitism. First, assembly and annotation of a high-quality genome for *C. typhae* enabled us to characterize 27 proviral segments clustered in proviral loci. Using these data, we characterized large numbers of chromosomal integrations (from 12 to 85 events per host haploid genome) for all 16 bracovirus circles containing a HIM. Integrations were found in four *S. nonagrioides* tissues and in the body of a caterpillar in which parasitism had failed. The 12 remaining circles do not integrate but are maintained at high levels in host tissues. Surprisingly, we found that HIM-mediated chromosomal integration in the wasp germ line has occurred accidentally at least six times during evolution. Overall, our study furthers our understanding of wasp-host genome interactions and supports HIM-mediated chromosomal integration as a possible mechanism of horizontal transfer from wasps to their hosts.

IMPORTANCE Bracoviruses are endogenous domesticated viruses of parasitoid wasps that are injected together with wasp eggs into wasp host larvae during parasitism. Several studies have shown that some DNA circles packaged into bracovirus particles become integrated into host somatic genomes during parasitism, but the phenomenon has never been studied using nontargeted approaches. Here, we use bulk Illumina sequencing to systematically characterize and quantify bracovirus circle integrations that occur in four tissues of the Mediterranean corn borer (*Sesamia nonagrioides*) during parasitism by the *Cotesia typhae* wasp. Our analysis reveals that all circles containing a HIM integrate at substantial levels (from 12 to 85 integrations per host cell, in total) in all tissues, while other circles do not integrate. In addition to shedding new light on wasp-bracovirus-host interactions, our study supports HIM-mediated chromosomal integration of bracovirus as a possible source of wasp-to-host horizontal transfer, with long-term evolutionary consequences.

KEYWORDS bracovirus, chromosomal integration, genomics, horizontal transfer, host-parasite relationship, parasitoid wasps, polydnavirus

Citation Muller H, Chebbi MA, Bouzar C, Périquet G, Fortuna T, Calatayud P-A, Le Ru B, Obonyo J, Kaiser L, Drezen J-M, Huguet E, Gilbert C. 2021. Genome-wide patterns of bracovirus chromosomal integration into multiple host tissues during parasitism. J Virol 95:e00684-21. https://doi.org/10.1128/JVI.00684-21.

Editor Colin R. Parrish, Cornell University

Address correspondence to Elisabeth Huguet, elisabeth.huguet@univ-tours.fr, or Clément Gilbert, clement.gilbert@egce.cnrs-gif.fr.

Received 22 April 2021
Accepted 7 July 2021
Accepted manuscript posted online 28 July 2021
Published 27 October 2021

Polydnavirus genomes within parasitoid wasps (Hymenoptera) are composed of domesticated viral genes and genes of different origins involved in virulence (1–3). The domesticated viruses encode viral particles akin to gene transfer agents, which are injected during oviposition into the lepidopteran hosts of parasitoid wasps and are necessary for successful development of the wasp larvae. Polydnaviruses result from large double-stranded DNA virus endogenization events that took place during the course of parasitoid wasp evolution (1, 4–12). Polydnaviruses of the *Ichnovirus* genus identified in the genomes of certain Ichneumonidae Campopleginae and Banchinae wasps are thought to originate from closely related virus ancestors, the nature of which is still unknown but could possibly correspond to nucleocytoplasmic large DNA viruses (10). Polydnaviruses belonging to the *Bracovirus* genus result from endogenization of a nudivirus that occurred about 100 million years ago in the ancestor of the microgastroid complex of braconid wasps (7, 13, 14), a hyperdiversified monophyletic group estimated to contain at least 46,000 species (15). Once integrated in the ancestor of microgastroid wasps, the nudivirus genes were domesticated and inherited vertically in all branches of the microgastroid tree for 100 million years. All microgastroid species studied to date express nudivirus-derived bracovirus genes in specialized cells located in a region of the ovaries named the calyx. The products of these genes form viral particles containing virulence genes (2, 16). Although the evolutionary origin of virulence genes has in most cases not been uncovered, some are clearly of wasp origin (17, 18), while others derive from transposable elements (TEs) (3, 19). They are located on so-called proviral segments dispersed in multiple chromosomal regions in the wasp's genome (13, 20); the major one, named the macrolocus, spans 2 Mb and includes two-thirds of the proviral segments. Chromosome-scale assembly of the genome of *Cotesia congregata* (Microgastrinae) revealed that it contains 10 proviral loci (PL), each made of 1 to 18 proviral segments (PL2, 18 segments [3]). Comparisons between *Cotesia* species and *Microplitis demolitor* showed that the synteny of these PL is well conserved along the phylogeny of Braconidae in ~53 million years of evolution, suggesting strong evolutionary constraints associated with the function of the segments.

PL belong to units that are amplified in calyx cells during particle production (21). Among those amplified units (replication units [RU]), segments are excised and circularized through site-specific recombination, which involves direct repeat junctions (DRJs) located at their extremities (3, 22–26). The resulting double-stranded DNA circles are finally packaged into viral particles, which are released in the oviduct lumen and injected into the host together with the wasp's eggs during oviposition. Once in caterpillar host cells, DNA circle-borne virulence genes are expressed (27). Interestingly, several studies using cell culture or *in vivo* models have shown that at least some circles persist in cell lines or over the entire duration of wasp development in the form of chromosomally integrated forms (26, 28–30). Using a PCR-based approach with Sanger sequencing, integration of DNA circles was shown to occur upon parasitism for 2 circles in the hemocytes of the host of the *Microplitis demolitor* wasp (Microgastrinae) via a motif called the host integration motif (HIM) that is conserved in all *M. demolitor* bracovirus (MdBV) circles (26). Another study using primer extension capture followed by high-throughput sequencing unveiled several thousand chromosomal integrations for 8 circles of *Cotesia congregata* in the hemocytes of its host, the tobacco hornworm *Manduca sexta* (30). The 8 *C. congregata* proviral circles surveyed in this study contain HIM and, as reported for *M. demolitor*, all integrations of these circles involved two motifs, called junction 1 (J1) and junction 2 (J2), located within the HIM. J1 and J2 correspond to sequences that form the extremities of the viral sequences when integrated into lepidopteran host DNA. It was further shown that, as for MdBV circles, an ~50-bp sequence located between J1 and J2 is lost upon integration. Integration of polydnavirus circles is not limited to bracoviruses and was also recently described for ichnoviruses. This suggests that this phenomenon plays an important role in the parasitism

success of parasitoid wasps, and it reveals shared characteristics in the mechanisms underlying integration of ichnoviruses and bracoviruses (31).

In this study, we used a bulk, rather than targeted, sequencing approach to investigate polydnavirus circle persistence and integration in the host-parasitoid system involving the *Cotesia typhae* wasp (Microgastrinae, Braconidae) and its natural host, the Mediterranean corn borer (*Sesamiae nonagrioides*, Noctuidae). *Cotesia typhae* is a recently described species among the *Cotesia flavipes* species complex, and it is native to eastern sub-Saharan Africa (32). In its natural environment in East Africa, it exclusively parasitizes larvae of *S. nonagrioides* dwelling on Typhaceae plants. It is also able to parasitize *S. nonagrioides* larvae in cultivated maize fields from France; therefore, it is currently being studied as a possible biocontrol agent against this major agricultural pest (33). We first report a high-quality assembly of the whole *C. typhae* genome based on a hybrid sequencing approach. We found that it contains 27 typical bracovirus proviral segments as well as an unexpectedly large number of circle sequences (at least 6) that were duplicated through HIM-mediated integration. We then show that integration of all HIM-containing circles occurs systematically at high levels during parasitism in all *S. nonagrioides* tissues, not only in hemocytes as described for *M. sexta* parasitized by *C. congregata* (30). We further demonstrate that integration is not required for the persistence of circles during parasitism, as the quantity of nonintegrated circles is similar to that of most integrated circles in all host tissues 7 days after parasitism. Interestingly, high levels of bracovirus integration were also detected in the host's genome even when parasitism failed.

RESULTS

Assembly and annotation of the *C. typhae* genome. The genome of *C. typhae* was sequenced at about 45× depth with short paired-end reads (Illumina) and 350× depth with long reads (Oxford Nanopore Technologies [ONT]) (see Table S1 in the supplemental material). The size of the preliminary short-read assembly was 183 Mb. In agreement with this, the size of the hybrid (short- and long-read) assembly was 186,662,351 bp (see Table S2). This assembly was made of only 72 scaffolds and had an N_{50} value of 6.81 Mb (see Table S2). It is noteworthy that the assembly nearly reached the chromosome scale with a mean of 7.2 scaffolds per chromosome, since *C. typhae* has 10 chromosomes per haploid genome (34).

The completeness of the assembly was assessed using Benchmarking Universal Single-Copy Orthologs (BUSCO) (35). The results revealed that 1,639 (98.9%) of 1,658 conserved insect genes were present in the final assembly (see Table S3). Assembly visualization by Blobtools (36) using taxon-annotated GC-coverage plots showed a majority assignment to the *Polydnaviridae* family (131 Mb), which is due to the presence of bracovirus sequences dispersed in the wasp genome (3); the majority of large scaffolds were identified as containing a bracovirus sequence (see Fig. S1). Our automatic annotation revealed that 58.6% of the *C. typhae* genome is made of TEs. The most numerous TEs are large retrotransposon derivative (LARD) and terminal inverted repeat (TIR) elements, which represent 35 and 28% of the classified TEs, respectively (see Fig. S2). We automatically annotated a total of 8,591 genes in the genome of *C. typhae* (see Table S4). More than 90% of the predicted genes had over 50% of their exons supported by transcriptome sequencing (RNA-seq) from the species *Cotesia vestalis*, the closest species for which RNA-seq data were available. Genes exhibited a mean of 5 exons per transcript (see Table S4). The joint functional annotation procedure with InterProScan (37) and BLASTP (38) enabled us to annotate 6,488 gene models (75.5%). We were also able to transfer 781 *Cotesia congregata* manually curated genes (3) into the new annotation. Of note, the number of annotated genes is lower than that of *C. congregata* (~14,000 genes, among which ~12,000 were validated by *C. congregata* RNA-seq data), probably in part because of the divergence between *C. typhae* and *C. vestalis* RNA sequences used for annotation.

Annotation of *C. typhae* bracovirus proviral segments. In order to annotate the proviral segments of *C. typhae*, we used the 26 segments of *Cotesia sesamiae* and the

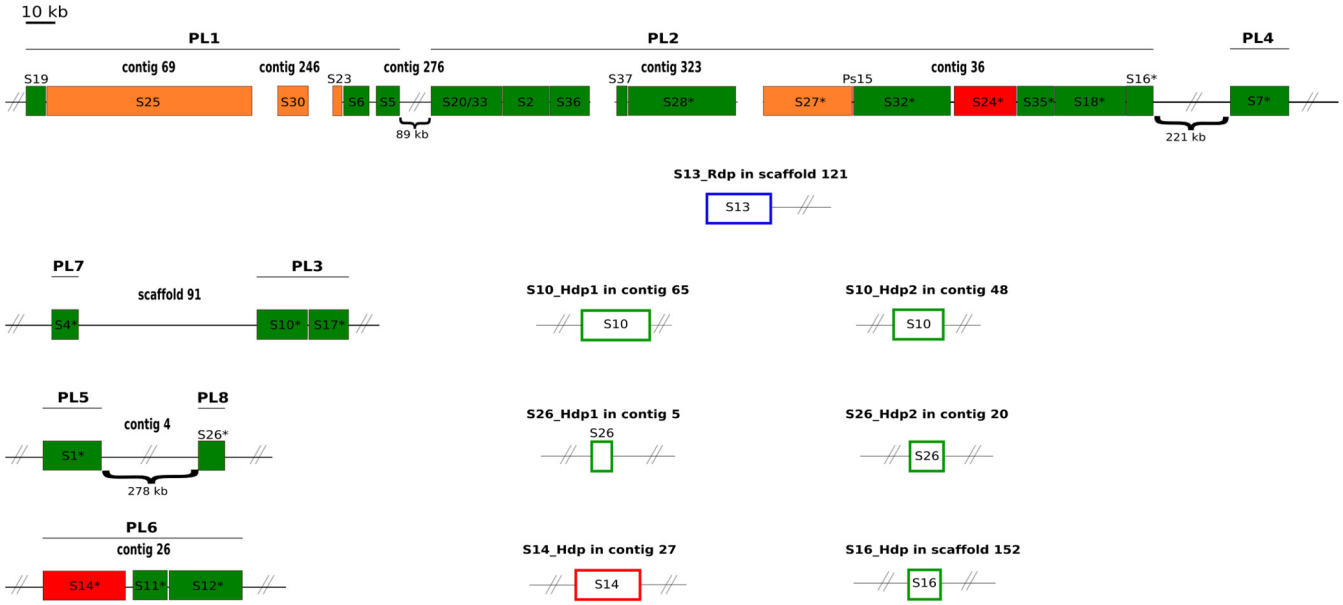


FIG 1 Map of CtBV proviral segments. Proviral segments are represented by filled rectangles. Segments duplicated after circularization are empty. Asterisks indicate HIM-bearing circles found to be integrated into the *S. nonagrioides* genome, corresponding precisely to all segments originating from the RU2.3 part of the macrolocus and isolated loci (PL3, PL4, PL5, PL6, PL7, and PL8). Each contig or scaffold in which the segments are located is indicated, and lines indicate segments that belong to the same PL. The size of the segments and the spaces between them are shown to scale, unless hash marks are present. The colors represent the quality of the annotation. Green indicates that we delimited both extremities of the segments with confidence (DRJs in proviral segments or J1 and J2 motifs in HIM-mediated duplications). Orange and red indicate that one or both extremities (see Table S5) have to be taken with caution. In the case of the orange ones, the contig was too short to identify the extremity, whereas in the case of red ones, the extremity was long enough but we were not able to find the motif. Although they are shown in green, the DRJs of S37 and S26 are truncated, probably due to sequencing or assembly issues. Blue indicates the segment duplicated after circularization by other means than HIM. In this case, there is no J1 and J2 motifs at the extremities, nor DRJ.

10 segments specific to *C. congregata* as queries to perform similarity searches for the *C. typhae* genome. Twenty-seven of a total of 38 segments described in *C. congregata* and/or *C. sesamiae* were clearly identified in *C. typhae* (Fig. 1; also see Table S5). Among the 11 segments not found in *C. typhae*, 9 are specific to *C. congregata*, which means that they are not present in *C. sesamiae* either. The segment S13, which was previously found in both *C. sesamiae* and *C. congregata*, is missing in *C. typhae*. As in *C. sesamiae*, S20 and S33 are fused to form S20/33. We found 3 segments (S10, S11, and S19) that are present in *C. congregata* but have not been found so far in *C. sesamiae*. In total, we annotated 27 proviral segments in *C. typhae* (Fig. 1; also see Table S5).

As expected, the synteny of the segments described in other species is conserved in *C. typhae* (3, 20). As described in previous studies on *Glyptapanteles* and *Cotesia* species, we found a macrolocus, here gathering 18 segments, divided into two PL, PL1 and PL2 (17, 20). The 9 other segments are dispersed across six dispersed loci, from PL3 to PL8. We found that PL4 and the macrolocus are on the same scaffold, consistent with their localization on the same chromosome in *C. congregata* (3). We also found that PL7 and PL3 on one hand and PL5 and PL8 on the other hand are on the same chromosome as in *C. congregata* (3). As for *C. sesamiae*, we did not identify PL10 in *C. typhae*, suggesting that this PL is specific to *C. congregata* and is recent. While PL9 is present in both *C. congregata* (2 proviral segments) and *C. sesamiae* (1 proviral segment), we did not find it in *C. typhae*.

Proviral segment characteristics and HIM identification. HIMs were previously described in 12 of 36 segments in *C. congregata* (30). These HIMs were used here as queries to perform BLASTN searches for *C. typhae* segments (see Data File S2 in the supplemental material). HIMs were found at their expected homologous loci in *C. typhae* for all except 1 segment, S15. The lack of HIM in S15 is likely due to the fact that this segment is undergoing degradation in *C. typhae*. Indeed, in contrast to *C. congregata*, in which S15 is 8,700 bp and contains 5 genes, this segment is residual in both *C.*

sesamiae and *C. typhae*, being 700 and 300 bp, respectively, and containing no gene (see below). In a second step of the analysis, we aligned the 11 HIMs identified in *C. typhae* against the 5 other *C. typhae* segments for which we found integrations in *S. nonagrioides* (S18, S24, S27, S28, and S32; see below). We were able to find HIM in all of these additional segments (see Data File S2). To note, our annotation of HIMs in *C. typhae* segments allowed us to refine the boundaries of the HIM for *C. congregata* S11 (see Data File S2).

Interestingly, our annotation identified seven other bracoviral sequences dispersed in the wasp genome. They share high levels of similarity with viral sequences but do not follow the organization of proviral segments with DRJs at their extremities. Without DRJs at their extremities, these segments cannot form circles and are thus not functional (see "Persistence of nonintegrated circles 7 days postoviposition" for more details). Five of these segments are clearly flanked by the J1 and J2 motifs, which normally lie within the HIM, itself located internally to the proviral segments (Fig. 1). Another segment has one-half an HIM (containing the J1 motif) at one extremity, but the contig is too short to identify the presence of the second half (containing the J2 motif) at the other extremity. Given that the structure of these 6 segments is identical to that observed after integration in the host genome (26, 30), we concluded that, as observed for 3 segments in *C. sesamiae* (39), these 6 *C. typhae* segments (namely, S10_Hdp1, S10_Hdp2, S14_Hdp, S16_Hdp, S26_Hdp1, and S26_Hdp2) originate from HIM-mediated duplication (Hdp). The 6 Hdp segments show between 97.13% and 98.9% similarity to their parental segment, suggesting that they result from relatively recent duplication events. The structure of the last nonproviral segment is atypical. It is highly similar to *C. sesamiae* segment 13 but it is not flanked by DRJs or HIMs. However, it possesses a single internal DRJ, presumably resulting from circularization of its parental segment via recombination of the 5′ DRJ and the 3′ DRJ. The presence of a large flanking sequence indicates that it is present as inserted into the wasp genome and not as a circle or an intermediate amplification form. Thus, we conclude that this segment is a rearranged duplication (Rdp) of S13 (S13_Rdp) that, in contrast to the duplications described above, was not mediated by HIM. To note, we were not able to find the parental segment of S13_Rdp. An explanation might be that S13 was lost after being duplicated in *C. typhae*. A second more plausible explanation might be that S13 is actually present in *C. typhae* but has not been sequenced/assembled. In this regard, according to the synteny of segments in other *Cotesia* species, S13 was expected to lie between S36 and S37 but S36 and S37 lie at the extremity of 2 different contigs (Fig. 1). In addition, sequencing depth data also suggest the presence of S13 in *C. typhae* (see "Persistence of nonintegrated circles 7 days postoviposition"). In this case, *C. typhae* would have 28 proviral segments in total. It is also noteworthy that we identified 3 other segments that we considered potentially resulting from assembly errors. These segments are highly similar to segments S1, S14, and S20/33 and thus could be real duplications of these segments. However, the contigs on which they lie (contig_14, contig_294, and contig_143) are short and do not contain any other wasp sequence (i.e., the segments are partial and not flanked by any other wasp sequence). Therefore, we decided not to include them in the annotation.

Annotation of HIM-mediated duplications in other *Cotesia* species. The finding of 6 HIM-mediated duplications of bracoviral segments was striking, given the absence of such duplications in high-quality genomes of *M. demolitor* and *C. congregata* (3, 26). To assess whether this feature is specific to *C. typhae*, we searched for HIM-mediated duplications in all other available *Cotesia* genomes (*C. sesamiae*, *C. flavipes*, *Cotesia rubecula*, *C. vestalis*, and *Cotesia glomerata*). The highly fragmented nature of these additional genomes prevented us from reaching a high level of confidence in the annotation of all segments. Therefore, the results of this search should be considered preliminary. Of the 6 HIM-mediated duplications, we were able to investigate whether they are shared by other *Cotesia* species for 5 of them (S16_Hdp, S10_Hdp1, S10_Hdp2, S26_Hdp1, and S26_Hdp2). Indeed, our approach relies on both J1 and J2

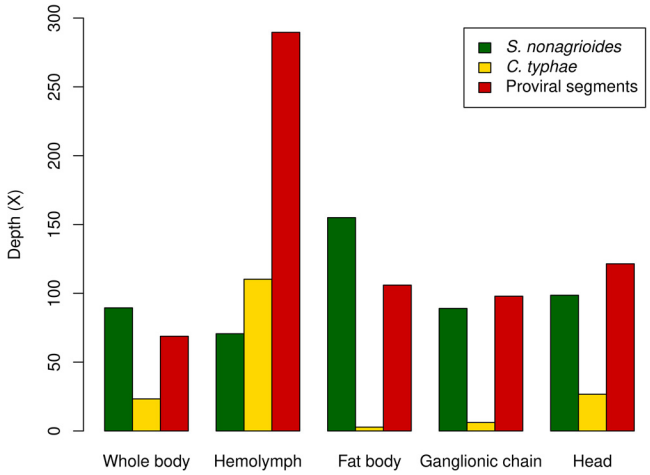


FIG 2 Average sequencing depths in the 5 samples. Green, yellow, and red indicate the average sequencing depths over the whole genome of *S. nonagrioides*, the whole genome of *C. typhae*, and the 27 *C. typhae* proviral segments, respectively.

being clearly annotated. We identified 4 segments and 1 segment orthologous to these 5 duplications in *C. sesamiae* and *C. flavipes*, respectively. More precisely, *C. sesamiae* shares S10_Hdp1, S10_Hdp2, S26_Hdp1, and S26_Hdp2, whereas *C. flavipes* shares only S26_Hdp1. *C. typhae*, *C. sesamiae*, and *C. flavipes* form a monophyletic clade sister to the three other *Cotesia* species included in our search (3). In this clade, *C. typhae* is more closely related to *C. sesamiae* than to *C. flavipes*. These phylogenetic relationships imply that segment S26 underwent a first HIM-mediated duplication in the ancestor of the three species. Regarding the duplication of S26 and the two of S10, they occurred prior to the split between *C. typhae* and *C. sesamiae* and may even be older, as we cannot draw a conclusion about their absence in the fragmented genome of *C. flavipes*. We were not able to find any orthologous HIM-mediated duplications in the other *Cotesia* wasps, and we could identify only 2 candidate *de novo* HIM-mediated duplications, both in *C. sesamiae*. Those involve the parental segments S11 and S18.

Sequencing depth and coverage of the genomes of *C. typhae* and *S. nonagrioides*. We assessed the amount of host versus parasitoid DNA we sequenced in the 5 samples of parasitized *S. nonagrioides* (heads, hemocytes, fat body, ganglionic chain, and whole body) by separately mapping trimmed reads to the genomes of *S. nonagrioides* and *C. typhae*. We obtained a total of 335 million to 595 million trimmed reads, depending on the sample, which covered 97.9% to 99.3% of the 1,021-Mbp *S. nonagrioides* genome. The average sequencing depth along the *S. nonagrioides* genome varied between 71× and 155×, depending on the sample (Fig. 2). The percentage of reads mapping to the *S. nonagrioides* genome varied from 73.05% in the hemocytes to 91.94% in the fat body. Mirroring this variation, between 20.89% and 0.31% of the reads mapped to the *C. typhae* genome in the hemocytes and in the fat body, respectively. Thus, the vast majority of reads (92.57% to 93.94%) mapped either onto the genome of *S. nonagrioides* or onto that of *C. typhae*. The proportion of the *C. typhae* genome covered by the reads was high (94% to 99%) in 4 of the samples, while it dropped to 41% in the fat body. Importantly, the average coverage was higher on proviral segments (68.8× to 289.6×, depending on the sample) than on the rest of the genome (2.8× to 110.2×, depending on the sample) in all samples (Fig. 2). This is consistent with the presence of a greater proportion of integrated and/or nonintegrated wasp bracoviral circles versus other wasp genomic regions in our DNA extracts.

HIM-mediated integration of *C. typhae* bracovirus DNA circles into the *S. nonagrioides* genome. To identify and quantify integrations of *C. typhae* bracovirus (CtBV) DNA circles into the *S. nonagrioides* genome, we searched for chimeric reads for which a region aligns on a CtBV proviral segment and the other region aligns on the caterpillar genome. We identified chimeric reads mapping to all 16 *C. typhae* proviral

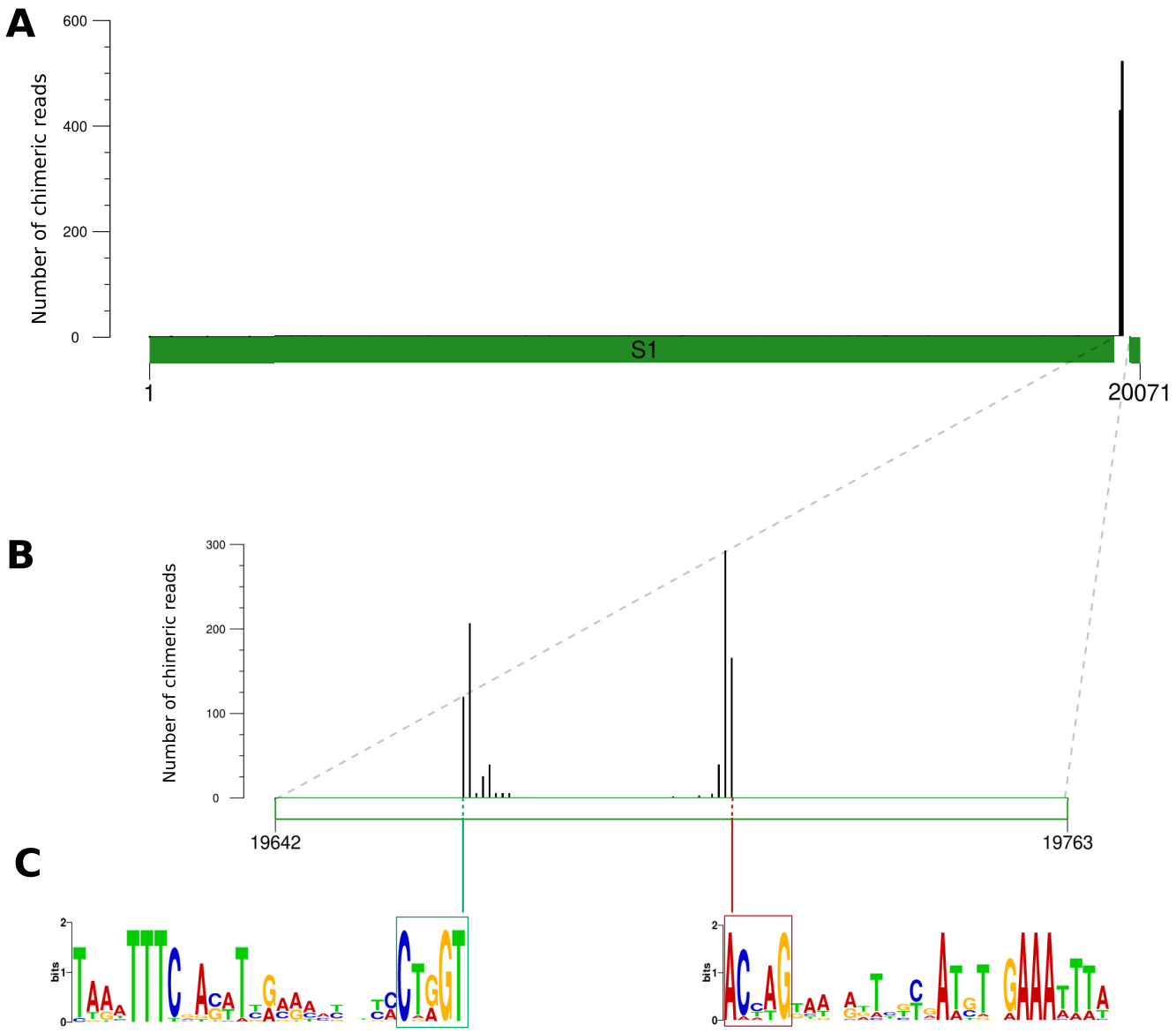


FIG 3 Map of chimeric reads indicating HIM-mediated chromosomal integration of segment 1. (A) Number of chimeric reads along segment 1 in hemocytes, oriented from the 5′ DRJ to the 3′ DRJ. The white portion represents the HIM (not to scale) near the 3′ DRJ. (B) Magnification of the 121-bp HIM, showing two regions with many chimeric reads, called J2 (left) and J1 (right). (C) Sequence logo of J2 and J1 generated with weblogo.berkeley.edu, using an alignment of the HIMs of the 16 segments that integrated into the *S. nonagrioides* genome. For J2, we used the 30 bp upstream from the minimum position at which we observed >2 chimeric reads; for J1, we used the 30 bp downstream from the maximum position at which we observed >2 chimeric reads. The highly conserved motif J1 is framed in red and J2 in green.

segments containing a HIM in all 5 DNA samples. The total number of chimeric reads (excluding PCR duplicates) on these segments varied from 4 on segment 16 in the head to 947 on segment 1 in the hemolymph. Importantly, between 87.5% and 100% of chimeric reads mapping to the 16 HIM-containing segments were located in HIMs (see Fig. S3). In fact, the vast majority of chimeric reads mapped to two short regions located within HIMs and spaced by 41 to 73 bp (see an example of segment 1 in Fig. 3; also see Fig. S3). Alignment of all 16 HIMs allowed us to identify a conserved motif under each of these regions that corresponded to the J1 and J2 junctions previously characterized in *C. congregata* and *M. demolitor* (Fig. 3C) (26, 30). Overall, the pattern we observed confirms that HIMs split during circle integration, that the 41- to 73-bp region between J1 and J2 is lost, and that J1 and J2 end up at the extremities of the linearized circle once in the host. Our results also show that the 16 HIM-containing circles

integrate in all host tissues surveyed and that most integrations of these circles into the *S. nonagrioides* genome are mediated by HIMs.

Potential role of microhomologies in CtBV circle integration. Interestingly, wasp-moth junctions in chimeric reads do not all map at the same position within the J1 or J2 motifs. Rather, they are distributed over 2- to 12 bp-long regions, depending on the segments and samples (Fig. 3; also see Fig. S3). This pattern could be due to biological variation in the position of the breakpoint within the J1 and J2 motifs. It could also reflect imprecision in our mapping of wasp-host junctions caused by the presence of microhomologies between CtBV and host sequences at the junction. Indeed, at the CtBV-host junction, there is a 1 in 4 chance that the base following the junction position in the CtBV circle would be the same as that following the junction in the moth genome (see Fig. S4a). In our approach, the position of the CtBV-host junction corresponds to that of the BLASTN alignment end coordinate on the wasp, regardless of the presence of any overlap (see Fig. S4b). There is thus a 1 in 4 chance for the true position of the CtBV-host junction to be shifted by 1 bp for an overlap of 1 bp. For an overlap of 2 bp, there is a 1 in 16 chance that the wasp-host junction would be shifted by 2 bp. Interestingly, when there is no overlap, we observed that the CtBV-host junction almost always occurs at the same exact position in J1 and J2 for all segments, with some very rare chimeric reads shifted by 1 bp. Thus, it appears that junctions devoid of microhomology involve CtBV circles that all underwent a double-strand break at the same exact position, as expected under the hypothesis that bracovirus circle integration is mediated by a site-specific recombinase (30). Among junctions with microhomology, we found more chimeric reads with shifted CtBV-host junctions than expected by chance, suggesting that the imprecision of the breakpoint may be at least partly biological.

To further assess whether bracovirus-moth microhomologies at the junctions may somehow foster integration of DNA circles, we compared the expected numbers of chimeric reads for each microhomology length to the observed values (see Materials and Methods) (Fig. 4). We did this for chimeric reads falling specifically in J1 or J2 and for chimeric reads falling outside J1 and J2 but still in the HIM regions. Regarding chimeric reads falling in J1 or J2, we found that the number of observed microhomology lengths was close to that expected by chance for microhomology lengths of $>$3 bp. Thus, while certainly affecting the precision of our junction-mapping pipeline, these microhomologies are unlikely to have biological underpinnings. In contrast, the numbers of 0-bp, 1-bp, and 2-bp microhomologies differed markedly from what is expected by chance, with the observed 0-bp microhomologies being 1.8 times less numerous and 1-bp and 2-bp microhomologies being 1.5 times more numerous than expected by chance (Fig. 4a). Like 3-bp-long microhomologies, 1-bp- and 2-bp-long microhomologies affect the precision of our junction-mapping pipeline. However, their overrepresentation indicates that they likely have biological underpinnings. For chimeric reads falling outside J1 and J2 but still in the HIMs, we observed a major underrepresentation of 0- to 2-bp microhomologies (65 versus 478 reads), mirroring a major overrepresentation of 4- to 13-bp microhomologies (478 versus 105 reads). This suggests that, when the breakpoint is located further away from the canonical positions of J1 and J2, the presence of microhomology between CtBV and moth sequences may be crucial for successful integration.

Few integrations of CtBV DNA circles outside HIMs. Our search for chimeric reads also yielded a number of reads mapping outside HIMs in HIM-containing segments, as well as reads mapping to segments that did not contain a HIM. The number of such reads was low. In HIM-containing segments, the number varied from 0 (for 11 segments in all or some tissues, depending on the segment) to 23 (for segment 1 in the hemolymph). In segments devoid of HIM, this number varied from 1 (for multiple segments in multiple tissues) to 11 (for segment 20/33 in the fat body). In contrast to chimeric reads mapping to HIMs, which are clustered in two regions corresponding to J1 and J2 motifs, reads mapping outside HIMs are dispersed over the circles, with no circle position outside HIMs being mapped by more than one bracovirus-host junction,

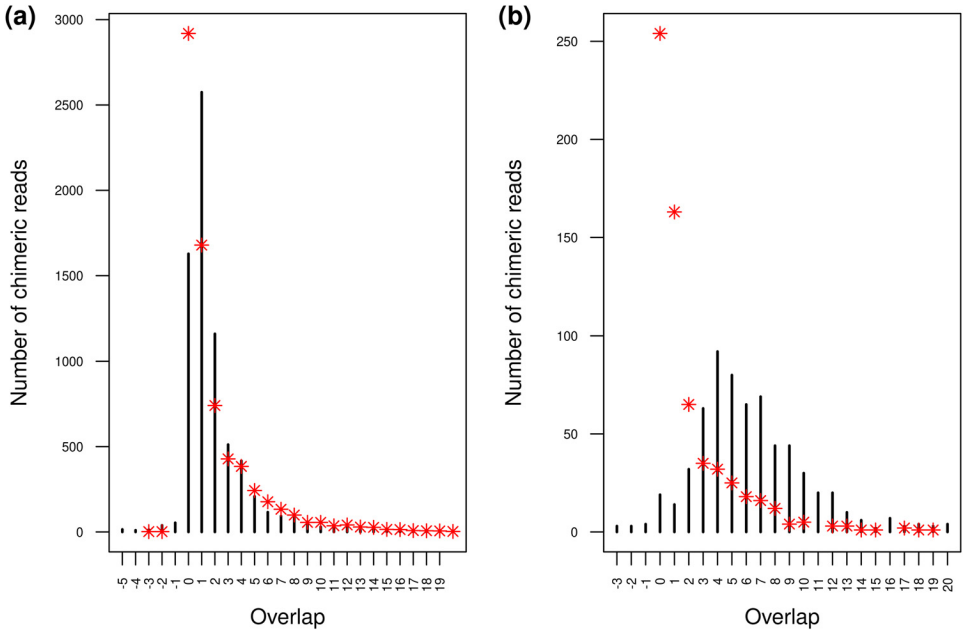


FIG 4 Distribution of microhomology lengths at wasp-host junctions in chimeric reads. Black bars correspond to the numbers of observed chimeric reads for each microhomology length. Red asterisks correspond to the expected numbers of chimeric reads for each microhomology length. (a) Distribution of microhomology lengths for CtBV-host junctions mapped in J1 or J2. (b) Distribution of microhomology lengths for CtBV-host junctions mapped within HIM but outside J1 or J2.

except for two junctions covered by 2 reads each. This pattern could suggest that, in addition to HIM-mediated integration, circles could integrate into the *S. nonagrioides* genome through other mechanisms, possibly involving host DNA repair pathways, as suggested by Wang et al. (31) for *Diadegma semiclausum* ichnovirus (DsIV). In agreement with this, we found that the number of chimeras mapping outside HIMs in bracovirus circles was always higher than expected, given the number of wasp-host chimeras involving exons of wasp BUSCO genes (see Materials and Methods). However, given the small number of such non-HIM, bracovirus circle integrations, this possible alternative circle integration mechanism is unlikely to play a significant role in parasitism for *C. typhae*.

Gene content of integrated segments. To assess whether circle integration is associated with circle gene content, we compared gene family content for integrated versus nonintegrated circles (Fig. 5). This comparison was done for all gene families with known predicted domains and more than 2 genes. Overall, it appears that the integration of a circle is associated with its content in gene families (Fisher exact test, $P < 0.01$). Three gene families present on at least 3 segments seem to explain this observation, i.e., viral ankyrin (VANK), serine-rich, and protein tyrosine phosphatase (PTP). These gene families contain 5, 4, and 24 genes distributed over 4, 3, and 7 segments, respectively, all found integrated in the *S. nonagrioides* genome. This observation suggests that integration of these three gene families is important for parasitism success.

Quantification of integrated bracovirus circles in the *S. nonagrioides* genome. We then set out to quantify the number of integrations of CtBV circles that occurred during parasitism of *S. nonagrioides* larvae in our experiment. Parsimoniously, we considered only chimeric reads that fell in the J1 and J2 motifs of the HIMs, that is, between 730 and 3,126 chimeric reads per sample. We found that the vast majority of integrations in the moth's genome (6,784 [98%] of 6,940 integration events [IEs]) were supported by 1 chimeric read only. Among IEs supported by more than 1 read (2%), 3 were supported by 3 chimeric reads and the rest by 2 chimeric reads. This pattern indicates that most chimeric reads correspond to independent IEs. Thus, among the host cells we sequenced, almost no cells harbored a shared IE that would originate from a

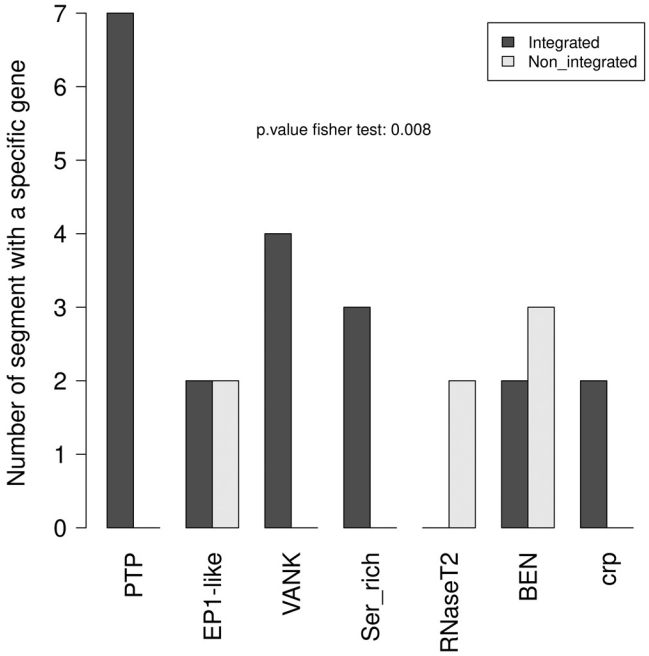


FIG 5 Integration capacity of segments containing ≥1 gene belonging to seven gene families: PTP (protein tyrosine phosphatase), EP1-like (early parasitism-specific protein 1), VANK (viral ankyrin), Ser_rich, RNaseT2, BEN (BEN-domain proteins), and crp (cysteine-rich proteins). Segments containing genes belonging to several gene families are counted for each family. Black bars correspond to segments that integrate into the genome of *S. nonagrioides*, while white bars correspond to segments that do not integrate.

cell division. Figure 6 shows the number of IEs we inferred per segment and per sample by counting each integration position only once. This led to 714 to 3,064 IEs, depending on the samples and segments. In order to be able to compare the number of IEs for each segment between samples, we turned absolute numbers of IEs into relative numbers normalized to 1 million reads mapping to the genome of *S. nonagrioides* (see Table S6). Considering all samples together, S1 is the segment with the most integrations, with 6.64 IEs per million reads mapping to the host (*S. nonagrioides*) (IPMH), followed by S7, with 3.38 IPMH and then by S26 with 1.76 IPMH (Fig. 6a). All of the other segments have less than 1.45 IPMH. For all segments, the hemocyte sample is the sample with the most chimeric reads (Fig. 6a). In total, we infer about 12.5 IPMH in the hemocytes, about 3 IPMH in the ganglionic chain and the head, and about 2 in the whole body and the fat body (Fig. 6b). Given the haploid genome size of *S. nonagrioides* (1,021 Mpb [40]), the read size (150 bp) and the number of IPMH, we estimate the average number of IEs per genome as follows: (IPMH/read size) × genome size (in mega-base pairs). This yielded an average of 85, 25, and 12 IEs per genome in the hemocytes, ganglionic chain, and fat body, respectively.

Quantification of HIM-containing CtBV circles in their integrated versus circularized forms. We assessed how many of the HIM-containing CtBV circles we sequenced were integrated into the *S. nonagrioides* genome versus how many there were in total, regardless of their form (circular or integrated). For this, we compared the numbers of IEs (as an approximation of the number of sequenced integrated circles) to the average circle sequencing depths (as an approximation of the total quantity of CtBV) for each circle in each sample. We found that the numbers of IEs per circle were strongly correlated with sequencing depths for all samples (Spearman rho of 0.7 to 0.9; $P < 0.01$) (Fig. 7). This indicates that the number of integrated circles depends to a relatively large extent on the total amount of circles that are injected by wasps into their host. Interestingly, we also found that the ratio of any forms to integrated circles varied depending on the circles, with these variations being similar among samples. For

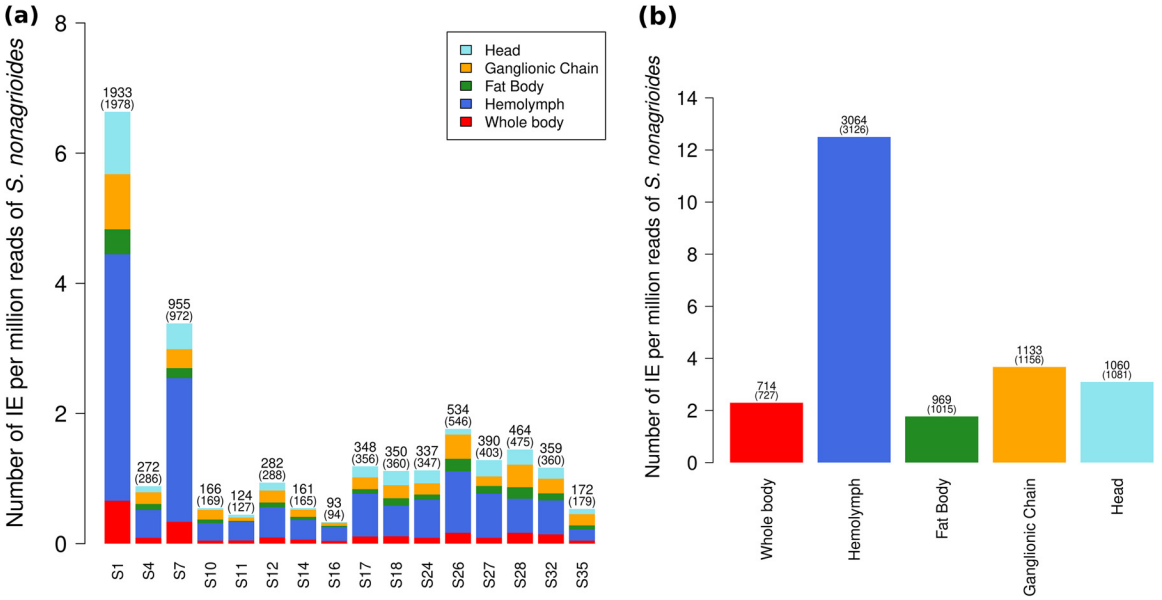


FIG 6 Number of IEs for each segment and sample. Absolute numbers of IEs and of chimeric reads (in parentheses) are shown at the top of each bar. (a) Barplot comparing the numbers of IEs for each segment. (b) Barplot comparing the total numbers of IEs of all segments in each sample.

example, circle 1 is characterized by the lowest ratio in 3 of 5 samples, while circle 16 has the highest ratio in all samples (Fig. 7). This likely reflects variation in the efficiency of integration among circles. In addition, it suggests that a significant part of the circles remain nonintegrated, at least for the circles with a high ratio. Thus, in addition to being determined by the overall quantity of circles injected by the wasp, the propensity of a circle to integrate depends on other factors, possibly the binding affinity of the integrase/recombinase to HIM sequences, which may have more or less diverged from the optimal HIM.

Persistence of nonintegrated circles 7 days postoviposition. We then used the average sequencing depth per circle to compare the quantity of HIM-containing circles (or circles that integrate into host genomes) to the quantity of other circles, which do not integrate into host genomes (Fig. 8; also see Fig. S5). The sequencing depth of segment 15 is close to the average depth on the *C. typhae* genome, suggesting that this segment is present only in the form of proviral segments in *C. typhae* cells that are present at different levels in the different tissues. This is in accordance with the annotation of this segment, for which we did not find any DRJs, suggesting that segment 15 is not able to form DNA circles and should thus be considered a pseudosegment (Ps15 in Fig. 1). All other segments display higher coverage than the *C. typhae* genome, suggesting that, in addition to their proviral form present in *C. typhae* cells, they are present in the circular form and/or in the integrated form. We found that, with the exception of circles 1 and 7, which are characterized by very high sequencing depths and large numbers of IEs, the ranges of sequencing depths were similar between HIM-containing circles (from 314 to 946) and other circles (from 311 to 937, leaving Ps15 aside) (Fig. 8; also see Fig. S5). Thus, the number of integrating circles found in host tissues 7 days postparasitism is similar to that of nonintegrating circles.

In addition, it is worth noting that the sequencing depth of S13_Rdp in the parasitized caterpillars was in the range of that of the functional segments (Fig. 8). This observation supports the presence of the parental S13 in *C. typhae*, which we were not able to assemble. The high sequencing depth of S13_Rdp is likely due to the fact that reads that would map onto S13 if it were in our assembly instead map to the very similar S13_Rdp. Importantly, the sequencing depths of all other Rdp and Hdp segments were in the range of the sequencing depths of the other regions of the *C. typhae*

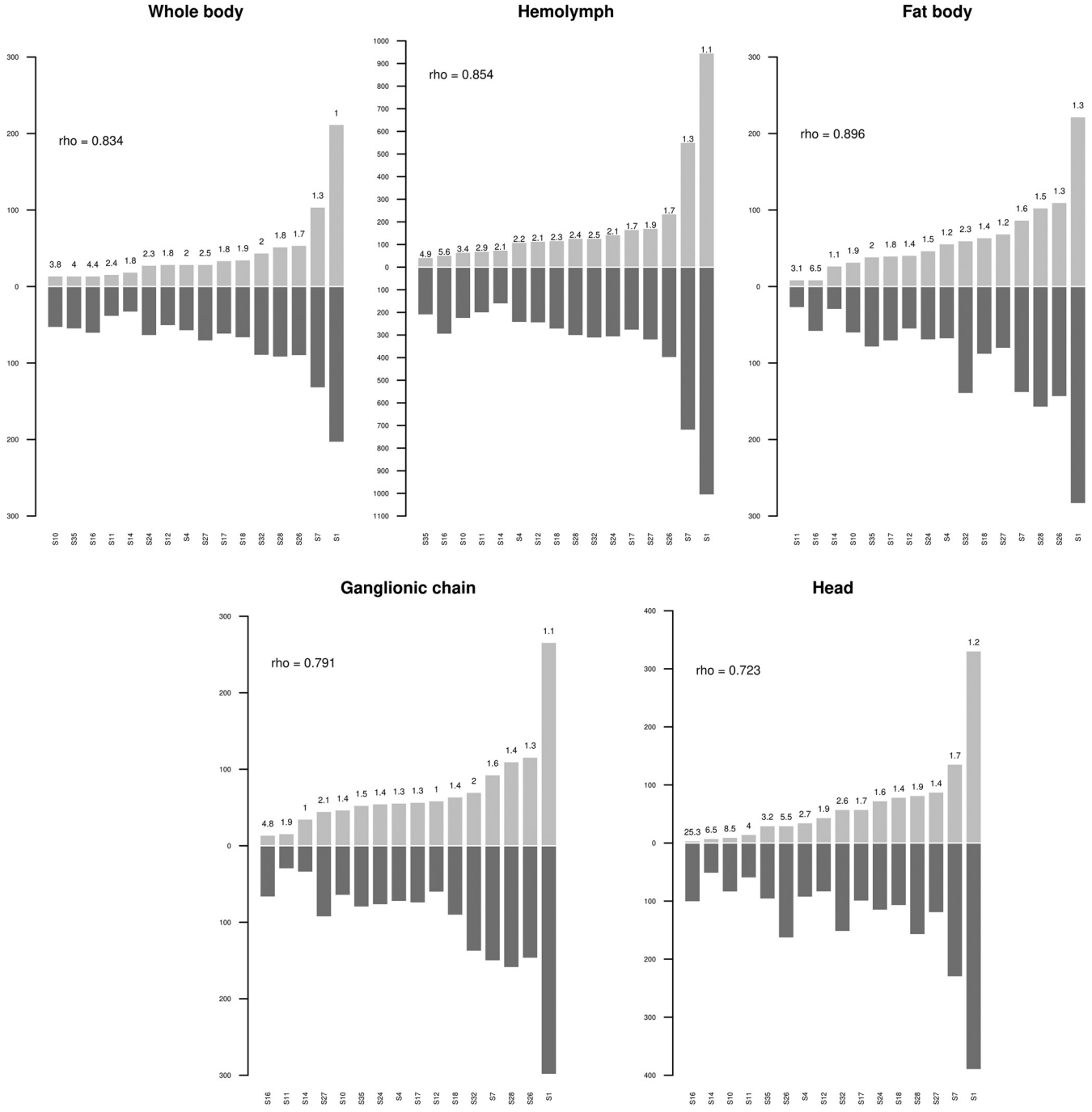


FIG 7 Histograms showing the number of chimeric reads and the sequencing depth for each HIM-containing segment. Light gray bars show the number of chimeric reads, while dark gray bars show the sequencing depth. The ratio of sequencing depth to chimeric reads is indicated at the top of each light gray bar. The Spearman rho values indicate the correlation between sequencing depth and the number of chimeric reads for each sample.

genome, supporting the idea that they are present only as proviral segments in teratocytes and other residual wasp cells. This is in line with the nonfunctional nature of these duplicated segments, which are not expected to generate circles (Fig. 8).

Distribution of wasp segments throughout the genome of *S. nonagrioides*. We investigated whether DNA circles integrate randomly along the *S. nonagrioides* genome. To do so, we split the genome into 100,000-bp windows and assessed whether some windows were subjected to more integrations than expected by chance. We chose to not have any windows with a mixture of contigs, which led to 2,553 windows smaller than 100,000 bp that we eliminated from our analysis. The remaining 9,121

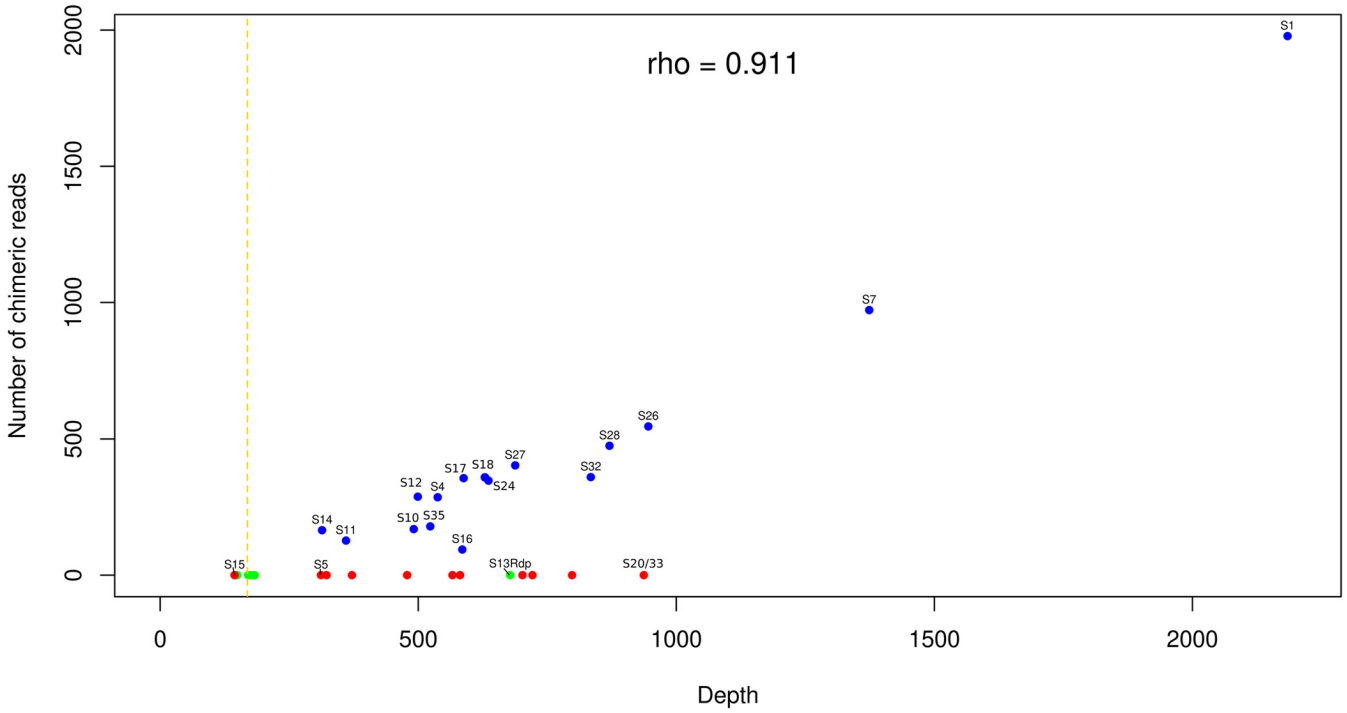


FIG 8 Plot of the sequencing depth versus the number of chimeric reads for each of the CtBV segments. Sequencing depths and numbers of chimeric reads were summed for all samples. The same plots are shown for each sample in Fig. S5 in the supplemental material. Blue dots represent proviral segments that do integrate into the *S. nonagrioides* genome, and red dots represent proviral segments that do not integrate. Green dots represent duplicated segments, i.e., Rdp and Hdp segments. The identification numbers of the segments are shown near each blue dot. For red dots, only the identification numbers S15, S5, and S20/33 are shown; for green dots, only S13_Rdp is indicated. The yellow dashed line shows the average depth on the *C. typhae* genome when all samples are summed. The Spearman rho value indicates the correlation between sequencing depth and the number of IEs for segments that do integrate into the *S. nonagrioides* genome.

windows of 100,000 bp covered 89.3% of the genome and bore 88.3% to 90.7% of the IEs, depending on the samples. Under the null hypothesis that segments integrate at random in the *S. nonagrioides* genome, we assumed that the number of windows bearing integrations followed a Poisson distribution. We compared observed versus expected numbers of IEs under our null hypothesis for each sample separately as well as for a pool of all samples. Observed distributions departed significantly from the expected ones in all 6 cases ($P < 0.001$) (see Fig. S6). In the case of the pooled data set, we observed an excess of windows with no IEs or $\geq$3 IEs and we observed a deficit of windows with 1 or 2 IEs. Under a Poisson distribution, we did not expect any windows with more than 7 IEs. However, some windows have up to 23 IEs. This result suggests that segments do not integrate entirely randomly into the genome, as observed for *C. congregata* bracovirus (CcBV) in the genome of *M. sexta* (30). Interestingly, two windows had $\geq$2 IEs in all 5 samples, with a maximum of 8 IEs. These IEs come from 9 and 11 segments; therefore, the overrepresentation of IEs in these two windows is not due to 1 specific segment targeting them. We then tested whether several factors, including variation of sequencing depth, GC content, TE content, or gene content, could explain the distribution of IEs along the *S. nonagrioides* genome. We found that none of these variables was strongly correlated with the IE density (data not shown).

DISCUSSION

HIM-mediated duplications of CtBV in the wasp genome. In this study, we assembled a high-quality genome for the braconid wasp *C. typhae*, which allowed us to annotate 27 bracovirus proviral segments plus 7 other duplicated segments, among which at least 6 resulted from HIM-mediated duplications. Such a high number of HIM-mediated duplications is noteworthy, given that none was found in the high-quality

assembly of *C. congregata*. These duplications imply that, after circularization, a segment can integrate into the genome of wasp germ line cells, as suggested by Serbielle et al. (39), who identified 3 HIM-mediated duplications in *C. sesamiae*. Several scenarios could explain such accidental integration. First, HIM-mediated duplication in the wasp could occur through ectopic circularization of a segment in germ cells associated with ectopic expression of wasp factors required for integration in these cells. This scenario appears unlikely, because it would imply that the entire set of complex processes leading to particle production would accidentally occur in the germ line. Second, circle-containing bracoviral particles could sometimes accidentally enter germ cells of the wasp that produced them. This scenario also seems unlikely, because virus particles are released in the calyx lumen, which is located in the posterior part of the ovaries, whereas germ line cells are located in the upper part, in the ovarioles. Third, circle-containing bracoviral particles could enter another wasp individual during accidental oviposition in this individual. This could occur when a wasp oviposits in an already parasitized host larva containing a high density of wasp embryos, as observed for *Microplitis croceipes* (41). Such behavior could occur more frequently in wasps parasitizing stem borer hosts, because these wasps follow the galleries made in the plant stem by their host, instead of ovipositing from outside the plant (42). The aggressive response of stem borer hosts may indeed impose a higher pressure on female wasps in the confined, plant stem environment, which may be more conducive to behaviors such as oviposition into host larvae that have already been parasitized by another wasp. In this respect, it is noteworthy that we found HIM-mediated duplications only in *Cotesia* wasps parasitizing stem borers (*C. typhae*, *C. sesamiae*, and *C. flavipes*). No such duplications were found in the three other *Cotesia* species (*C. rubecula*, *C. glomerata*, and *C. congregata*), which are known to parasitize lepidopteran hosts dwelling on plant leaves (43). Whether the type of host and its habitat have an impact on the likelihood of HIM-mediated duplications will have to be reappraised when higher-quality genomes are available for other wasps.

Several clues suggest that HIM-mediated duplications may participate in the dynamic evolution of wasp bracoviral segments. Indeed, there are striking similarities in the gene content between segments producing DNA circles in the PL2 region and isolated loci (such as S1 in PL2 and S17 in PL3) in *Cotesia* species, which suggests that dispersed loci may originate from duplications, whether these duplications are mediated by HIM or not (39). However, the generation of a new segment by HIM-mediated duplication is probably rare, since such a segment would need complex genome rearrangements in order to acquire the ability to form DNA circles. Indeed, after HIM-mediated duplication, the segment contains a single DRJ (whereas both the 5′ DRJ and the 3′ DRJ are required for circularization) and none of the regulatory sequences allowing bracovirus DNA amplification, which are located at the extremities of the amplified regions (RUs) outside proviral segments (3, 21). Duplications could also participate in the dynamic evolution of wasp bracoviral segments through gene conversion or other mechanisms.

Chromosomal integration of CtBV in multiple host tissues. This study shows that all 16 HIM-containing circles of CtBV undergo chromosomal integration in *S. nonagrioides* cells during parasitism. Previous studies characterizing chromosomal integration of polydnavirus DNA circles focused only on one tissue type (hemocytes) and/or were limited in terms of the number of circles studied (26, 28–31). Here, we used bulk Illumina sequencing of DNA extracted from hemocytes, fat bodies, ganglionic chains, and heads of parasitized *S. nonagrioides* larvae, which shows that chromosomal integration of DNA circles is not limited to hemocytes and extends to all other surveyed tissues. Interestingly, the *C. typhae* genome has been sequenced deeply (110×) in the hemolymph, which indicates the presence of numerous wasp cells in this tissue. These wasp cells may be teratocytes, which are known to be released from the wasp embryonic membranes into the host when eggs hatch. In *C. congregata*, which is gregarious at the larval stage, like *C. typhae*, the number of teratocytes reaches about 140

teratocytes per wasp embryo (44). Teratocytes play various roles during parasitism, including host immunosuppression, production of antimicrobial peptides, and nutritional functions (45, 46). They undergo physiological and morphological changes during development of the wasp embryos, including an increase in ploidy level (47, 48), which likely explains the high sequencing depth obtained over the *C. typhae* genome in the hemolymph. In the other tissues, however, the sequencing depth for *C. typhae* was much lower (down to 3× in the fat body), indicating that few if any wasp teratocytes were sequenced in those samples. This in turn points to a low level of contamination of other tissues by hemolymph, indicating that the majority of CtBV circle IEs we identified in nonhemolymph tissues are *bona fide* chromosomal integrations in cells constituting those tissues. Although we did not perform replicates for each tissue, our study seemingly indicates that the number of circle IEs is higher in hemocytes than in other tissues, which is in line with the role of hemocytes in immunity (49, 50) and the known effect of circle-borne virulence genes in thwarting host immunity responses (2). This could also indicate that hemocytes are preferentially infected by CtBV, resulting in greater abundance of viral circles in these cells, as already suggested by Beck et al. (51). Given the multiple effects of bracovirus on host physiology, development, and behavior (29), it is likely that integration in a wide range of tissues contributes to parasitism success and is not merely a by-product of the capacity of bracoviral particles to enter many cell types.

Persistence of integrated versus nonintegrated CtBV circles. It was traditionally assumed that integration of polydnavirus circles was beneficial to the wasp because it allowed persistence and expression of these circles throughout the duration of wasp embryonic and larval development, which can last between 7 and 14 days under laboratory conditions, depending on the species considered (26, 30, 52). Here, we estimated that, at 7 days postoviposition, parasitized hosts contain between 12 and 85 integrated circles per haploid genome, depending on the tissue. Most IEs characterized in this study are supported by only 1 chimeric read, indicating that most integrations are specific to one of the *S. nonagrioides* cells we sequenced. At first sight, this may seem unexpected, because an IE occurring early after parasitism may be expected to be shared by many cells at 7 days postparasitism as a result of successive divisions of the original IE-bearing cell. However, given the range of haploid genome sequencing depths (70× to 155×, depending on the samples), we estimated that we sequenced a very small number of *S. nonagrioides* cells (maximum of 35 to 77 cells). Thus, the probability of sequencing 2 cells with the same IE was very low. Therefore, the fact that we find very few IEs supported by more than 1 chimeric read cannot be taken as an indication of limited persistence of integrated circles in a given host cell lineage through successive mitotic divisions. Measuring such persistence would require sequencing the host genome more deeply. However, an interesting observation we made regarding the persistence of circles throughout parasitism is that, with the exception of circles 1 and 7, the 14 other integrating circles are not present in greater quantities than nonintegrating circles, a trend that holds for all tissues (Fig. 8; also see Fig. S5 in the supplemental material). Thus, it appears that similar quantities of integrating and nonintegrating circles can persist over at least 7 days during parasitism. It follows that integration is not a requirement for persistence during at least one-half of the duration of *C. typhae* embryonic development. Interestingly, our data confirm that integrating and nonintegrating circles clearly differ in terms of gene content, with genes such as VANK and PTP being present exclusively on integrating circles (Fig. 5) (30). Further studies are needed to shed light on the role of integration during parasitism (30) and on the link between CtBV gene content and integration.

Mechanism of CtBV integration. Our study confirms that bracovirus DNA circles integrate into the genome of wasp hosts through site-specific recombination involving HIMs (26, 30). As proposed earlier, *vlf-1* and *int-1*, two candidate genes of nudiviral origin encoding an integrase domain (of the phage integrase family also known as tyrosine recombinases) may be involved in chromosomal integration (30). These two proteins are loaded into bracovirus particles (16, 53) and delivered to the host, and they were shown by RNA interference experiments to be involved in circle excision (54). Interestingly, our study of

microhomologies between wasp and moth sequences at bracovirus-moth junctions reveals variation in the mechanism of integration. On one hand, we found that all bracovirus-moth junctions devoid of microhomology, which represent 21% of all junctions uncovered in this study, occur at the exact same positions in the J1 and J2 motifs within the HIM. This may indicate that, for a relatively large fraction of integrations, the occurrence of a double-strand break at a canonical position can be resolved without microhomology. On the other hand, we found an excess of 1-bp and 2-bp microhomologies in the junctions located in J1 and J2 motifs and an excess of 3-bp to 12-bp microhomologies in the junctions located outside J1 and J2 motifs (641 [8%] of 7,746 reads). Thus, our results indicate that integration can occur with or without wasp-moth base pairing, but it is unclear whether microhomology-mediated integrations are generated through the same mechanism as integrations devoid of microhomology (30) or whether they may occur through DNA repair mechanisms (55).

Evolution of HIM in braconid and ichneumonid wasps. Irrespective of whether chromosomal integration of wasp circles involves a single mechanism or multiple mechanisms, it appears that the vast majority of IEs, if not all of them, occur at double-strand breaks generated within HIMs. Our study thus confirms the central role played by HIMs in integration. While these motifs have been found in all Microgastrinae wasps studies so far, they could not be identified in *Chelonus inanitus*, a bracovirus-containing microgastroid wasp belonging to another subfamily (Cheloninae) (30). It was thus proposed that HIMs might have been acquired by the ancestor of Microgastrinae about 54 million years ago, independently and well after the domestication of the nudivirus shared by all microgastroid wasps (30). The recent finding that ichnovirus circles from the ichneumonid wasp *Diadegma semiclausum* undergo chromosomal integration into their host via HIM-like motifs raises the question of the evolutionary link between these motifs in bracoviruses and ichnoviruses. Structurally, ichnovirus and bracovirus HIMs are similarly made of J1 and J2 motifs separated by a stretch of sequence that is deleted upon circle integration. The size of the sequence between J1 and J2 is relatively homogeneous in most bracovirus and ichnovirus segments (33 to 78 bp), although some ichnovirus segments have longer intervening sequences (e.g., DsIV-38 and *Tranosema rostrale* ichnovirus F1, which have 311-bp-long and 1,781-bp-long intervening sequences, respectively). Like bracovirus HIMs, which do not seem to be ubiquitous among microgastroids (i.e., they were not found in *Chelonus inanitus* bracovirus segments), ichnovirus HIMs were not found in all Campopleginae wasps known to harbor an ichnovirus that were searched (31). Indeed, in addition to DsIV, Wang et al. found HIMs in *Tranosema rostrale* and *Hyposoter fugitivus* ichnoviruses but not in *Campoletis sonorensis* ichnovirus (31). We think that three evolutionary scenarios could explain the presence of HIMs in both bracovirus and ichnovirus segments. The first scenario posits that HIMs and other integration factors were present in the ancestor viruses (bracovirus and ichnovirus) and were lost in several wasp lineages. Supporting this scenario, nudivirus HzNV1 is known to integrate into the DNA of cultured cells and to persist during a latent phase both as an integrated form and as an episomal form (56). Nudivirus integration properties might have favored the recurrent domestication of nudiviruses by parasitic wasps (9, 11). However, the mechanism of HzNV1 integration has not yet been characterized. Concerning ichnoviruses, because the ancestor belongs to a virus family that is possibly extinct, nothing is known regarding potential ancient integration properties. The second scenario assumes that integration of DNA circles evolved after viral domestication. It implies that HIMs would have been acquired in Braconidae and Ichneumonidae after viral domestication. This acquisition could have occurred through independent recruitment of recombinase sites and proteins from related viral elements or TEs present in both braconid and ichneumonid wasp genomes. In agreement with this scenario, HIM-like motifs that contain inverted terminal repeats and are involved in site-specific recombination are common in prokaryotes, yeast, and viral genomes and TEs (57). Recombination sites of site-specific recombinases involved in DNA insertions, inversions, or circularizations are typically between 30 and 200 nucleotides in length and consist of two motifs with a partial inverted repeat symmetry, to which the recombinase binds and

which flank a central crossover sequence at which the recombination takes place (58). HIM sites correspond fairly well to that description. In eukaryotes, several examples of recombinases originating from TEs have been reported, such as the RAG1 protein, which is responsible for shuffling immunoglobulin genes in vertebrates (19), and transposases that are involved in the maturation of paramecium nuclei (59). Finally, a third scenario would imply that HIMs were acquired only once, by either Ichneumonidae or Braconidae wasps, and then were transferred between the two polydnaviruses. Such transfer could have been favored by the integration properties of polydnavirus circles and by the fact that some wasps from the two families are known to parasitize the same host species (60, 61). This seems rather unlikely, however, since it is not sufficient to transfer the HIM sequence to provide a functional mechanism, and the recombinase gene needs to be transferred at the same time; however, the latter is not present on a bracovirus circle (it is packaged as a protein in polydnavirus particles). Characterization of the different proteins involved in circle integration and of the integrases/recombinases encoded in parasitoid wasp genomes will probably be helpful to shed further light on the evolutionary history of HIMs and polydnaviruses at large.

Possible long-term impact of polydnavirus integration in wasp hosts. Previous studies uncovered bracovirus circle sequences in the genomes of several species of lepidopterans, indicating that such sequences were horizontally transferred from wasps to lepidopterans at some point during the evolutionary history of these insects (18, 62). Although we did not include host germ line tissues in this study, our finding that bracovirus circles can integrate into host tissues other than hemocytes suggests they may also integrate into host germ line cells. In this context, it is remarkable that a fairly large number of bracovirus integrations were found in the whole-body sample (Fig. 6), i.e., a host larva in which no wasp larvae were present 7 days postparasitism. The absence of wasp embryos in this larva and the 5 larvae that we did not sequence could be due either to active resistance of the host, which would have prevented the development of these embryos, or to the fact that the wasp injected venom but no eggs into these larvae (63). Relatively high sequencing depths over the entire *C. typhae* genome in the sequenced larva (Fig. 2) are in agreement with a possible presence of teratocytes, in turn suggesting that eggs were indeed injected by the wasp. Although we could not assess whether the sequenced larva would have developed into an adult and been fertile, we have verified by PCR the presence of bracovirus circles in several adults of *S. nonagrioides* that survived parasitism by *C. typhae* in our laboratory (data not shown). Altogether, these results tend to support the hypothesis according to which wasp-to-lepidopteran horizontal transfer of bracovirus segments can occur through HIM-mediated integration.

MATERIALS AND METHODS

DNA extraction, library preparation, and sequencing of the *C. typhae* genome. The DNA extraction was performed on *C. typhae* individuals from an isofemale line that has been reared in the Evolution, Génomes, Comportement, et Écologie (EGCE) laboratory (Gif-sur-Yvette, France) since 2015, from a strain reared at the International Centre of Insect Physiology and Ecology (Nairobi, Kenya) since 2013, when it was initially collected from the Kobodo locality in Kenya (0.679S, 34.412E). In order to obtain high-quality DNA, several individuals were pooled and ground in liquid nitrogen to give 100 mg of fine dry powder. The DNA was then extracted using Nucleobond AXG100 columns and buffer set IV from Macherey-Nagel, following the manufacturer's protocol. We obtained 26 μg of DNA, quantified with a Qubit fluorometer (Thermo Fisher Scientific). The integrity of DNA was checked on an agarose gel, and Nanodrop measurements were performed to confirm the absence of proteins and other contaminants. For whole-*C. typhae* genome sequencing, we subcontracted the French National Sequencing Center (Genoscope, Evry, France) to prepare two types of DNA libraries according to the requirements for Illumina and ONT sequencing. The Illumina library was sequenced on a MiSeq platform using the 300-bp paired-end sequencing mode with a targeted mean insert size of 350-bp (see Table S1 in the supplemental material). Paired-end reads were trimmed of adapters and low-quality bases and then merged into single reads using the BBMerge tool (64). For Nanopore sequencing, preparation of libraries was carried out with a 1D genomic DNA ligation protocol (SQK-LSK109; ONT) and sequenced using R9.4.1 flow cells on both MinION and PromethION sequencers (ONT) (see Table S1).

Assembly of the *C. typhae* genome. The genome size was first estimated from a preliminary assembly obtained from Illumina reads with ABySS v2.0 (65) using a k-mer length of 96. The genome assembly was then performed *de novo* with Flye v2.5 (66) using 30× the longest ONT reads (see Table S1). The resulting Nanopore assembly was polished using Racon v1.5.7 (67) after mapping about 2 Gb of the

longest raw ONT reads (see Table S1) with Minimap2 v2.17-r941 (68) and then Pilon v1.23 (69) using the merged Illumina reads mapped with BBMap v37.62 (70). The completeness of the genome assembly was assessed by searching for similarities to highly conserved genes among insects. For this purpose, we ran BUSCO v3.0.1 in genome mode, specifying a profile library of 1,658 single-copy core genes (April 2019 release) (35). Finally, scaffolds were checked for potential contamination by sequences from other organisms by visualizing them with Blobtools v1.1.1 using taxon-annotated GC-coverage plots. Blobtools assigns scaffolds to taxonomic ranks depending on their homologies, using both BLASTN (NCBI nucleotide database downloaded in November 2019) and BLASTX (UniRef90 protein database downloaded in November 2019). For each scaffold, Blobtools sums up scores of all hits by taxonomic rank and retains the best rank for the taxonomic assignment.

Annotation of the *C. typhae* genome. TEs were *de novo* identified and annotated in genomic sequences using TEdenovo and TEannot pipelines, respectively, included in the REPET package v2.5 (71). To construct a *de novo* repeat library, repeats were first screened using Recon (73), Grouper (72), and Piler (74). Consensus repeats were then classified into families using PASTEClassifier and filtered for all potential wasp genes corresponding to multigenic families. The TE library built by TEdenovo (71, 72) was then applied to perform a homology-based repeat search in the genome using TEannot (71, 72). Gene annotation was then performed on the repeat-masked assembly by running two iterations of MAKER v2.31.10 (75). The first iteration of MAKER used alignments of *C. vestalis* transcriptome assembly (est2genome = 1), and both reviewed Hexapoda and *Polydnaviridae* UniProt-Swiss-Prot proteins (January 2020 release) (protein2genome = 1) as sources of evidence for homology-based gene prediction. The resulting gene prediction was then used to train SNAP v2006-07-28 (77) and AUGUSTUS v3.3.2 (37) in order to construct *ab initio* gene models. The second run of MAKER allowed refinement of all of these gene models in a GFF3 output file. Predicted genes were functionally annotated with InterProScan v5.39-77.0 (76) using the PfamA database v32.0 (38) and with BLASTP v2.7.1+ using the UniProt-Swiss-Prot database (January 2020 release). Finally, functional annotations obtained were integrated in the final GFF3 file by using ipr_update_gff and maker_functional_gff modules distributed by MAKER.

Annotation of CtBV proviral segments. The localization of bracovirus proviral segments is relatively well conserved between species of the *Cotesia* genus and even with *M. demolitor*, which is more distantly related (3, 78). We annotated the proviral segments of *C. typhae* based on similarity searches using the proviral segments of its closest relative species (*C. sesamiae* and *C. congregata*) as queries. In *Cotesia congregata*, proviral segments are numbered from S1 to S37, including a segment that is no longer functional (pseudosegment 34 [ps34]) (20, 79). *C. congregata* has 36 proviral segments, and *C. sesamiae* has at least 26 proviral segments. The higher number of proviral segments in *C. congregata* results in part from extensions by duplications (responsible for 7 new segments at the macrolocus, for example) (20) and possibly from some losses in *C. sesamiae*.

The coding regions of the 26 segments of *C. sesamiae* (80) were aligned to the *C. typhae* genome using BLASTN to identify genes of each segment. DRJs of each *C. congregata* segment (see Data File S1 in the supplemental material) were then aligned using BLASTN searches for each homologous candidate segment in *C. typhae* to determine precisely the segment coordinates. The coding regions and DRJs of 10 segments present in *C. congregata* but not in *C. sesamiae* (segment S37new reported by Gauthier et al. [3], segments S3, S9, S19, S22, S29, and S31 in the macrolocus, and segments S10, S11, S21, and ps34 in dispersed loci [79]) were also aligned on the *C. typhae* genome. The synteny between segments and some other genes flanking the segments also helped to resolve ambiguous locations of the segments (3, 20).

Annotation of HIM-mediated duplications of viral circle sequences in other *Cotesia* species. We investigated whether any HIM-mediated duplications in *C. typhae* are shared with other *Cotesia* species, which would indicate that such duplications occurred before speciation. We used the chromosome-scale genome available for *C. congregata* and the more fragmented genomes of *C. sesamiae*, *C. flavipes*, *C. rubecula*, *C. vestalis*, and *C. glomerata* (3). In order to perform this analysis, we used the outputs of two BLASTN searches, (i) a similarity search between the *Cotesia* genomes and HIMs (HIMs of CtBV or CcBV, depending on whether the *Cotesia* species is more related to *C. typhae* or *C. congregata*) and (ii) a similarity search between the *Cotesia* genomes and the HIM-wasp genome junctions in *C. typhae* (options -max_target_seqs 5 -evalue $10e^{-6}$ for both searches). In the case of shared HIM-mediated duplications, we expect to obtain (i) hits on one-half of the HIM sequences for the first similarity search and (ii) hits on most of the length of the junctions for the second similarity search. Moreover, these two outputs should overlap; therefore, we filtered such cases with Rscript. This pipeline is applicable only to HIM-mediated duplications for which both extremities are identified and for which we can obtain the junctions. Thus, we were able to look for shared HIM-mediated duplications for 5 segments, i.e., S16_Hdp, S10_Hdp1, S10_Hdp2, S26_Hdp1, and S26_Hdp2. We also searched for additional candidate HIM-mediated duplications that would be specific to each genome. For this, we used the result of the first BLASTN output and that of a BLASTN similarity search between *Cotesia* genomes and DRJs (same options as for the two first searches). This third output allowed us to identify cases in which the 5′ DRJ and the 3′ DRJ of the same segment aligned next to each other (and not at the extremities of the segments, in contrast to proviral segments), as expected for HIM-mediated integrations (30).

Sequencing of *S. nonagrioides* larvae parasitized by *C. typhae*. *C. typhae* individuals used for this experiment were taken from the strain of Kobodo coming from International Centre of Insect Physiology and Ecology rearing (see "DNA extraction, library preparation, and sequencing of the *C. typhae* genome") and reared at EGCE with a protocol set up to limit inbreeding. *S. nonagrioides* larvae came from a strain reared at EGCE since 2010 from individuals collected in several localities in southwest France and refreshed yearly with such individuals. Eighteen *S. nonagrioides* larvae were each parasitized by a different *C. typhae* female. Ovipositions were confirmed by visual observations for all of them. During

oviposition, *C. typhae* lays a relatively large number of eggs in its host, generally ranging between 70 and 110 eggs (34). Larval development typically takes about 14 days under laboratory conditions until wasp larvae emerge from their host and pupate (52). Here, we placed the larvae at −80°C 7 days after oviposition. We then dissected the 18 larvae to check for the presence of wasp larvae, which at this stage measure about 5 mm and can be easily spotted by eye. The apparent success of wasp larval development before their storage was observed in 12 caterpillars. We then collected hemolymph, heads, ganglionic chains, and fat bodies from 6, 3, 9, and 1, respectively, of these 12 caterpillars. In total, we collected 780 μl of hemolymph. The minimum amount of each tissue necessary to extract sufficient amounts of DNA for Illumina sequencing (at least 500 ng at a concentration of at least 50 ng/μl) was determined in a separate experiment. Except for the hemolymph, all samples were rinsed multiple times with phosphate-buffered saline. DNA was then extracted from a pool of each tissue (except the fat body) using the DNeasy blood and tissue kit (Qiagen). We also extracted DNA from 1 of the 6 whole larvae in which we were unable to find any wasp embryos. We subcontracted Novogen to build a paired-end library (2×150 bp; insert size, 350 bp) for each sample. Each sample was then sequenced on an Illumina platform to produce a targeted amount of 100 Gbp.

Assessment of sequencing coverage on the genome of *C. typhae* and *S. nonagrioides*. Sequencing coverage was assessed on the genome of *C. typhae* assembled in this study, as well as on that of *S. nonagrioides* described by Muller et al. (40) (GenBank accession number JADWQK000000000). In brief, the genome was assembled using short Illumina reads and long ONT reads using the MaSurCA assembler (81), followed by a run of the purge_dup pipeline (82) to remove scaffolds with low coverage, partial overlaps, and haplotigs. The resulting assembly is composed of 2,253 scaffolds with an N_{50} value of 1,105 kbp and a total size of 1,021 Mpb. It contains 96% of Lepidoptera BUSCO genes, 2.7% of which are duplicated (40).

Adapters were removed and reads were quality trimmed with Trimmomatic v0.38 (options LEADING:20, TRAILING:20, SLIDINGWINDOW:4:15, and MINLEN:36) (83). Raw and trimmed read quality was assessed using FastQC v0.11.8 (84). To obtain statistics on sequencing depth, we aligned trimmed paired-end reads from the 5 samples using Bowtie2 v2.3.4.2 in end-to-end mode separately on the wasp and moth genomes (85). The resulting SAM files were sorted and converted into BAM files with SAMtools v1.7. Finally, sequencing depth was calculated with bedtools genomecov v2.26.0 for each sample for both *C. typhae* and *S. nonagrioides* genomes.

Characterization of CtBV circle integrations into the genome of *S. nonagrioides*. Raw fastq files were converted into fasta files with the seqtk seq command (option -a). Resulting fasta files were aligned on the *C. typhae* genome with BLASTN v2.6.0 (options -task megablast, -max_target_seqs 2 -outfmt 6). Reads that aligned on *C. typhae* were extracted and aligned on the *S. nonagrioides* genome, with the same options. The resulting outputs contained alignment coordinates and other information for each read aligning on both reference genomes.

We used these outputs to identify integrations of CtBV DNA circles throughout the *S. nonagrioides* genome. For that, we searched for sequencing reads for which a portion aligned on the *S. nonagrioides* genome only and the other portion aligned on CtBV proviral segments only. Such chimeric reads were identified using an R pipeline that was previously used to identify recombination events within a single genome and that we slightly adapted for our study (86). After this pipeline, we filtered out the PCR duplicates. Briefly, wasp-caterpillar chimeric reads are identified based on the tabular BLASTN outputs as follows: (i) at least 16 bases must align only on *C. typhae*, and a minimum of 16 other bases must align only on *S. nonagrioides*; (ii) less than 10% of the read length is allowed to map to neither reference genome; (iii) no more than 20 bases can align simultaneously on both reference genomes; and (iv) no more than 5 bases must be inserted between the two genomes at the integration point. The two latter filters imply that aligned read regions are allowed to overlap by up to 20 bp or to be separated by at most 5 bp. The overlap corresponds to microhomology between CtBV DNA circles and the host genome at the integration point, whereas the separation corresponds to nontemplated addition of nucleotides at the integration point (86, 87). To check whether the microhomology lengths at integration points were consistent with those expected by chance, we simulated expected distributions following the approach described by Peccoud et al. (86). Briefly, considering the sequences of the CtBV circles and the *S. nonagrioides* genome, the distribution of homology lengths was compared to that of random chimeric reads generated *in silico*. Each *in silico* read was made of two regions extracted from random locations of CtBV circles and the *S. nonagrioides* genome. The lengths of the two regions were chosen at random, with the conditions that both were at least 28 bp and their sum was the size of a read (150 bp). These reads were then subjected to a BLAST search against the sequences from which they were generated, and the BLAST outputs were subjected to the same analysis as that performed on real data.

Localization of chimeric reads in CtBV circles. Chimeric reads mapping to CtBV circles were assigned to three categories depending on the position of the wasp-host junction, i.e., (i) chimeric reads for which the CtBV-host junction falls within HIMs, (ii) reads for which the junction falls in circles devoid of HIMs or outside HIMs in circles containing HIMs, and (iii) reads for which the wasp-host junction falls precisely in the J1 and J2 regions. The last category is included in the first one. The J1 and J2 regions were defined as the positions supported by the most chimeric reads plus the positions around that point until a position was supported by <2 reads. We defined J1 and J2 independently for each sample. To assess whether integration not involving HIMs was specific to bracovirus circles or whether it also occurred for any wasp genome regions, we compared the number of chimeras falling outside HIMs in bracovirus circles to those found in exons of wasp BUSCO genes. Considering the length and sequencing depth of BUSCO gene exons and bracovirus circles, we calculated an expected number of chimeric reads

for each circle in each sample. We then compared these expected numbers to the observed numbers of chimeras falling outside HIMs.

Data availability. The assembly and annotation of the *C. typhae* genome are available in GenBank under the accession number JAAOIC000000000.2 and at the BioInformatics Platform for Agroecosystem Arthropods (BIPAA) (https://bipaa.genouest.org/sp/cotesia_typhae/). The raw sequencing reads for the 5 samples of *S. nonagrioides* parasitized by *C. typhae* are available in the NCBI database under BioProject number PRJNA718433.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

SUPPLEMENTAL FILE 1, PDF file, 5.9 MB.

SUPPLEMENTAL FILE 2, XLSX file, 0.01 MB.

ACKNOWLEDGMENTS

We thank Claire Capdevielle-Dulac for extracting the DNA of *Cotesia typhae* that was used for sequencing, as well as Rémi Jeannette, Odile Giraudier, Hubert Marteau, and Sylvie Nortier for their assistance with insect rearing. We thank Arnaud Lemainque and Jean-Marc Aury from the National Sequencing Center (Genoscope) for obtaining the bulk of the *Cotesia typhae* genome raw sequences. We are grateful to Fabrice Legeai and Anthony Bretaudeau for making the genome of *Cotesia typhae* publicly available on the BIPAA.

C. typhae individuals used in this study originated from the International Centre of Insect Physiology and Ecology under the juridical framework of material transfer agreement CNRS 072057/IRD 302227/00. *C. typhae* genome sequencing was funded by the French National Research Agency (ANR) (grant CoteBio ANR17-CE32-0015-02 to L.K.). The work was also supported by ANR project TransVir ANR-15-CE32-0011-01 (to C.G.).

REFERENCES

1. Drezen J-M, Leobold M, Bézier A, Huguet E, Volkoff A-N, Herniou EA. 2017. Endogenous viruses of parasitic wasps: variations on a common theme. Curr Opin Virol 25:41–48. https://doi.org/10.1016/j.coviro.2017.07.002.
2. Strand MR, Burke GR. 2014. Polydnaviruses: nature's genetic engineers. Annu Rev Virol 1:333–354. https://doi.org/10.1146/annurev-virology-031413-085451.
3. Gauthier J, Boulain H, van Vugt JJFA, Baudry L, Persyn E, Aury J-M, Noel B, Bretaudeau A, Legeai F, Warris S, Chebbi MA, Dubreuil G, Duvic B, Kremer N, Gayral P, Musset K, Josse T, Bigot D, Bressac C, Moreau S, Periquet G, Harry M, Montagné N, Boulogne I, Sabeti-Azad M, Maïbèche M, Chertemps T, Hilliou F, Siaussat D, Amselem J, Luyten I, Capdevielle-Dulac C, Labadie K, Merlin BL, Barbe V, de Boer JG, Marbouty M, Cônsoli FL, Dupas S, Hua-Van A, Le Goff G, Bézier A, Jacquin-Joly E, Whitfield JB, Vet LEM, Smid HM, Kaiser L, Koszul R, Huguet E, Herniou EA, Drezen JM. 2021. Chromosomal scale assembly of parasitic wasp genome reveals symbiotic virus colonization. Commun Biol 4:104–115. https://doi.org/10.1038/s42003-020-01623-8.
4. Strand MR, Burke GR. 2013. Polydnavirus-wasp associations: evolution, genome organization, and function. Curr Opin Virol 3:587–594. https://doi.org/10.1016/j.coviro.2013.06.004.
5. Sharanowski BJ, Ridenbaugh RD, Piekarski PK, Broad GR, Burke GR, Deans AR, Lemmon AR, Moriarty Lemmon EC, Diehl GJ, Whitfield JB, Hines HM. 2021. Phylogenomics of Ichneumonoidea (Hymenoptera) and implications for evolution of mode of parasitism and viral endogenization. Mol Phylogenet Evol 156:107023. https://doi.org/10.1016/j.ympev.2020.107023.
6. Legeai F, Santos BF, Robin S, Bretaudeau A, Dikow RB, Lemaitre C, Jouan V, Ravallec M, Drezen J-M, Tagu D, Baudat F, Gyapay G, Zhou X, Liu S, Webb BA, Brady SG, Volkoff A-N. 2020. Genomic architecture of endogenous ichnoviruses reveals distinct evolutionary pathways leading to virus domestication in parasitic wasps. BMC Biol 18:89. https://doi.org/10.1186/s12915-020-00822-3.
7. Bézier A, Annaheim M, Herbinière J, Wetterwald C, Gyapay G, Bernard-Samain S, Wincker P, Roditi I, Heller M, Belghazi M, Pfister-Wilhem R, Periquet G, Dupuy C, Huguet E, Volkoff A-N, Lanzrein B, Drezen J-M. 2009. Polydnaviruses of braconid wasps derive from an ancestral nudivirus. Science 323:926–930. https://doi.org/10.1126/science.1166788.
8. Volkoff A-N, Jouan V, Urbach S, Samain S, Bergoin M, Wincker P, Demettre E, Cousserans F, Provost B, Coulibaly F, Legeai F, Béliveau C, Cusson M, Gyapay G, Drezen J-M. 2010. Analysis of virion structural components reveals vestiges of the ancestral ichnovirus genome. PLoS Pathog 6:e1000923. https://doi.org/10.1371/journal.ppat.1000923.
9. Pichon A, Bézier A, Urbach S, Aury J-M, Jouan V, Ravallec M, Guy J, Cousserans F, Thézé J, Gauthier J, Demettre E, Schmieder S, Wurmser F, Sibut V, Poirié M, Colinet D, da Silva C, Couloux A, Barbe V, Drezen J-M, Volkoff A-N. 2015. Recurrent DNA virus domestication leading to different parasite virulence strategies. Sci Adv 1:e1501150. https://doi.org/10.1126/sciadv.1501150.
10. Béliveau C, Cohen A, Stewart D, Periquet G, Djoumad A, Kuhn L, Stoltz D, Boyle B, Volkoff A-N, Herniou EA, Drezen J-M, Cusson M. 2015. Genomic and proteomic analyses indicate that banchine and campoplegine polydnaviruses have similar, if not identical, viral ancestors. J Virol 89:8909–8921. https://doi.org/10.1128/JVI.01001-15.
11. Burke GR, Simmonds TJ, Sharanowski BJ, Geib SM. 2018. Rapid viral symbiogenesis via changes in parasitoid wasp genome architecture. Mol Biol Evol 35:2463–2474. https://doi.org/10.1093/molbev/msy148.
12. Burke GR. 2019. Common themes in three independently derived endogenous nudivirus elements in parasitoid wasps. Curr Opin Insect Sci 32:28–35. https://doi.org/10.1016/j.cois.2018.10.005.
13. Herniou EA, Huguet E, Thézé J, Bézier A, Periquet G, Drezen J-M. 2013. When parasitic wasps hijacked viruses: genomic and functional evolution of polydnaviruses. Philos Trans R Soc Lond B Biol Sci 368:20130051. https://doi.org/10.1098/rstb.2013.0051.
14. Murphy N, Banks JC, Whitfield JB, Austin AD. 2008. Phylogeny of the parasitic microgastroid subfamilies (Hymenoptera: Braconidae) based on sequence data from seven genes, with an improved time estimate of the origin of the lineage. Mol Phylogenet Evol 47:378–395. https://doi.org/10.1016/j.ympev.2008.01.022.
15. Rodriguez JJ, Fernández-Triana JL, Smith MA, Janzen DH, Hallwachs W, Erwin TL, Whitfield JB. 2013. Extrapolations from field studies and known faunas converge on dramatically increased estimates of global microgastrine parasitoid wasp species richness (Hymenoptera: Braconidae). Insect Conserv Divers 6:530–536. https://doi.org/10.1111/icad.12003.

16. Wetterwald C, Roth T, Kaeslin M, Annaheim M, Wespi G, Heller M, Mäser P, Roditi I, Pfister-Wilhelm R, Bézier A, Gyapay G, Drezen J-M, Lanzrein B. 2010. Identification of bracovirus particle proteins and analysis of their transcript levels at the stage of virion formation. J Gen Virol 91:2610–2619. https://doi.org/10.1099/vir.0.022699-0.
17. Desjardins CA, Gundersen-Rindal DE, Hostetler JB, Tallon LJ, Fadrosh DW, Fuester RW, Pedroni MJ, Haas BJ, Schatz MC, Jones KM, Crabtree J, Forberger H, Nene V. 2008. Comparative genomics of mutualistic viruses of *Glyptapanteles* parasitic wasps. Genome Biol 9:R183. https://doi.org/10.1186/gb-2008-9-12-r183.
18. Gasmi L, Boulain H, Gauthier J, Hua-Van A, Musset K, Jakubowska AK, Aury J-M, Volkoff A-N, Huguet E, Herrero S, Drezen J-M. 2015. Recurrent domestication by Lepidoptera of genes from their parasites mediated by bracoviruses. PLoS Genet 11:e1005470. https://doi.org/10.1371/journal.pgen.1005470.
19. Zhang F, Zhang J, Yang Y, Wu Y. 2019. A chromosome-level genome assembly for the beet armyworm (*Spodoptera exigua*) using PacBio and Hi-C sequencing. bioRxiv 2019.12.26.889121. https://doi.org/10.1101/2019.12.26.889121.
20. Bézier A, Louis F, Jancek S, Periquet G, Thézé J, Gyapay G, Musset K, Lesobre J, Lenoble P, Dupuy C, Gundersen-Rindal D, Herniou EA, Drezen J-M. 2013. Functional endogenous viral elements in the genome of the parasitoid wasp *Cotesia congregata*: insights into the evolutionary dynamics of bracoviruses. Philos Trans R Soc Lond B Biol Sci 368:20130047. https://doi.org/10.1098/rstb.2013.0047.
21. Louis F, Bézier A, Periquet G, Ferras C, Drezen J-M, Dupuy C. 2013. The bracovirus genome of the parasitoid wasp *Cotesia congregata* is amplified within 13 replication units, including sequences not packaged in the particles. J Virol 87:9649–9660. https://doi.org/10.1128/JVI.00886-13.
22. Gruber A, Stettler P, Heiniger P, Schümperli D, Lanzrein B. 1996. Polydnavirus DNA of the braconid wasp *Chelonus inanitus* is integrated in the wasp's genome and excised only in later pupal and adult stages of the female. J Gen Virol 77:2873–2879. https://doi.org/10.1099/0022-1317-77-11-2873.
23. Savary S, Beckage N, Tan F, Periquet G, Drezen JM. 1997. Excision of the polydnavirus chromosomal integrated EP1 sequence of the parasitoid wasp *Cotesia congregata* (Braconidae, Microgastinae) at potential recombinase binding sites. J Gen Virol 78:3125–3134. https://doi.org/10.1099/0022-1317-78-12-3125.
24. Pasquier-Barre F, Dupuy C, Huguet E, Monteiro F, Moreau A, Poirié M, Drezen J-M. 2002. Polydnavirus replication: the EP1 segment of the parasitoid wasp *Cotesia congregata* is amplified within a larger precursor molecule. J Gen Virol 83:2035–2045. https://doi.org/10.1099/0022-1317-83-8-2035.
25. Desjardins CA, Gundersen-Rindal DE, Hostetler JB, Tallon LJ, Fuester RW, Schatz MC, Pedroni MJ, Fadrosh DW, Haas BJ, Toms BS, Chen D, Nene V. 2007. Structure and evolution of a proviral locus of *Glyptapanteles indiensis* bracovirus. BMC Microbiol 7:61. https://doi.org/10.1186/1471-2180-7-61.
26. Beck MH, Zhang S, Bitra K, Burke GR, Strand MR. 2011. The encapsidated genome of *Microplitis demolitor* bracovirus integrates into the host *Pseudoplusia includens*. J Virol 85:11685–11696. https://doi.org/10.1128/JVI.05726-11.
27. Chevignon G, Thézé J, Cambier S, Poulain J, Da Silva C, Bézier A, Musset K, Moreau SJM, Drezen J-M, Huguet E. 2014. Functional annotation of *Cotesia congregata* bracovirus: identification of viral genes expressed in parasitized host immune tissues. J Virol 88:8795–8812. https://doi.org/10.1128/JVI.00209-14.
28. McKelvey TA, Lynn DE, Gundersen-Rindal D, Guzo D, Stoltz DA, Guthrie KP, Taylor PB, Dougherty EM. 1996. Transformation of gypsy moth (*Lymantria dispar*) cell lines by infection with *Glyptapanteles indiensis* polydnavirus. Biochem Biophys Res Commun 225:764–770. https://doi.org/10.1006/bbrc.1996.1248.
29. Gundersen-Rindal DE, Lynn DE. 2003. Polydnavirus integration in lepidopteran host cells in vitro. J Insect Physiol 49:453–462. https://doi.org/10.1016/s0022-1910(03)00062-3.
30. Chevignon G, Periquet G, Gyapay G, Vega-Czarny N, Musset K, Drezen J-M, Huguet E. 2018. *Cotesia congregata* bracovirus circles encoding PTP and ankyrin genes integrate into the DNA of parasitized *Manduca sexta* hemocytes. J Virol 92:e00438-18. https://doi.org/10.1128/JVI.00438-18.
31. Wang Z, Zhou Y, Yang J, Ye X, Shi M, Huang J, Chen X. 2020. Genome-wide profiling of *Diadegma semiclausum* ichnovirus integration in parasitized *Plutella xylostella* hemocytes identifies host integration motifs and insertion sites. Front Microbiol 11:608346. https://doi.org/10.3389/fmicb.2020.608346.
32. Kaiser L, Fernandez-Triana J, Capdevielle-Dulac C, Chantre C, Bodet M, Kaoula F, Benoist R, Calatayud P-A, Dupas S, Herniou EA, Jeannette R, Obonyo J, Silvain J-F, Ru BL. 2017. Systematics and biology of *Cotesia typhae* sp. n. (Hymenoptera, Braconidae, Microgastrinae), a potential biological control agent against the noctuid Mediterranean corn borer, *Sesamia nonagrioides*. Zookeys 105–136. https://doi.org/10.3897/zookeys.682.13016.
33. Kaiser L, Ru BPL, Kaoula F, Paillusson C, Capdevielle-Dulac C, Obonyo JO, Herniou EA, Jancek S, Branca A, Calatayud P-A, Silvain J-F, Dupas S. 2015. Ongoing ecological speciation in *Cotesia sesamiae*, a biological control agent of cereal stem borers. Evol Appl 8:807–820. https://doi.org/10.1111/eva.12260.
34. Benoist R, Capdevielle-Dulac C, Chantre C, Jeannette R, Calatayud P-A, Drezen J-M, Dupas S, Rouzic AL, Ru BL, Moreau L, Dijk EV, Kaiser L, Mougel F. 2020. Quantitative trait loci involved in the reproductive success of a parasitoid wasp. Mol Ecol 29:3476–3493. https://doi.org/10.1111/mec.15567.
35. Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM. 2018. BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol Biol Evol 35:543–548. https://doi.org/10.1093/molbev/msx319.
36. Kumar S, Jones M, Koutsovoulos G, Clarke M, Blaxter M. 2013. Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. Front Genet 4:237. https://doi.org/10.3389/fgene.2013.00237.
37. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. 2006. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res 34:W435–W439. https://doi.org/10.1093/nar/gkl200.
38. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, Bateman A. 2016. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res 44:D279–D285. https://doi.org/10.1093/nar/gkv1344.
39. Serbielle C, Dupas S, Perdereau E, Héricourt F, Dupuy C, Huguet E, Drezen J-M. 2012. Evolutionary mechanisms driving the evolution of a large polydnavirus gene family coding for protein tyrosine phosphatases. BMC Evol Biol 12:253. https://doi.org/10.1186/1471-2148-12-253.
40. Muller H, Ogereau D, Da-Lage J-L, Capdevielle C, Pollet N, Fortuna T, Jeannette R, Kaiser L, Gilbert C. 2021. Draft nuclear genome and complete mitogenome of the Mediterranean corn borer, *Sesamia nonagrioides*, a major pest of maize. G3 (Bethesda) 11:jkab155. https://doi.org/10.1093/g3journal/jkab155.
41. Takasu K, Hoang Le K. 2007. The larval parasitoid *Microplitis croceipes* oviposits in conspecific adults. Naturwissenschaften 94:200–206. https://doi.org/10.1007/s00114-006-0181-3.
42. Potting RPJ, Overholt WA, Danso FO, Takasu K. 1997. Foraging behavior and life history of the stemborer parasitoid *Cotesia flavipes* (Hymenoptera: Braconidae). J Insect Behav 10:13–29. https://doi.org/10.1007/BF02765472.
43. Kester KM, Barbosa P. 1994. Behavioral responses to host foodplants of two populations of the insect parasitoid *Cotesia congregata* (Say). Oecologia 99:151–157. https://doi.org/10.1007/BF00317096.
44. Beckage NE, de Buron I. 1997. Developmental changes in teratocytes of the braconid wasp *Cotesia congregata* in larvae of the tobacco hornworm, *Manduca sexta*. J Insect Physiol 43:915–930. https://doi.org/10.1016/s0022-1910(97)00056-5.
45. Strand MR, Wong EA. 1991. The growth and role of *Microplitis demolitor* teratocytes in parasitism of *Pseudoplusia includens*. J Insect Physiol 37:503–515. https://doi.org/10.1016/0022-1910(91)90027-W.
46. Okuda T, Kadono-Okuda K. 1995. *Perilitus coccinellae* teratocyte polypeptide: evidence for production of a teratocyte-specific 540 kDa protein. J Insect Physiol 41:819–825. https://doi.org/10.1016/0022-1910(95)00009-J.
47. Hotta M, Okuda T, Tanaka T. 2001. *Cotesia kariyai* teratocytes: growth and development. J Insect Physiol 47:31–41. https://doi.org/10.1016/s0022-1910(00)00089-5.
48. Mancini D, Garonna AP, Pedata PA. 2016. To divide or not to divide: an alternative behavior for teratocytes in *Encarsia pergandiella* (Hymenoptera: Aphelinidae). Arthropod Struct Dev 45:57–63. https://doi.org/10.1016/j.asd.2015.10.003.
49. Jiang H, Vilcinskas A, Kanost MR. 2010. Immunity in lepidopteran insects, p 181–204. *In* Söderhäll K (ed), Invertebrate immunity. Springer, Boston, MA.
50. Lavine MD, Strand MR. 2002. Insect hemocytes and their role in immunity. Insect Biochem Mol Biol 32:1295–1309. https://doi.org/10.1016/s0965-1748(02)00092-9.

51. Beck MH, Inman RB, Strand MR. 2007. *Microplitis demolitor* bracovirus genome segments vary in abundance and are individually packaged in virions. Virology 359:179–189. https://doi.org/10.1016/j.virol.2006.09.002.
52. Benoist R, Chantre C, Capdevielle-Dulac C, Bodet M, Mougel F, Calatayud PA, Dupas S, Huguet E, Jeannette R, Obonyo J, Odorico C, Silvain JF, Le Ru B, Kaiser L. 2017. Relationship between oviposition, virulence gene expression and parasitism success in *Cotesia typhae* nov. sp. parasitoid strains. Genetica 145:469–479. https://doi.org/10.1007/s10709-017-9987-5.
53. Bézier A, Herbinière J, Lanzrein B, Drezen J-M. 2009. Polydnavirus hidden face: the genes producing virus particles of parasitic wasps. J Invertebr Pathol 101:194–203. https://doi.org/10.1016/j.jip.2009.04.006.
54. Burke GR, Thomas SA, Eum JH, Strand MR. 2013. Mutualistic polydnaviruses share essential replication gene functions with pathogenic ancestors. PLoS Pathog 9:e1003348. https://doi.org/10.1371/journal.ppat.1003348.
55. Seol J-H, Shim EY, Lee SE. 2018. Microhomology-mediated end joining: good, bad and ugly. Mutat Res 809:81–87. https://doi.org/10.1016/j.mrfmmm.2017.07.002.
56. Lin CL, Lee JC, Chen SS, Wood HA, Li ML, Li CF, Chao YC. 1999. Persistent Hz-1 virus infection in insect cells: evidence for insertion of viral DNA into host chromosomes and viral infection in a latent status. J Virol 73:128–139. https://doi.org/10.1128/JVI.73.1.128-139.1999.
57. Wang J, Liu Y, Liu Y, Du K, Xu S, Wang Y, Krupovic M, Chen X. 2018. A novel family of tyrosine integrases encoded by the temperate pleolipovirus SNJ2. Nucleic Acids Res 46:2521–2536. https://doi.org/10.1093/nar/gky005.
58. Grindley NDF, Whiteson KL, Rice PA. 2006. Mechanisms of site-specific recombination. Annu Rev Biochem 75:567–605. https://doi.org/10.1146/annurev.biochem.73.011303.073908.
59. Bischerour J, Bhullar S, Denby Wilkes C, Régnier V, Mathy N, Dubois E, Singh A, Swart E, Arnaiz O, Sperling L, Nowacki M, Bétermier M. 2018. Six domesticated PiggyBac transposases together carry out programmed DNA elimination in paramecium. Elife 7:e37927. https://doi.org/10.7554/eLife.37927.
60. Sarfraz M, Dosdall LM, Keddie BA. 2007. Resistance of some cultivated Brassicaceae to infestations by *Plutella xylostella* (Lepidoptera: Plutellidae). J Econ Entomol 100:215–224. https://doi.org/10.1603/0022-0493(2007)100[215:ROSCBT]2.0.CO;2.
61. Kahuthia-Gathu R, Othim STO. 2019. Effects of two cultivated *Brassica* spp. on the development and performance of *Diadegma semiclausum* (Hymenoptera: Ichneumonidae) and *Cotesia vestalis* (Hymenoptera: Braconidae) parasitizing *Plutella xylostella* (Lepidoptera: Plutellidae) in Kenya. J Econ Entomol 112:2094–2102. https://doi.org/10.1093/jee/toz144.
62. Schneider SE, Thomas JH. 2014. Accidental genetic engineers: horizontal sequence transfer from parasitoid wasps to their lepidopteran hosts. PLoS One 9:e109446. https://doi.org/10.1371/journal.pone.0109446.
63. Mabiala-Moundoungou ADN, Doury G, Eslin P, Cherqui A, Prévost G. 2010. Deadly venom of *Asobara japonica* parasitoid needs ovarian antidote to regulate host physiology. J Insect Physiol 56:35–41. https://doi.org/10.1016/j.jinsphys.2009.09.001.
64. Bushnell B, Rood J, Singer E. 2017. BBMerge: accurate paired shotgun read merging via overlap. PLoS One 12:e0185056. https://doi.org/10.1371/journal.pone.0185056.
65. Jackman SD, Vandervalk BP, Mohamadi H, Chu J, Yeo S, Hammond SA, Jahesh G, Khan H, Coombe L, Warren RL, Birol I. 2017. ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. Genome Res 27:768–777. https://doi.org/10.1101/gr.214346.116.
66. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol 37:540–546. https://doi.org/10.1038/s41587-019-0072-8.
67. Vaser R, Sović I, Nagarajan N, Šikić M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res 27:737–746. https://doi.org/10.1101/gr.214270.116.
68. Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094–3100. https://doi.org/10.1093/bioinformatics/bty191.
69. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9:e112963. https://doi.org/10.1371/journal.pone.0112963.
70. Bushnell B. 2014. BBMap: a fast, accurate, splice-aware aligner. lbnl-7065e. Lawrence Berkeley National Laboratory, Berkeley, CA.
71. Flutre T, Duprat E, Feuillet C, Quesneville H. 2011. Considering transposable element diversification in de novo annotation approaches. PLoS One 6:e16526. https://doi.org/10.1371/journal.pone.0016526.
72. Quesneville H, Bergman CM, Andrieu O, Autard D, Nouaud D, Ashburner M, Anxolabehere D. 2005. Combined evidence annotation of transposable elements in genome sequences. PLoS Comput Biol 1:e22. https://doi.org/10.1371/journal.pcbi.0010022.
73. Bao Z, Eddy SR. 2002. Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res 12:1269–1276. https://doi.org/10.1101/gr.88502.
74. Edgar RC, Myers EW. 2005. PILER: identification and classification of genomic repeats. Bioinformatics 21(Suppl 1):i152–158. https://doi.org/10.1093/bioinformatics/bti1003.
75. Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 12:491. https://doi.org/10.1186/1471-2105-12-491.
76. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R. 2005. InterProScan: protein domains identifier. Nucleic Acids Res 33:W116–W120.
77. Korf I. 2004. Gene finding in novel genomes. BMC Bioinformatics 5:59. https://doi.org/10.1186/1471-2105-5-59.
78. Burke GR, Walden KKO, Whitfield JB, Robertson HM, Strand MR. 2014. Widespread genome reorganization of an obligate virus mutualist. PLoS Genet 10:e1004660. https://doi.org/10.1371/journal.pgen.1004660.
79. Espagne E, Dupuy C, Huguet E, Cattolico L, Provost B, Martins N, Poirié M, Periquet G, Drezen JM. 2004. Genome sequence of a polydnavirus: insights into symbiotic virus evolution. Science 306:286–289. https://doi.org/10.1126/science.1103066.
80. Jancek S, Bézier A, Gayral P, Paillusson C, Kaiser L, Dupas S, Le Ru BP, Barbe V, Periquet G, Drezen J-M, Herniou EA. 2013. Adaptive selection on bracovirus genomes drives the specialization of *Cotesia* parasitoid wasps. PLoS One 8:e64432. https://doi.org/10.1371/journal.pone.0064432.
81. Zimin AV, Puiu D, Luo M-C, Zhu T, Koren S, Marçais G, Yorke JA, Dvořák J, Salzberg SL. 2017. Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. Genome Res 27:787–792. https://doi.org/10.1101/gr.213405.116.
82. Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. 2020. Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics 36:2896–2898. https://doi.org/10.1093/bioinformatics/btaa025.
83. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. https://doi.org/10.1093/bioinformatics/btu170.
84. Andrews S. 2011. FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc.
85. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. https://doi.org/10.1038/nmeth.1923.
86. Peccoud J, Lequime S, Moltini-Conclois I, Giraud I, Lambrechts L, Gilbert C. 2018. A survey of virus recombination uncovers canonical features of artificial chimeras generated during deep sequencing library preparation. G3 (Bethesda) 8:1129–1138. https://doi.org/10.1534/g3.117.300468.
87. Gilbert C, Peccoud J, Chateigner A, Moumen B, Cordaux R, Herniou EA. 2016. Continuous influx of genetic material from host to virus populations. PLoS Genet 12:e1005838. https://doi.org/10.1371/journal.pgen.1005838.

10 - Article n°4: Investigating bracovirus chromosomal integration and inheritance in lepidopteran host and non-target species

After the results of the article n°3, in which we found that DNA circles of *C. typhae* can integrate in all the tissues of *S. nonagrioides* we studied, and even in the caterpillar in which parasitism failed, we were wondering whether these integrations could be transmitted to the next generation of moths that would reproduce after surviving to parasitism. For this to happen, some integrations have to take place in the germinal cells, to persist until adult stage, and finally to be transmitted to viable offspring. To answer this question, I designed with colleagues a study in which we investigated CtBV integration in two adults that survived parasitism and in 500 offspring of surviving caterpillars. Here, I was involved in all the experiments and analyses. Although we found DNA integrations in the two adults, in the same proportion as in the caterpillar of Article n°3 for one of them, we could not find any integration in the 500 offspring. This result means that if transmission is possible, it happens at a frequency too low to be detectable by our pipeline. This result may be perceived as encouraging in terms of risk linked to uncontrolled HT during biocontrol using *C. typhae*. However, at the scale of a biocontrol campaign that would involve numerous parasitized caterpillars every year, even a low probability of HT could lead to numerous BV transmissions.

In addition, we investigated the genetic risks of HT of CtBV to non-target lepidopteran hosts that share the same ecological niche as *S. nonagroides* in France: *Nonagria typhae*, *Globia sparganii*, and *Chilo phragmitella*. In this second part, my contribution to experiments only consisted in extracting DNA of *C. phragmitella*, I did not collect the non-target caterpillars in the field nor did I parasitize them. Then I did all the analyses, except for the assemblies of the reference genomes of *Nonagria typhae* and *Globia sparganii*, which were done by Camille Heisserer, whom I co-supervised during her M2 internship. In this study, we found massive CtBV integrations in the genomes of the three non-target species which means that CtBV are able to recognize and enter cells of species other than *S. nonagrioides*, even in species that can be quite divergent (*C. phragmitella* diverged from *S. nonagrioides* more than 100 million years). Although integrations took place via HIM in all species, it seems that a second mechanism is involved in *C. phragmitella* suggesting that host factors are involved in CtBV integration. In the context of bio-control, this massive integration in these non-target species is not encouraging, although the likelihood with which *C. typhae* may be able to parasitize them in nature appears very low according to other experiments performed in parallel by other members of our laboratory.

Because the published version is not freely available, I included in this manuscript the postprint version. The published version can be found at the following link:
https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.16685.

Investigating bracovirus chromosomal integration and inheritance in lepidopteran host and non-target species

Héloïse Muller[1], Camille Heisserer[1,2], Taiadjana Fortuna[1], Florence Mougel[1], Elisabeth Huguet[2], Laure Kaiser[1], Clément Gilbert[1*]

[1]Université Paris-Saclay, CNRS, IRD, UMR Évolution, Génomes, Comportement et Écologie, 91190 Gif-sur-Yvette, France.

[2]UMR 7261 CNRS, Institut de Recherche sur la Biologie de l'Insecte, Faculté des Sciences et Techniques, Université de Tours, Tours, France

*Author for Correspondence: Clément Gilbert, Université Paris-Saclay, CNRS, IRD, UMR Évolution, Génomes, Comportement et Écologie, 91190 Gif-sur-Yvette, France; email: clement.gilbert@egce.cnrs-gif.fr

Keywords: bio-control | Braconidae | bracoviruses | evolutionnary genomics | horizontal transfer | parasitoid wasps

Abstract

Bracoviruses (BVs) are domesticated viruses found in braconid parasitoid wasp genomes. They are composed of domesticated genes from a nudivrius, coding viral particles in which wasp DNA circles are packaged. BVs are viewed as possible vectors of horizontal transfer of genetic material (HT) from wasp to their hosts because they are injected, together with wasp eggs, by female wasps in their host larvae, and because they undergo massive chromosomal integration in multiple host tissues. Here, we show that chromosomal integrations of the *Cotesia typhae* BV (CtBV) persist up to the adult stage in individuals of its natural host, *S. nonagrioides*, that survived parasitism. However, while reproducing host adults can bear an average of nearly two CtBV integrations per haploid genome, we were unable to retrieve any of these integrations in 500 of their offspring using Illumina sequencing. This suggests either that host gametes are less targeted by CtBVs than somatic cells or that gametes bearing BV integrations are non-functional. We further show that CtBV can massively integrate into the chromosomes of other lepidopteran species that are not normally targeted by the wasp in the wild, including one which is at least 100 million years divergent from the natural host. Cell entry and chromosomal integration of BVs are thus unlikely to be major factors shaping wasp host range. Together, our results shed new light on the conditions under which BV-mediated wasp-to-host HT may occur and provide information that may be helpful to evaluate the potential risks of uncontrolled HT associated with the use of parasitoid wasps as biocontrol agents.

Introduction

Horizontal transfer (HT) of genetic material involves transmission of DNA by means other than reproduction (Keeling & Palmer, 2008). Widespread in prokaryotes, HT is also increasingly recognized as an important evolutionary process in eukaryotes (Aubin et al., 2021; Boto, 2014; Van Etten & Bhattacharya, 2020). Many studies report remarkable examples of HT of genes with important consequences in multiple lineages (Danchin et al., 2016; Simion et al., 2021; Wybouw et al., 2016; Xia et al., 2021; Li et al. 2022). In addition, large scale studies of transposable elements (TEs) inferred dozens to thousands of horizontal transfers of transposable elements in various multicellular eukaryotes such as plants (Baidouri et al., 2014), insects (Peccoud et al., 2017; Heringer et al. 2022; Wallau et al. 2016), and vertebrates (Zhang et al., 2020). Given the profound impact TEs have on genome evolution, HT of TEs can be viewed as a key process shaping eukaryote genomes (Gilbert & Feschotte, 2018). Contrasting with our detailed knowledge of HT in prokaryotes, the mechanisms and possible vectors of HT remain elusive in eukaryotes. The finding that viruses and extracellular vesicles can carry host genetic material suggests that they may accidentally act as vectors of HT (Gilbert & Cordaux, 2017; Ono et al., 2019). Here we further evaluate the conditions and frequency at which a specific type of viruses, the bracoviruses (BVs) of Braconidae endoparasitoid wasps (Hymenoptera), may mediate HT between wasps and lepidopterans.

Braconidae endoparasitoid wasps lay their eggs in insect hosts, mostly lepidopterans (see the picture of Fig. 1 as an example), which are used by

wasp larvae as substrate and food source. In addition to their eggs, females of many Braconidae wasp species inject BVs into their host, which interfere with host developmental and immunological pathways, facilitating the development of wasp embryos (Beckage & Drezen, 2011; Strand & Burke, 2014). Genes encoding structural components of these viruses originate from domestication of viral genes that became integrated into the genome of braconid wasp ancestors about 100 million years ago (labeled 1c in Fig. 1) (Gauthier et al., 2018; Herniou et al., 2013). Bracovirus structural genes, of nudiviral origin, are specifically expressed in the calyx of the ovaries, resulting in the production of viral particles into which circular double stranded DNA molecules are packaged (labeled 2 in Fig. 1). These packaged DNA circles result from amplification, excision and circularization of so-called “proviral segments” (labeled 1a and 1b in Fig. 1), which contain wasp genes and other genes of unknown evolutionary origin (Bézier et al., 2013; Drezen et al., 2014; Herniou et al., 2013). Circularization takes place in calyx cells *via* site-specific recombination at Direct Repeat Junction (DRJ) motifs, which are conserved between proviral segments and between Braconidae species (Beck et al., 2011; Desjardins et al., 2008; Gauthier et al., 2021; Gruber et al., 1996; Muller, Chebbi, et al., 2021; Pasquier-Barre et al., 2002; Savary et al., 1997). A remarkable feature of BVs is that once released in host’s cells, DNA circles containing a specific motif called Host Integration Motif (HIM) integrate massively into the host genome (labeled 3 in Fig. 1) (Beck et al., 2011; Chevignon et al., 2018). We use the term “HIM-containing circles” to refer to these circles and we call other circles, which do not contain HIM “circles devoid of HIM”. Integration involves site-specific recombination between HIMs and the host

genome, likely mediated by BV-encoded enzymes. In the wasp *Cotesia typhae*, it was shown that the 16 HIM-containing circles (out of 27 circles) undergo chromosomal integration in at least four host tissues/organs (hemocytes, fat body, ganglionic chain, and head), yielding an average of 12 to 85 integrated circles per host haploid genome (Muller, Chebbi, et al., 2021). Integration involves linearization of DNA circles through two DNA double strand breaks within HIMs and loss of a small ≃50-bp region lying between the two breaks (labeled 3 in Fig. 1). The role played by BV DNA circles integration in wasp-host interactions remains to be characterized. So far, it appears that integration does not enhance expression of circle-borne genes (Chevignon et al., 2014, 2018), nor does it enhance persistence of DNA circles throughout wasp development (Muller, Chebbi, et al., 2021).

Systematic and massive DNA circle integration in host genomes during parasitism may facilitate wasp-to-host HT with long-term consequences. Some of BV circle sequences have been found in various lepidopteran genomes that clearly result from HT from wasp to lepidopterans (Gasmi et al., 2015; Schneider & Thomas, 2014). Remarkably, some BV genes acquired by *Spodoptera* species have been domesticated and now play a role in anti-bacterial immune response (Di Lelio et al., 2019; Gasmi et al., 2015). Thus, BV DNA circles can integrate into the genome of the host germline and some host individuals surviving parasitism can transmit these integrations to their offspring (Drezen et al., 2017). These studies also demonstrated that horizontally transferred BV genes can be an important source of genetic novelty and adaptation in lepidopterans.

Integration in the host germline is consistent with the known tropism of BV particles for a large range of tissues (Beck et al., 2011; Wyder et al., 2003) and the capacity of BV DNA circles to integrate in the genome of various cell types (Muller, Chebbi, et al., 2021), although at different levels depending on the species and tissues. Survival of parasitized lepidopterans is a rather common phenomenon in nature, as the success of endoparasitoid wasps' development often strongly varies depending on the geographical origin of host populations (Gitau et al., 2007). To which extent resistant hosts retain BV DNA integrated in their genome at the adult stage, and in which proportion these wasp sequences can be transmitted vertically to the next host generation has never been investigated. Yet, given that some endoparasitoid wasps are used as biocontrol agents (Polaszek & Walker, 1992), measuring the frequency of such possible HT is relevant not only from a basic science point of view, but also to evaluate the risks of uncontrolled HT potentially associated with strategies that use exotic wasps bearing BV.

Another possible way chromosomally integrated BV circles may be vertically transmitted in host populations is through accidental parasitism of a non-permissive host species, in which by definition parasitism fails. In laboratory conditions, endoparasitoid wasps are indeed able to lay their eggs into species not known to be targeted in the wild, with varying degrees of success (Harwood et al., 1998). Whether BV DNA circle integration occurs in non-natural hosts, supporting such hosts as possible facilitators of wasp-to-host HT of BV DNA, has never been investigated so far.

In this study, we first evaluate the frequency under which integration of BV DNA circles in host germline followed by transmission to the next generation may occur in the wasp-host system involving the wasp *Cotesia typhae* and its host, the Mediterranean corn borer *Sesamia nonagrioides* (Lepidoptera: Noctuidae). S*. nonagrioides* is a major pest of corn in Mediterranean regions (Bosque-Perez & Schulthess, 1998; Kankonda et al., 2018; Moyal et al., 2011). *C. typhae* (Fernandez-Triana) (Hymenoptera: Braconidae) is a newly described species closely related to *Cotesia sesamiae* found in east Africa (Kaiser et al., 2015, 2017). The two wasp species are reproductively isolated and although the former is a generalist endoparasitoid, *C. typhae* is a specialist at both host (*S. nonagrioides*) and plant (*Typha domingensis*) levels. Due to such a specialization, in addition to inducing a high host death rate during parasitism, *C. typhae* is currently under study as a potential biological control agent against *S. nonagrioides* in Europe (Kaiser et al., 2015, 2017). We then assessed whether *C. typhae* BV DNA circles integrate into the genome of three other stemborer lepidopteran species: two Noctuidae (*Globia sparganii* and *Nonagria typhae*) and one Crambidae (*Chilo phragmitella*). These species are normally not targeted in the wild by *C. typhae* but since their distribution overlaps with that of *S. nonagrioides* in France and they share similar ecological niches (Galichet et al., 1992; Teder & Tammaru, 2002; Tewksbury et al., 2002), they might suffer non-target effects during a putative biocontrol campaign, the consequences of which are important to assess (Louda et al., 2003; Sands & Driesche, 2000).

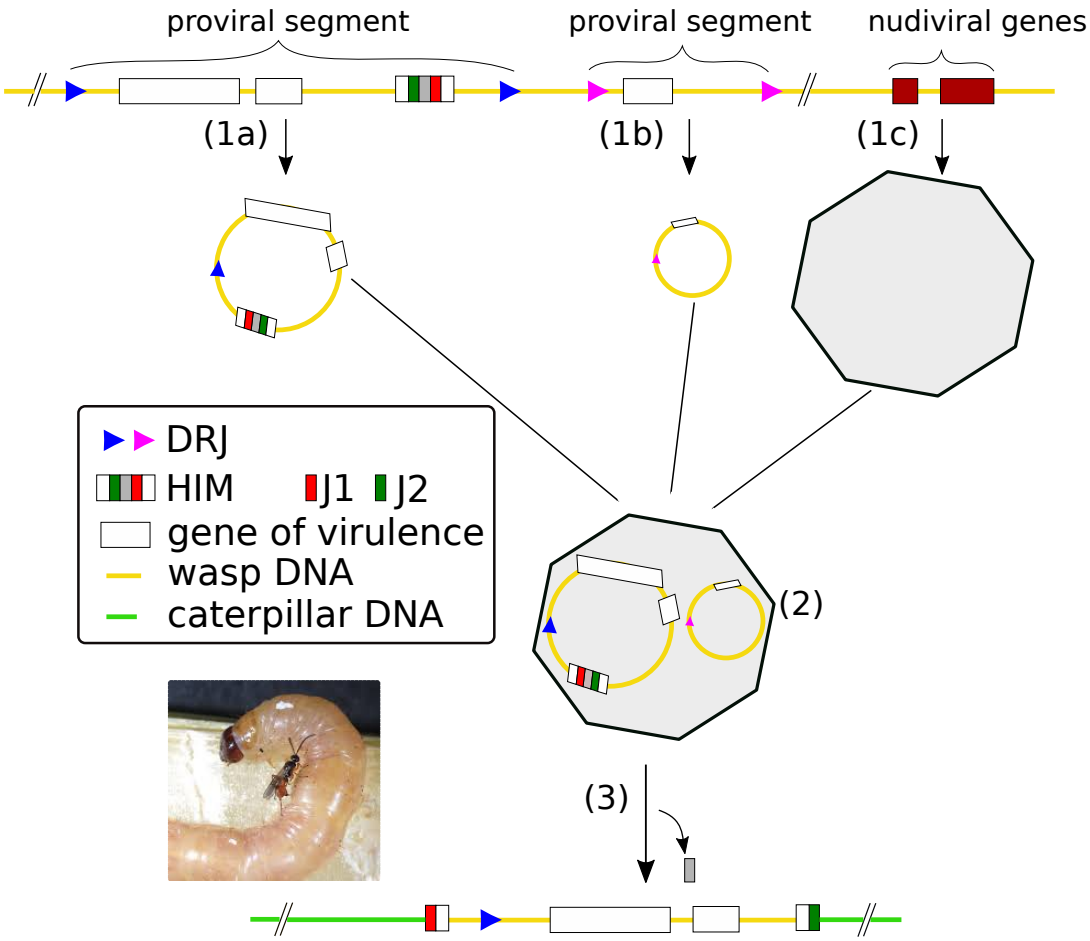


Figure 1: Structure of bracoviruses. In the wasp genome, (1a) is a HIM-containing segment, (1b) is a segment devoid of HIM, and (1c) corresponds to genes of viral origin. White rectangles are virulence genes. In the wasp calyx, (1a) and (1b) form DNA circles thanks to their DRJ, whereas (1c) form viral particles. In (2), DNA circles are packaged into viral particles, which are injected in the caterpillar host at the same time as the wasp eggs (see the picture as an example, where C. typhae is parasiting S. nonagrioides). In (3), the HIM-containing segment integrates via its HIM into the caterpillar genome, losing a ~50bp region located between J1 and J2.

Materials and methods

Parasitism and sequencing of 500 *S. nonagrioides* larvae and two adults

Sixty-two French *S. nonagrioides* larvae at the L4/L5 stage were each parasitized by a different *C. typhae* female. *S. nonagrioides* larvae used for this experiment came either from the wild (collected in France at Sorbets and Coublucq) or from a strain reared at EGCE since 2010 which initially comes from the southwest of France and which is refreshed yearly with French individuals from the wild. In order to maximize the number of resistant *S. nonagrioides*, we chose to parasitized them with the *C. typhae* strain “Makindu”, which shows low virulence on French *S. nonagrioides* larvae, with about 60% of individuals resisting to parasitism (Benoist et al., 2017). This “Makindu” wasp strain is reared at EGCE since 2015, but it initially comes from 10 individuals who emerged from *S. nonagrioides* collected in 2010-2011 from the wild in the Makindu locality (2.278S, 37.825E; South-East Kenya). An isofemale line was then produced and maintained by parasiting at each generation three-weeks-old larvae from the Makindu *S. nonagrioides* strain, on which the wasp virulence is higher than on the french strain used for the experiments of this article (Benoist et al., 2020).

Oviposition events were all confirmed by visual observation. Here, 37 larvae (59.7%) resisted to parasitism, *i.e.* no wasp pupae emerged from them and they reached adult stage. We obtained a total of 20 males and 17 females, but because of important delays in the development of some

individuals, we were able to form 15 couples only. A total of 13 females involved in these couples produced some eggs (sometimes less than 10 eggs), but most of these eggs were either unfertilized or non-viable. Numerous and viable eggs leading to the development of offspring could be recovered for only five females (Suppl. Fig. 1). The names of these F1 offspring refer to their mothers: F1-Adult1, F1-Adult2, F1-Adult3, F1-Adult4, and F1-Adult5. We let all these five progenies develop until the larval stage L3/L4 and we sacrificed them by freezing at -80°C. We then randomly picked a total of 500 caterpillars for the experiment. For technical reasons (high mortality of eggs in some progenies because of mold in our rearing tubes), we obtained less than 100 caterpillars for some couples, impeding us to sample identical numbers of individuals from each crossing. We ended up with 110 offspring individuals for F1-Adult1, 80 for F1-Adult2, 115 for F1-Adult3, 70 for F1-Adult4, and 125 for F1-Adult5 (Suppl. Fig. 1).

A piece from each of these caterpillars was then cut with a razor blade and weighed. Pieces of equal weight were pooled five by five for each set of offspring, *i.e.* 100 pools of 5 offspring. The rest of the bodies were individually frozen at -80°C to be able to identify which individuals would bear the eventual BV integration identified in pools of F1 (see below). We extracted DNA from the 100 pools with the kit DNeasy Purification (QIAGEN). We then pooled these DNA extractions 20 by 20 with equimolar ratio, leading to five libraries composed of 100 caterpillars each (Suppl. Fig. 1): libraries PoolF1-A to PoolF1-E. We sub-contracted Novogen to build a paired-end library (2 x 150 bp; insert size = 350 bp) for each of these five libraries. Each library was then sequenced on an Illumina platform to

produce a targeted amount of 100 Gbp each. In addition, the DNA of two resistant *S. nonagrioides* adults was also sequenced. One of these adults (Adult-1) is a female who produced offspring (progeny included in this experiment under the name F1-Adult1), whereas the other (Adult-6) is a female who did not produce any offspring. DNA was extracted from the thorax of these two resistant *S. nonagrioides* females, using the DNeasy Purification kit (QIAGEN) after their natural death. Library construction and sequencing was done by Novogen following the same protocol as the one used for the five other libraries, except that the targeted number of bases produced was 50 Gbp (*i.e.*, a depth of about 50X on the *S. nonagrioides* genome).

Parasitism of non-target species

C. typhae females from the strain “Kobodo” reared at EGCE since 2019, also as an isofemale line but from six parasitized caterpillars collected in the field at Kobodo locality (0.679S, 34.412E; West Kenya), were set to parasitize two caterpillars from the bulrush wainscot *Nonagria typhae* (Lepidoptera: Noctuidae), five from the Webb's wainscot *Globia sparganii* (Lepidoptera: Noctuidae), and eight from the Reed Veneer *Chilo phragmitella* (Lepidoptera: Crambidae). Each of these caterpillars was parasitized by a different wasp, and ovipositon was confirmed by visual observation. These non-natural host species were sampled in the wild in different regions in France in June 2020. The caterpillars of *N. typhae* and *G. sparganii* were collected in Saint Michel en l’Herm (46.354N, 1.260W), and the caterpillars of *C. phragmitella* in Sacy-le-Grand (49.347N, 2.537E). *C. typhae* larvae emerged from four out of these 15 parasitized non-

natural hosts, all from *C. phragmitella*. Irrespective of whether parasitism was successful or not, all parasitized individuals died between five and 26 days post parasitism, all at late larval or pre-pupal stage. Using the kit NucleoBond AXG20 (Macherey-Nagel), we extracted DNA from the whole body of three larvae, from which no wasp larvae emerged: one *N. typhae* larva, one *G. sparganii* larva, and one *C. phragmitella,* which died 17, 19, and 25 days post-parasitism, respectively. We sub-contracted Novogen to build a paired-end library (2 x 150 bp; insert size = 350 bp) for these three samples for sequencing on an Illumina platform to produce a targeted amount of 100 Gb each.

Genome assemblies

The genome of *C. typhae* and *S. nonagrioides* were retrieved from GenBank under accession numbers JAAOIC000000000 (Muller, Chebbi, et al., 2021) and JADWQK000000000 (Muller, Ogereau et al., 2021). The genomes of *N. typhae, G. sparganii,* and *C. phragmitell*a were *de novo* assembled for this study with the following procedure: (i) raw Illumina reads produced from parasitized larvae were mapped against the *C. typhae* genome with Bowtie2 in order to remove all reads from *C. typhae*, (ii) the filtered reads were assembled with MaSuRCA v4.0.1 with all parameters set to default, except for jellyfish hash size (JF_SIZE = 20000000000), and we activated USE_LINKING_MATES since we did not use any long-reads, (iii) any trace of *C. typhae* genome remaining in the assembly were identified by similarity search (blastn) between the newly assembled genomes and *C. typhae* genome. We filtered out from the assemblies all contigs with more than 99% identity with a *C. typhae* contig

on more than 90% of its length. This last step led to remove four contigs from *N. typhae,* none from *G. sparganii*, and 24 from *C. phragmitella*. We also ran purge_dups on *C. phragmitella*, whose assembly initially harbored 9.5% of duplicated BUSCO (Guan et al., 2020). We assessed the quality of these three new genomes with Busco v5.0.0 using 5286 lepidopterean core genes.

Sequencing data processing

To search for the presence of BV sequences integrated into *G. sparganii*, *N. typhae*, *C. phragmitella* and *S. nonagrioides* genomes, we developed the workflow WorkflowBowBlast (github/HeloiseMuller/WorkflowBowBlast). This workflow proceeds all the data the same way as described in Muller, Chebbi, et al. (2021), in which we characterized and quantified the genome-wide patterns of *C. typhae* BV (CtBV) integration in various tissues of parasitized *S. nonagrioides*. We activated all the tools of the workflow (integrity with fastqc v.0.11.8, trimmomatic v.0.38, bowtie2 v2-2.3.4.2, coverage with bedtools v.2.26.0 and blastn v.2.6.0) with all default parameters. We used the appropriate lepidopteran genomes (*S. nonagrioides*, *N. typhae*, *G. sparganii,* or *C. phragmitell*a) for the first reference genome and *C. typhae* for the second optional genome. For the analyses, we used two outputs of WorkflowBowBlast. First, we used the bedtools genomecov output of WorkflowBowBlast that indicates the sequencing depth at each position of the *C. typhae* genome. We then used CovWindows (github/HeloiseMuller/CovWindows) to get sequencing depth on each CtBV DNA (corresponding to any of the three CtBV forms) from this file. WorkflowBowBlast also automatically gives the average

sequencing depth on both genomes, and the percentage of the genomes covered at least once by Illumina reads. Second, we used the two blastn outputs of WorkflowBowBlast, one on each genome (Lepidoptera and wasp), to identify chimeric reads involving *C. typhae* proviral segments and the lepidopteran DNA using an R pipeline (Peccoud et al., 2018). This pipeline, specifically adapted to find chromosomally integrated BV in Muller, Chebbi, et al. (2021), identifies a read as chimeric if (i) at least 16 bases align only on the first reference genome, and a minimum of 16 other bases align only on the second reference genome; (ii) less than 10% of the read length maps to neither reference genome; (iii) no more than 20 bases align simultaneously on both reference genomes; and (iv) no more than 5 bases are inserted between the two genomes at the integration point. The two latter filters imply that aligned read regions are allowed to overlap by up to 20 bp or to be separated by at most 5 bp. After this pipeline, we filtered out chimeric reads resulting from PCR duplicates, except when comparing the number of chimeric reads to the sequencing depth, since PCR duplicates are not filtered out in the latest.

In addition to our ten samples (five libraries of 100 pooled *S. nonagrioides* offspring, two adult *S. nonagrioides* females, one *G. sparganii* larva, one *N. typhae* larva, and one *C. phragmitella* larva), we also processed in the exact same way two samples that we generated in earlier studies and used here for comparison: Caterpillar and Control. Both samples were sequenced in the same conditions as our ten samples. Caterpillar corresponds to a *S. nonagrioides* caterpillar parasitized by *C. typhae* and sequenced seven days post parasitism, for which no wasp larvae were

present (sample called “whole body” in the original publication (Muller, Chebbi, et al., 2021)). This sample was included to compare the number of CtBV chromosomal integrations in the natural host to those in the non-natural hosts. Control corresponds to the Illumina trimmed reads used to assemble the genome of non-parasitized *S. nonagrioides* (Muller, Ogereau, et al., 2021). We included this last sample in order to estimate the proportion of reads that would map on *C. typhae* genome when *S. nonagrioides* did not undergo any contact with C*. typhae*. In addition, this control also allowed us to determine that our pipeline returned no false positive *S. nonagrioides-C. typhae* chimeras. All analyses and figures were done with R v4.0.5.

Estimations and normalization of integration events

Integration events of CtBV DNA into host genomes were counted for each sample, as each position in the host genome supported by at least one read (*i.e.*, CtBV-host junctions supported by more than one read were counted as one integration event). We estimated the average number of integration events per haploid genome with IPMH*G/k, where IMPH is the normalized number of Integration events Per Million of Host reads, G is the genome size of the host in megabases (1,021Mb for *S. nonagrioides*), and k is the read length (150bp).

When assessing the transmission of CtBV from resistant parasitized moths to their offspring, we did not sequence 100% of the 1000 gametes we sampled (we sampled 500 offspring, and each inherited 2 gametes from their resistant parents). Thus, we had to estimate the number of gametes we sequenced (noted S). To do this, we calculated an upper limit, which is

the maximum number of gametes we might have sequenced, with: $S=\sum_{k=1}^{n}(d_k \times p_k)$,where d is the sequencing depth in library k (n=5 libraries), and p is the proportion of the genome sequenced on average in one offspring for library k. To estimate p, we randomly sampled one hundredth reads of the library (because each library contains 100 offspring), and we calculated the proportion of the host genome covered when looking only at this subset of reads. This proportion is an estimation of the coverage of one of the 100 offspring of the library. We repeated this sampling ten times and we calculated the average, *i.e.* the proportion of genome sequenced for one individual on average in library k (p_k).

Results

Wasp cells do not persist in adult *S. nonagrioides* surviving to parasitism by *C. typhae*

In a previous study, we showed that seven days after a female *C. typhae* wasp oviposited into larvae of its host *S. nonagrioides*, CtBV DNA persisted under three forms in parasitized caterpillars: the circular form, the chromosomally integrated form, and the proviral segments located in the wasp genome of wasp teratocyte cells present in the caterpillars (Muller, Chebbi, et al., 2021). Teratocytes are cells released from the wasp embryonic membranes into the host when eggs hatch. They are involved in various functions, including suppressing the host immune response (Buron & Beckage, 1997). Since insects undergo important cellular changes during metamorphosis (Tettamanti & Casartelli, 2019), we wondered whether some CtBV DNA and/or other wasp DNA can persist at

the adult stage in host individuals surviving to parasitism, in one form or another. Addressing this question is important to assess the extent to which transmission of germline-integrated CtBV and/or free CtBV to host offspring can occur. For this end, we Illumina-sequenced the whole genome of two adult *S. nonagrioides* out of the 37 that resisted to parasitism in our experiment. Both sequenced adult *S. nonagrioides* were female, but while Adult-1 produced many offspring, Adult-6 did not produce any.

To assess the extent to which wasp cells may persist in adult hosts resisting to parasitism, we looked whether the wasp genome was covered by the Illumina reads. For this, we mapped trimmed Illumina reads obtained from the two *S. nonagrioides* females on the whole genomes of both *C. typhae* and *S. nonagrioides*. Average sequencing depths over the *S. nonagrioides* genome were in the targeted range (41X and 49X) and 97% of the moth genome was sequenced at least once for the two individuals (Fig. 2). By contrast, average sequencing depths over the *C. typhae* genome were very low (0.5X and 0.86X) and reads mapped to only a small fraction of the wasp genome (<1% and 7.4%) (Fig. 2). To assess whether these sequencing depth and coverage, although low, indicate the presence of residual amounts of wasp cells, we compared these statistics to a control dataset corresponding to trimmed Illumina reads generated in an earlier study aiming at sequencing the genome of non-parasitized *S. nonagrioides* individuals that were never in contact with *C. typhae (Muller, Ogereau, et al., 2021).* For this dataset, we obtained an average sequencing depth of 0.43X over the *C. typhae* genome with less than 1%

of the wasp genome being covered at least once. The average sequencing depth on the genome of *S. nonagrioides* was 38.9X. This control indicates that in a *S. nonagrioides* individual devoid of any *C. typhae* DNA, we can expect an average sequencing depth on the latter at about 1.1% of the former, and less than 1% of the wasp genome is expected to be covered at least once. For Adult-1, the average depth over the *C. typhae* genome was 1.23% of the one on *S. nonagrioides* and less than 1% of the wasp genome was covered, which is close to what we obtained for the control dataset. This suggests that teratocytes and other wasp cells were absent in this individual. However, Adult-6 clearly had higher amounts of wasp DNA than expected in the absence of *C. typhae* DNA, with an average depth on the *C. typhae* genome at 1.77% of the one computed on the *S. nonagrioides* genome, and 7.4% of the wasp genome covered at least once. Although these figures could be interpreted as being due to the presence of small amounts of wasp cells in this *S. nonagrioides* adult individual, we found that the wasp genome regions covered by sequencing reads were very small, although *C. typhae* scaffolds are large (83% of these regions were between 100 and 300 bp). This suggests that the wasp DNA sequenced from this adult individual is highly fragmented and unlikely to originate from complete whole wasp genomes. Thus, while the origin of this wasp DNA remains to be determined, our results tend to indicate that teratocytes and other wasp cells are unlikely to persist in host adult individuals surviving to parasitism.

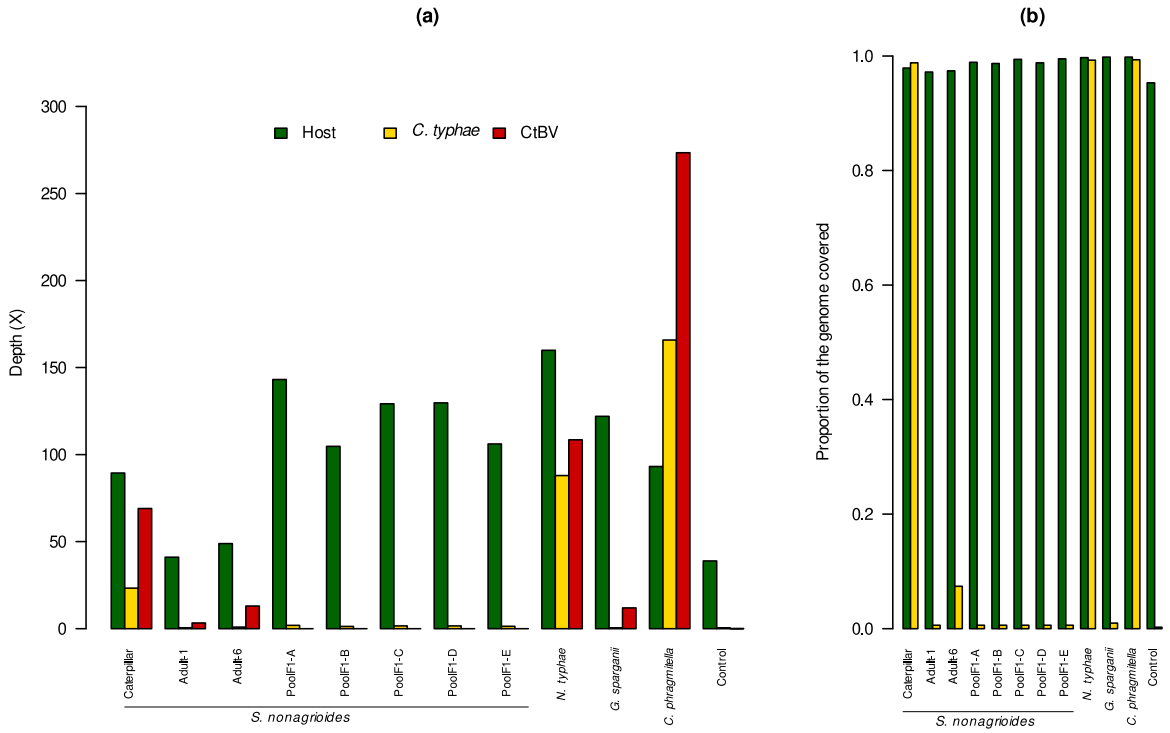


Figure 2: Sequencing depth and coverage over the genome of the wasp (Cotesia typhae), the bracovirus (CtBV) and the various moth hosts. (a) Barplot of the average sequencing depths. **(b)** Barplot of the proportion of the genomes of C. typhae and the hosts covered at least once. The green, yellow and red colors indicate the sequencing statistics over the whole genome of the host (S. nonagrioides, N. typhae, G. sparganii, or C. phragmitella depending on the samples), the whole genome of C. typhae, and the 27 CtBV DNA, respectively. The sequencing metrics are given for the ten samples produced for this study, but also for two other samples: Caterpillar is the sample corresponding to the whole body of a S. nonagrioides caterpillar sequenced seven days post-parasitism, and Control corresponds to the Illumina reads used for the assembly of the reference genome of S. nonagrioides. This sample has never been in contact with C. typhae.

Persistence of CtBV DNA in adults *S. nonagrioides* surviving to parasitism by *C. typhae*

We next focused on evaluating whether CtBV DNA may persist in adult *S. nonagrioides* surviving to parasitism, in one form or another. We found that the average sequencing depth over the 27 CtBV circles was higher than the average depth on the entire wasp genome in the two *S.*

nonagrioides adult individuals (3.7X *versus* 0.5X in Adult-1 and 12.46X *versus* 0.86X in Adult-6). This suggests that contrary to wasp cells, CtBV persisted in adult hosts, in one form or another. To assess whether some of the CtBV circles persisted in their chromosomally integrated form, we looked for chimeric reads between CtBV and *S. nonagrioides.* We found 43 and 303 chimeric reads corresponding to HIM-mediated integration of the 16 HIM-containing CtBV segments in Adult-1 and Adult-6, respectively. Thus, the 16 HIM-containing CtBV segments were able to persist at least in their integrated forms at the adult stage in both adult hosts (Fig. 3). More precisely, if we consider only independent integration events (IEs), *i.e.* counting only once chimeric reads with the exact same integration coordinate, we found 41 HIM-mediated IEs in Adult-1 and 297 HIM-mediated IEs in Adult-6. In order to compare samples, we normalized the number of IEs with the number of host reads (Integration events Per Million reads of Host, or IPMH). Adult-1 had 0.29 IPMH and Adult-6 had 1.77 IPMH (Fig. 3). From this normalized number of IEs, we can also express the average number of IE per haploid genome (see materials and methods), which allowed us to estimate 1.97 and 12.03 IEs per haploid genome for Adult-1 and Adult-6, respectively.

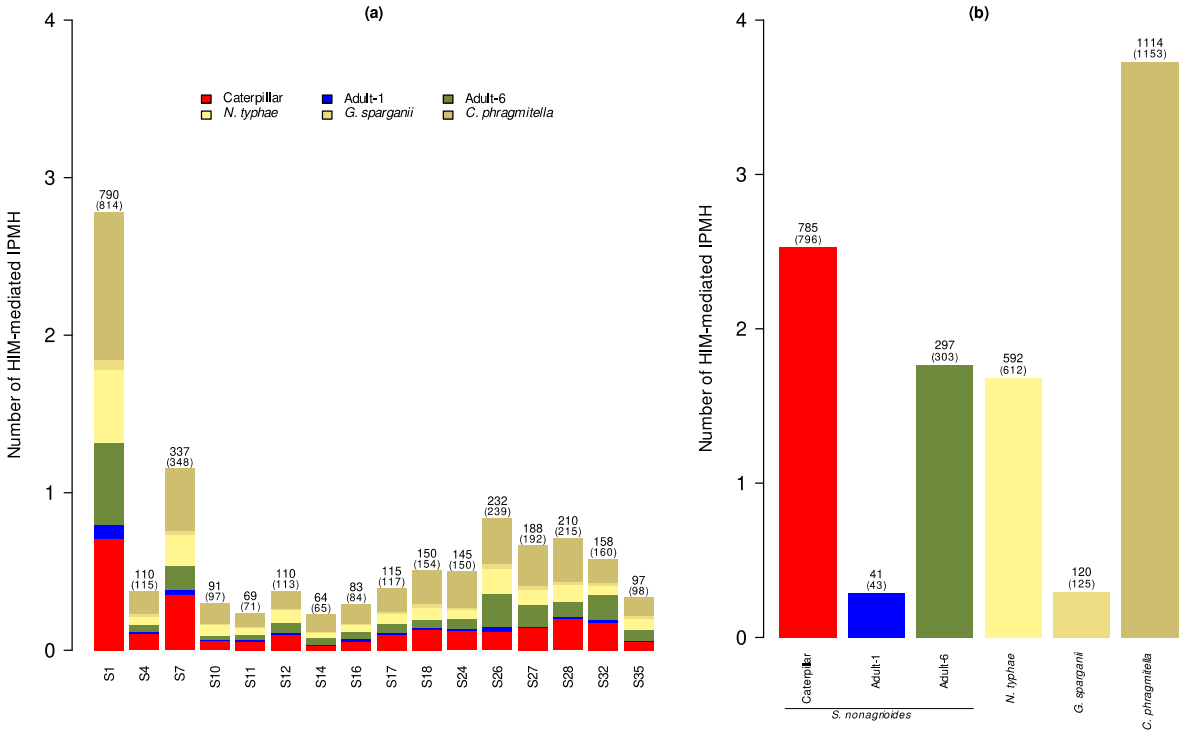


Figure 3: Number of HIM-mediated CtBV integration events per millions reads mapped on the host genome (IPMH) for each sample and CtBV circle. Absolute numbers of HIM-mediated integration events (IEs) and chimeric reads (in parentheses) are shown on top of each bar (without PCR duplicates). **(a)** Barplot comparing HIM-containing segments. **(b)** Barplot comparing samples.

We then compared the normalized number of IEs of both adult *S. nonagrioides* to the one found in the whole body of a parasitized caterpillar in which no wasp embryos were observed seven days after oviposition (called “Caterpillar”) (Muller, Chebbi, et al., 2021). We observed that the number of integrated CtBV circles can be quite high in adults, since Adult-6 is in the same order of magnitude as in a caterpillar (Fig. 3b). This raises the question of whether some CtBV circles might be integrated into the germline genome of *S. nonagrioides* and transmitted to the next generation of moth.

Before addressing this question, we investigated whether CtBV could also persist under their circular forms. Based on the current knowledge of the system, factors involved in bracovirus replication are not injected in the

host (Louis et al., 2013), so bracoviruses are only known to replicate in the calyx of the wasp, not in the caterpillar. This is why, one can expect bracovirus circles to be degraded after some times in the caterpillar. However, Muller, Chebbi, et al. (2021) found that these circles still persist in high amounts in *S. nonagrioides* caterpillars seven days post-parasitism, and Chevignon et al. (2018) found a persistence of bracovirus circles in *Manducta sexta* 12 days post-parasitism. To determine whether CtBV can persist under circular form in adult moths, we measured the sequencing depth over circles devoid of HIM, for which no or very few integrations have been found in the present study (zero and one IE in total in Adult-1 and Adult-6, respectively). While in Adult-1 all CtBV circles devoid of HIM had a sequencing depth close to the one of the *C. typhae* genome (0.5X), in Adult-6 at least two CtBV devoid of HIM were sequenced at a depth markedly higher than the *C. typhae* genome (0.86X): CtBV S20/33 (4.8X) and S23 (5.9X) (see the two red diamonds in Fig. 4b). Although limited, the quantity of persisting S20/33 and S23 is as high as that of three HIM-containing CtBV (S10, S11, and S14), indicating that at least some CtBV circles can persist up to the adult stage of the host and during its entire lifespan (Adult-6 died 46 days post-oviposition).

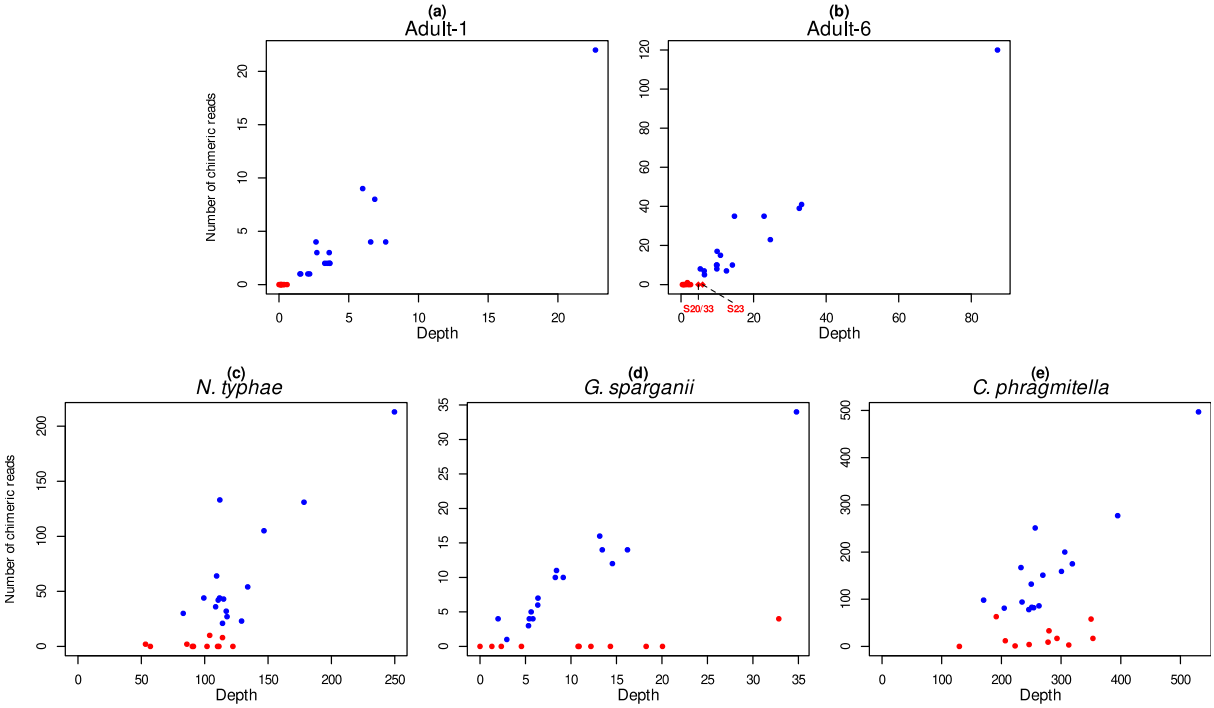


Figure 4: Sequencing depth over CtBV DNA as a function of the number of chimeric reads. Since all reads are taken into account to calculate the sequencing depth, all chimeric reads are also taken into account in this plot: the ones that result from a HIM-mediated integration or not, and also the PCR duplicates. Each dot shows a CtBV (whose depth can result from sequencing its proviral segment, its circular form or its integrated form), in blue for HIM-containing CtBV, and in red for CtBV devoid of HIM. Results are shown by sample, with S. nonagrioides Adult-1 in **(a)**, S. nonagrioides Adult-6 in **(b)** where the two red diamonds represent S20/33 and S23, N. typhae in **(c)**, G. sparganii in **(d)**, and C. phragmitella in (e).

No vertical transmission of CtBV circles by *S. nonagrioides* adults surviving to parasitism by *C. typhae*.

To assess the extent to which CtBV DNA persisting in adults *S. nonagrioides* can be transmitted vertically to the next generation of moth, we screened for the presence of CtBV in F1 individuals resulting from crosses of *S. nonagrioides* individuals that had survived parasitism. In total we Illumina-sequenced 500 offspring individuals produced by five couples. Each couple was made of a male and a female *S. nonagrioides* that both

survived parasitism. The DNA of the 500 individuals was distributed in five libraries each containing 100 individuals (Suppl. Fig. 1).

The obtained sequencing depth on the *S. nonagrioides* genome was between 105 and 143X depending on the library, and 98.7% to 99.5% of the moth genome was covered at least once in the five libraries (Fig. 2). As expected, the sequencing depth and coverage of the *C. typhae* genome were very low for the five libraries (1.2% to 1.3% of the sequencing coverage on the host, and <1% of the wasp genome covered at least once). These figures are similar to those obtained for the control dataset of reads produced from *S. nonagrioides* larvae that never encounter *C. typhae* (1.1% of the sequencing coverage on the host, and <1% of the wasp genome covered). Thus, as expected, no wasp whole genome was transmitted to moth F1 offspring individuals.

Importantly, the average sequencing depth on the 27 CtBV circles was close to zero (between 0 and 0.01X) for the five libraries of offspring DNA, suggesting that no CtBV DNA, in any forms, were sequenced (Fig. 2A). In agreement with this result, we could not identify any chimeric read between CtBV and *S. nonagrioides.* Thus, we could not identify any vertical transmission of CtBV, neither in their circular nor integrated forms, to any of the 500 offspring individuals we sequenced. This result provides a way to estimate an upper limit of the frequency at which vertical transmission of bracoviruses may be transmitted in host populations. We estimated that among the 1000 gametes transmitted to these offspring, we did not sequence more than 374 gametes (see methods). This gives a probability to detect the vertical transmission of a CtBV lower than 1/374 gametes.

This probability might have been different, if we had chosen a different sampling strategy. Here we chose to sequence 100 offspring individuals from five couples but we cannot tell whether sequencing more offspring from less couples, or sequencing less offspring from more couples could have increased our chances to find a transmission.

Chromosomal integration of CtBV in non-target species

In the context of biocontrol, it is worth to evaluate the possible consequences, in terms of BV chromosomal integration, of the introduction of an exotic biocontrol agent on autochthonous lepidopteran species that are not present in the native area of the wasp. Since such species have never interacted with the introduced wasp, they may be less permissive hosts and thus more likely to survive parasitism (Mahmoud et al., 2012). In turn, chromosomal integration of BV DNA in such non- or less- permissive species may increase the likelihood that germline-integrated BV DNA are transmitted through host generations (Drezen et al., 2017). Yet, the capacity of BV circles to undergo chromosomal integration in species not known to be naturally targeted by an endoparasitoid wasp has never been studied.

Here, we investigated whether CtBV can undergo chromosomal integration in three stemborer lepidopteran species (*Nonagria typha*e [Noctuidae], *Globia sparganii* [Noctuidae], and *Chilo phragmitella* [Crambidae]), the geographical distribution of which overlaps with that of *S. nonagrioides* in France (Galichet et al., 1992; Teder & Tammaru, 2002; Tewksbury et al., 2002). Among the two *N. typhae*, the five *G. sparganii*, and the eight *C. phragmitella* individuals that were parasitized by *C. typhae* for this

experiment, only four *C. phragmitella* yielded wasp cocoons, yet all hosts died before reaching the adult stage. For each species, we Illumina-sequenced one individual in which no wasp cocoon developed. To study CtBV infection in these three caterpillars, we first assembled their reference genomes, which are not available in public databases. Before doing so, we filtered out all reads mapping to the genome of *C. typhae* (see materials & methods). The quality of the assemblies we obtained was heterogenous, with a higher completeness for *N. typhae* (647Mb) and *G. sparganii* (972Mb) (93.6% and 82.1% of complete BUSCO, respectively) than for *C. phragmitella* (942Mb; 67.4% complete BUSCO) (Table 1). These assemblies all turned out to be sufficient to test for chromosomal integration.

Table 1. Characteristics of the three *de novo* assemblies produced in this study.

	#Contigs	Genome size assembly (Mb)	N50 (kb)	Complete BUSCO (duplicated)	Missing BUSCO
N. typhae	72,086	647	29.6	93.6% (1.9%)	2.0%
G. sparganii	304,227	972	8.5	82.1% (3.3%)	7.1%
C. phragmitella	293,788	942	4.7	67.4% (5.6%)	19.0%

Looking at sequencing depth and coverage over the whole *C. typhae* genome in the three samples revealed different patterns. While virtually no wasp whole genome was sequenced in *G. sparganii* (0.5X), it was deeply sequenced in *N. typhae* and *C. phragmitella* (88X and 166X, respectively), and almost entirely covered at least once (99.3% in the two last samples) (Fig. 2). These figures indicate the absence of wasp cells in *G. sparganii* and their presence in *N. typhae* and *C. phragmitella*. This further suggests that although no cocoons emerged from *N. typhae* and *C.*

phragmitella larvae before they died, wasp embryos were in fact able to start their development in these two species.

Interestingly, we found that as observed when sequencing a parasitized *S. nonagrioides* larvae, the total sequencing depth over CtBV DNA (in any form) was higher than that over the whole wasp genome for the three species not normally targeted by *C. typhae* (Fig. 2). Subtracting the whole wasp genome from this total CtBV DNA depth (*i.e.*, subtracting the proviral segment form) provides an estimate of the depth at which circular plus integrated forms of CtBV were sequenced: 11.8X, 20.5X and 107.6X in *G. sparganii*, *N. typhae* and *C. phragmitella*, respectively (Fig. 2a). Moreover, we identified HIM-mediated integration events (IEs) of CtBV in the three samples, with 120, 592 and 1114 IEs in *G. sparganii*, *N. typhae* and *C. phragmitella*, respectively (Fig. 3). The normalized number of integration events was the highest for *C. phragmitella,* in which it is even higher than in the natural host *S. nonagrioides* (Fig. 3). It is noteworthy that sequencing depths over non-integrated CtBV circles (circles for which we did not identify any chimeric read) are in the same range as those over integrated CtBV circles for the three non-natural host species (Fig. 4c-e). This indicates that circles can persist in non-natural host species at similar levels in their non-integrated forms and in their integrated forms, even up to 25 days post-oviposition (for *C. phragmitella*).

We finally compared the molecular signatures associated with CtBV chromosomal integration in the three non-target species to those characterized in the natural host *S. nonagrioides*. We have shown that almost all integrations in parasitized *S. nonagrioides* involve HIM-

containing circles (99%) and are HIM-mediated (96.5%) (Muller, Chebbi, et al., 2021). In non-target species, we found similar signatures in *N. typhae* and *G. sparganii*, with 98.9% and 96.8% of the integration events involving HIM-containing circles, and 92.8% and 97.6% of these integrations occurring within the HIM, respectively. The pattern was markedly different in *C. phragmitella*, with only 91.6% of the integration events involving HIM-containing circles, and only 74.3% of them occurring within the HIM. This high number of integrations outside the HIM motif in *C. phragmitella* is not due to just one or few circles integrating differently, but to a common tendency of most circles: 14 circles out of the 16 HIM-containing circles have significantly more integration events outside HIM in *C. phragmitella* than in *S. nonagrioides* (Fisher test: p.value<0.05; see Fig. 5). Such differences may be explained by the fact that host factors are involved in BV integration and that these factors are expected to diverge substantially in *C. phragmitella*, the non-target species the most distantly related to *S. nonagrioides*. This would be consistent with a recent study showing that host integrases are involved in BV integration (Wang et al., 2021).

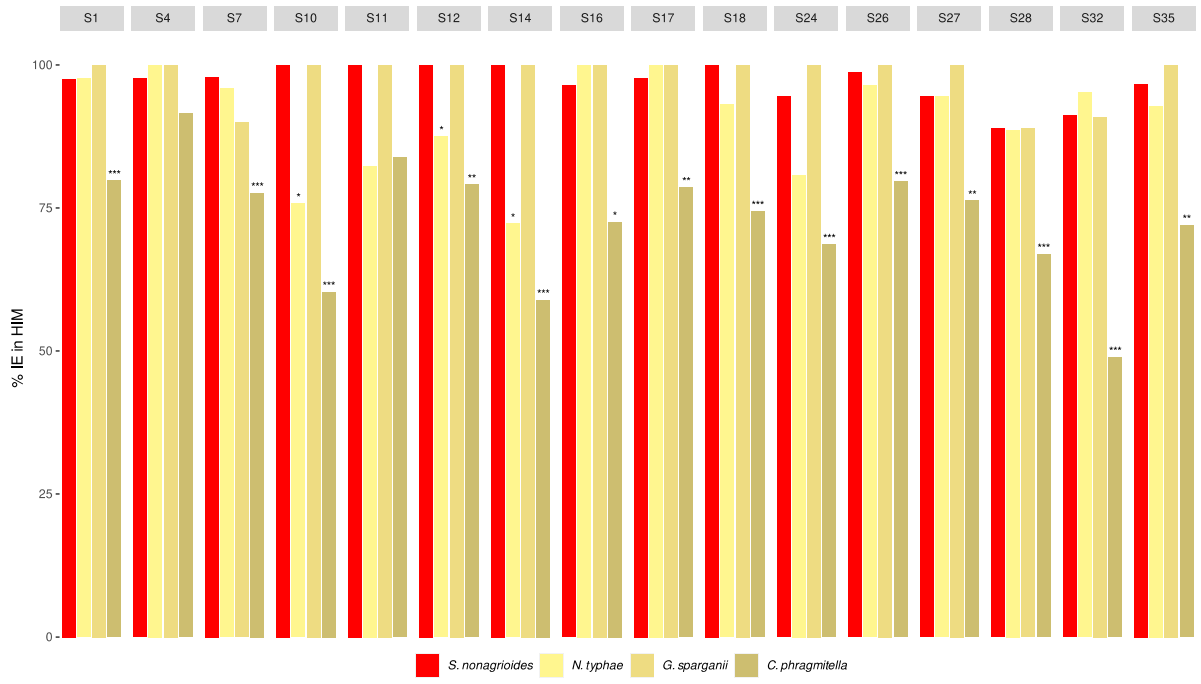


Figure 5: Percentage of HIM-mediated integration events (IEs) for the 16 HIM-containing circles. Data of the two adults S. nonagrioides and the sample Caterpillar are merged and compared to the three species not normally targeted by the wasp in the wild. Stars show whether the percentage of integration events which are HIM-mediated for the segment of interest is significantly different between the non-target species and S. nonagrioides (Fisher test; *: p.value<0.05, **: p.value<0.01; ***: p.value<0.001).

Discussion

In this study, we have characterized the presence of chromosomally-integrated and non-integrated forms of BV circles in *S. nonagrioides* adult individuals surviving parasitism by the endoparasitoid wasp *C. typhae*. We found that the two forms of CtBV can persist to the host adult stage in numbers that are similar to those identified in larval stages (Muller, Chebbi, et al., 2021), despite the tissue remodeling and other physiological changes undergone during metamorphosis by *S. nonagrioides* (Tettamanti & Casartelli, 2019). Interestingly, we found that circular forms of HIM-

containing CtBV persist in higher amounts at adult stage than CtBV circles devoid of HIM (sequenced at 12.2X and 1.2X on average respectively, see Fig. 4). In our earlier study (Muller, Chebbi, et al., 2021), we did not observe differences of persistence levels between HIM-containing CtBV circles and those devoid of HIM at seven days post oviposition (their sequencing depths were similar). The present results indicate that most DNA circles are lost/degraded, between seven days post-parasitism and the adult stage. A study monitoring DNA circles persistence in a permissive host from seven days post parasitism to the end of the wasp larvae development would shed additional light on segment copy number and persistence during development of wasp offspring, and therefore to some extent on the role of bracovirus circle integration. It is also striking that Adult-6, the adult with the highest number of CtBV integrations, did not produce any offspring, nor even any eggs, whereas Adult-1, which suffered less CtBV integration events, is the mother of hundreds of caterpillars (some of them being included in our analysis). We cannot conclude here on a possible link between the number of CtBV integrations and host fertility because of a lack of replicates. In the same way, it is possible that the reason why the majority of resisting adult couples (10 out of 15) did not produce any offspring is because of the effect of CtBV integration. Yet, these adults were not crossed as pools of many male and female individuals as routinely implemented in our lab, but as couples (*i.e.*, each made of one male and one female individual), which we previously noticed tends to decrease the likelihood of successful breeding, even when using non-parasitized individuals. It is thus difficult to assess whether parasitism or breeding conditions or both had an effect on

breeding success. This question certainly deserves to be addressed in the future as it would help to better understand the possible risks of transmission of integrated CtBV circles to the next host generation, and thus of horizontal transfer with possible long-term consequences on the host.

We tested for the possibility that CtBV DNA persisting in adult *S. nonagrioides* may be transmitted to F1 individuals resulting from crossings between host individuals surviving to parasitism. We were unable to identify any CtBV DNA in 500 of these F1. This indicates that the probability of transmission of CtBV DNA from one host generation to the next is lower than our detection limit. Our previous study showed that CtBV integrated in all *S. nonagrioides* tissues studied, which did not include gonads (Muller, Chebbi, et al., 2021), and we here report an average of 1.97 IEs per haploid genome in Adult-1 (the adult who produced offspring). Based on these findings, one could expect most gametes to harbor at least one IE. Nonetheless, the lack of detection of transmission we obtain in the present study means that at least a significant number of gametes do not harbor any IE. Although we found 1.97 IEs per haploid genome on average, we actually have no way to estimate the variance, that would indicate whether IEs are equitably shared by all cells or whether some cells carry most IEs. Regardless of the variance, gametes might be less subject to integration than other tissues because they may be less accessible. Another possibility is that CtBV may integrate in the germline at rates comparable to other tissues but it may impede proper gametogenesis and/or negatively impact the functionality

of gametes. Sequencing gonads of parasitized hosts would provide important information on this aspect.

Although we can only estimate an upper limit probability of transmission for resistant individuals, the probability of transmission at the population scale (including all parents, non-surviving, and sterile individuals) must be much lower. Indeed, out of the 62 parasitized caterpillars included in this study, 25 died before reaching adult stage, and 27 did not produce any offspring. Thus, the probability of transmission was equal to zero for 52 out of 62 parasitized *S. nonagrioides* in our experiment. It is well known that such transmission can occur in some species, as several cases of bracovirus-derived sequences resulting from ancient HT from wasp to lepidopterans have been characterized (Di Lelio et al., 2019; Gasmi et al., 2015; Schneider & Thomas, 2014). Yet, it is interesting to note that in spite of presumably being heavily targeted by *C. typhae* and other endoparasitoid wasps in the wild since many generations, *S. nonagrioides* does not contain any trace of bracovirus sequences in its (germline) genome (Muller, Chebbi, et al., 2021). Thus, the real probability of BV transmission in host populations is likely to be much lower than the upper limit we were able to estimate here. Although our upper limit estimate will be interesting for future studies aiming at investigating this rate, it is better to consider it as not so informative in the context of biocontrol strategies. Indeed, at the scale of biocontrol, even a low probability could lead to numerous BV transmissions since a large number of *S. nonagrioides* caterpillars (and potentially other lepidopteran species not known to be targeted by *C. typhae* in the wild) would be parasitized every

year. Future experiments should aim at refining estimates of the frequency of BV transmission to better inform policy makers in the context of wasp-based biocontrol strategies, as well as to better understand what impact wasp-to-lepidopteran HT may have had on the evolution of lepidopteran genomes.

Our study also shows that CtBV circles can persist as extrachromosomal molecules and undergo chromosomal integration in parasitized larvae of three lepidopteran species (*N. typhae*, *C. phragmitella*, *G. sparganii*) that are not naturally targeted by *C. typhae*, at rates comparable to those measured in *S. nonagrioides* (the native host). This shows that CtBV cell entry and chromosomal integration can occur in a wide range of lepidopteran species, including a Crambidae which diverged at least 100 million years ago from *S. nonagrioides*. Thus, these two processes are unlikely to play a major role in shaping the host range of Braconid wasps. This is in line with a study that showed that bracovirus gene expression levels, rather than cell entry and persistence of BV DNA, are more likely to be a key factor underlying host permissiveness (Bitra et al., 2016).

The extent to which the three non-natural host species included here are permissive to *C. typhae* is under study in our laboratory. It depends on ecological, behavioral and chemical factors that condition their attractiveness and acceptance as host. Yet, the outcome of the present study – only 4 host larvae out of 15 yielded wasp cocoons – together with other unpublished preliminary results suggest that these three species are much less permissive than *S. nonagrioides*. This suggests that the capacity of BV to enter in host cells and undergo chromosomal integration may not

be used as a good proxy for evaluating wasp permissive host range. In the context of using *C. typhae* as a biocontrol agent, our results are important because they show that autochthonous lepidopteran species occurring in the area in which the wasp could be released can suffer BV chromosomal integration and thus be affected in one way or another. This is, however, considering that interaction between *C. typhae* and these other autochthonous lepidopteran species in the wild are timely and ecologically possible, which remains to be assessed. On a more basic level, the fact that BV DNA can undergo chromosomal integration in a wide range of non-natural hosts, likely to be less permissive than native hosts, supports the idea that parasitizing non-natural hosts may increase the propensity of wasp-to-host HT of BV DNA with long term consequences (Drezen et al., 2017).

Acknowledgments

Research performed in this study was funded by French National Research Agency (ANR) (grants CoteBio ANR17-CE32-0015-02 to L.K. and TranspHorizon ANR-18-CE02-0021-01 to C.G.). We thank David Ogereau and Remi Jeanette for their help with DNA extraction and insect rearing, respectively.

Data Accessibility and Benefit-Sharing

Data accessibly statement

All raw sequencing reads produced during this study are available in NCBI under accession number PRJNA816685. The two previously published datasets that we used in this study are available under accession numbers SRX9975807 (Illumina trimmed reads used for the assembly of *S.*

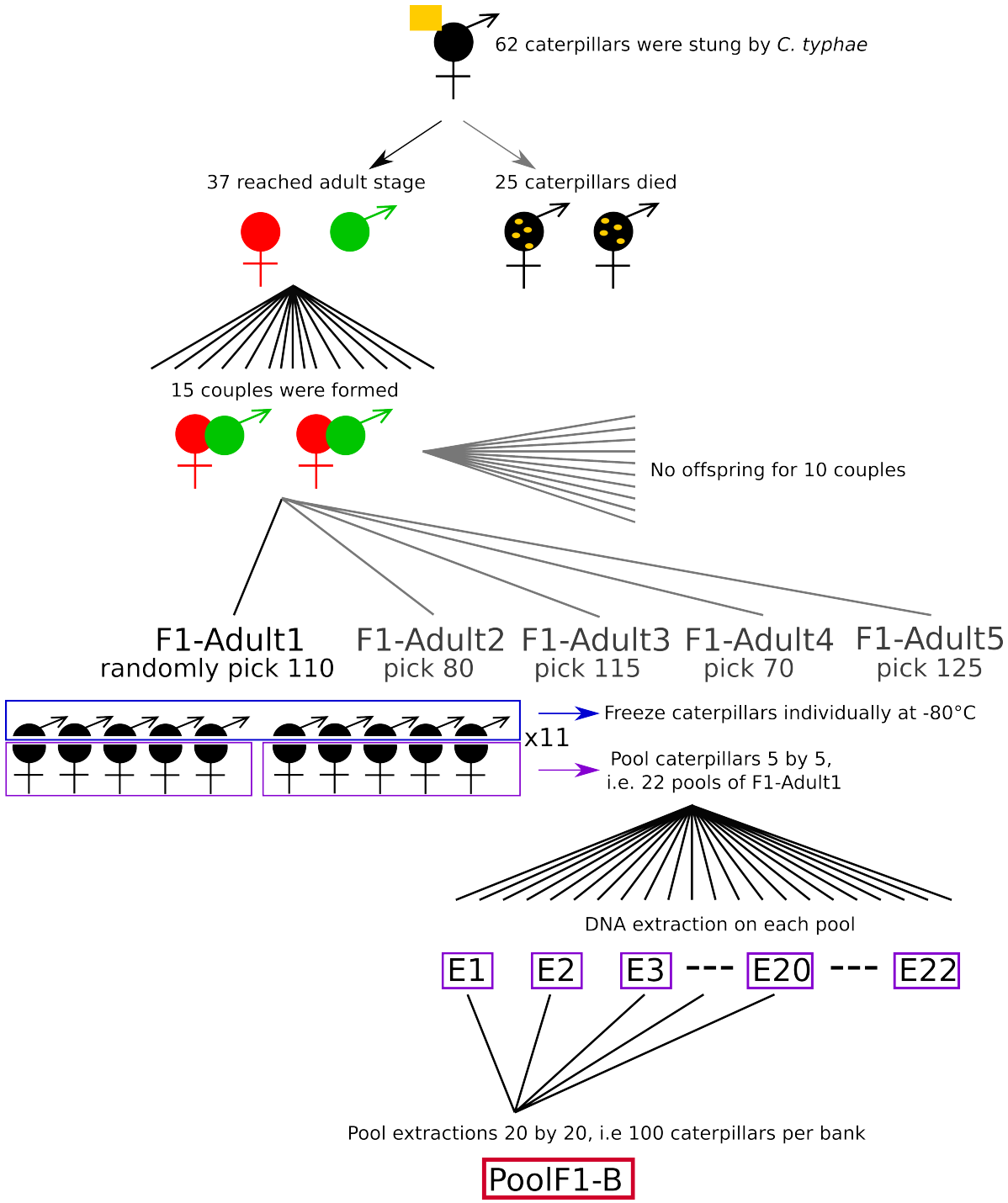
nonagrioides reference genome (Muller, Ogereau, et al., 2021)) and SAMN18533545 (“whole body” sample corresponding to a parasitized caterpillar for which no wasp larvae were present seven days post parasitism (Muller, Chebbi, et al., 2021)). Here, we called the former “Control” and the latter “Caterpillar”. The R scripts used for the analyses and to produce the figures of this project are available on GitHub (github/HeloiseMuller/Chimera).

Benefit-Sharing statement

C. typhae individuals used in this study originated from the International Centre of Insect Physiology and Ecology under the juridical framework of material transfer agreement CNRS 072057/IRD 302227/00. Access to genetic resources of the non-target lepidopteran species collected in France and benefit sharing have been done under the juridical framework of the Ministère de la Transition Ecologique (NOR: TRE2103218S/428, decision of March 26th, 2021).

Author contributions

T.F., C.G., L.K., F.M., and H.M. designed the study; T.F. and H.M. performed lab work; T.F. conducted field work and provided samples; C.H. and H.M performed data analysis with input from C.G. and F.M.; C.G. and H.M. wrote the manuscript, with input from E.H., F.M., and all other coauthors.



Supplementary Figure 1. Experimental design to obtain offspring of surviving parasitized caterpillars. In black are represented caterpillars of unidentified sex, in red the female adults, in green the male adults, and in yellow C. typhae (the square represents an adult in oviposition, and ovals represent wasp larvae). Although most moth couples produced some eggs, they lead to offspring only for 5 couples. Details of the experiment are shown for the progeny of Adult-1 (F1-Adult1), yet all the offspring were processed the same way. The DNA extractions (composed of 5 caterpillars each) were pooled 20 by 20 with equimolar ratio, leading to five libraries composed of 100 caterpillars each: library PoolF1-A to PoolF1-E. PoolF1-A was composed of 20 pools of F1-Adult5, PoolF1-B (shown on the figure as an example) of 20 pools of F1-Adult1, PoolF1-C of 16 pools of F1-Adult2 plus 1 pool of F1-Adult5 plus 3 pools of F1-Adult3, PoolF1-D of 20 pools of F1-Adult3 and PoolF1-E of 14 pools of F1-Adult4 plus 4 pools of F1-Adult5 plus 2 pools of F1-Adult1.

References

Aubin, E., El Baidouri, M., & Panaud, O. (2021). Horizontal Gene Transfers in Plants. *Life*, *11*(8), 857. https://doi.org/10.3390/life11080857

Baidouri, M. E., Carpentier, M.-C., Cooke, R., Gao, D., Lasserre, E., Llauro, C., Mirouze, M., Picault, N., Jackson, S. A., & Panaud, O. (2014). Widespread and frequent horizontal transfers of transposable elements in plants. *Genome Research*, *24*(5), 831–838. https://doi.org/10.1101/gr.164400.113

Beck, M. H., Zhang, S., Bitra, K., Burke, G. R., & Strand, M. R. (2011). The Encapsidated Genome of Microplitis demolitor Bracovirus Integrates into the Host Pseudoplusia includens. *Journal of Virology*, *85*(22), 11685–11696. https://doi.org/10.1128/JVI.05726-11

Beckage, N., & Drezen, J.-M. (Eds.). (2011). *Parasitoid Viruses: Symbionts and Pathogens*. Academic Press. https://doi.org/10.1016/C2009-0-64055-1

Benoist, R., Chantre, C., Capdevielle-Dulac, C., Bodet, M., Mougel, F., Calatayud, P. A., Dupas, S., Huguet, E., Jeannette, R., Obonyo, J., Odorico, C., Silvain, J. F., Le Ru, B., & Kaiser, L. (2017). Relationship between oviposition, virulence gene expression and parasitism success in Cotesia typhae nov. Sp. Parasitoid strains. *Genetica*, *145*(6), 469–479. https://doi.org/10.1007/s10709-017-9987-5

Benoist, R., Paquet, S., Decourcelle, F., Guez, J., Jeannette, R., Calatayud, P.-A., Le Ru, B., Mougel, F., & Kaiser, L. (2020). Role of egg-laying behavior, virulence and local adaptation in a parasitoid's chances of reproducing in a new host. *Journal of Insect Physiology*, *120*, 103987. https://doi.org/10.1016/j.jinsphys.2019.103987

Bézier, A., Louis, F., Jancek, S., Periquet, G., Thézé, J., Gyapay, G., Musset, K., Lesobre, J., Lenoble, P., Dupuy, C., Gundersen-Rindal, D., Herniou, E. A., & Drezen, J.-M. (2013). Functional endogenous viral elements in the genome of the parasitoid wasp Cotesia congregata: Insights into the evolutionary dynamics of bracoviruses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *368*(1626). https://doi.org/10.1098/rstb.2013.0047

Bitra, K., Burke, G. R., & Strand, M. R. (2016). Permissiveness of lepidopteran hosts is linked to differential expression of bracovirus genes. *Virology*, *492*, 259–272. https://doi.org/10.1016/j.virol.2016.02.023

Bosque-Perez, & Schulthess. (1998). *African cereal stem borers: Economic importance, taxonomy, natural enemies and control.* CAB INTERNATIONAL. https://www.cabi.org/cpc/abstract/19981108334

Boto, L. (2014). Horizontal gene transfer in the acquisition of novel traits by metazoans. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1777), 20132450. https://doi.org/10.1098/rspb.2013.2450

Buron, I. de, & Beckage, N. E. (1997). Developmental changes in teratocytes of the braconid wasp Cotesia congregata in larvae of the tobacco hornworm, Manduca sexta. *Journal of Insect Physiology*, *43*(10), 915–930. https://doi.org/10.1016/S0022-1910(97)00056-5

Chevignon, G., Periquet, G., Gyapay, G., Vega-Czarny, N., Musset, K., Drezen, J.-M., & Huguet, E. (2018). Cotesia congregata Bracovirus Circles Encoding PTP and Ankyrin Genes Integrate into the DNA of Parasitized Manduca sexta Hemocytes. *Journal of Virology*, *92*(15). https://doi.org/10.1128/JVI.00438-18

Chevignon, G., Thézé, J., Cambier, S., Poulain, J., Da Silva, C., Bézier, A., Musset, K., Moreau, S. J. M., Drezen, J.-M., & Huguet, E. (2014). Functional Annotation of Cotesia congregata Bracovirus: Identification of Viral Genes Expressed in Parasitized Host Immune Tissues. *Journal of Virology*, *88*(16), 8795–8812. https://doi.org/10.1128/JVI.00209-14

Danchin, E. G. J., Guzeeva, E. A., Mantelin, S., Berepiki, A., & Jones, J. T. (2016). Horizontal Gene Transfer from Bacteria Has Enabled the Plant-Parasitic Nematode Globodera pallida to Feed on Host-Derived Sucrose. *Molecular Biology and Evolution*, *33*(6), 1571–1579. https://doi.org/10.1093/molbev/msw041

Desjardins, C. A., Gundersen-Rindal, D. E., Hostetler, J. B., Tallon, L. J., Fadrosh, D. W., Fuester, R. W., Pedroni, M. J., Haas, B. J., Schatz, M. C., Jones, K. M., Crabtree, J., Forberger, H., & Nene, V. (2008). Comparative genomics of mutualistic viruses of

Glyptapanteles parasitic wasps. *Genome Biology*, *9*(12), R183. https://doi.org/10.1186/gb-2008-9-12-r183

Di Lelio, I., Illiano, A., Astarita, F., Gianfranceschi, L., Horner, D., Varricchio, P., Amoresano, A., Pucci, P., Pennacchio, F., & Caccia, S. (2019). Evolution of an insect immune barrier through horizontal gene transfer mediated by a parasitic wasp. *PLOS Genetics*, *15*(3), e1007998. https://doi.org/10.1371/journal.pgen.1007998

Drezen, J.-M., Chevignon, G., Louis, F., & Huguet, E. (2014). Origin and evolution of symbiotic viruses associated with parasitoid wasps. *Current Opinion in Insect Science*, *6*, 35–43. https://doi.org/10.1016/j.cois.2014.09.008

Drezen, J.-M., Josse, T., Bézier, A., Gauthier, J., Huguet, E., & Herniou, E. A. (2017). Impact of Lateral Transfers on the Genomes of Lepidoptera. *Genes*, *8*(11), 315. https://doi.org/10.3390/genes8110315

Galichet, P. F., Cousin, M., & Girard, R. (1992). Egg dormancy and synchronization of larval feeding with host plant development in three noctuid (Lepidoptera) species. *Acta Oecologica*, *13*(6), 701.

Gasmi, L., Boulain, H., Gauthier, J., Hua-Van, A., Musset, K., Jakubowska, A. K., Aury, J.-M., Volkoff, A.-N., Huguet, E., Herrero, S., & Drezen, J.-M. (2015). Recurrent Domestication by Lepidoptera of Genes from Their Parasites Mediated by Bracoviruses. *PLoS Genetics*, *11*(9). https://doi.org/10.1371/journal.pgen.1005470

Gauthier, J., Boulain, H., van Vugt, J. J. F. A., Baudry, L., Persyn, E., Aury, J.-M., Noel, B., Bretaudeau, A., Legeai, F., Warris, S., Chebbi, M. A., Dubreuil, G., Duvic, B., Kremer, N., Gayral, P., Musset, K., Josse, T., Bigot, D., Bressac, C., … Drezen, J.-M. (2021). Chromosomal scale assembly of parasitic wasp genome reveals symbiotic virus colonization. *Communications Biology*, *4*(1), 1–15. https://doi.org/10.1038/s42003-020-01623-8

Gauthier, J., Drezen, J.-M., & Herniou, E. A. (2018). The recurrent domestication of viruses: Major evolutionary transitions in parasitic wasps. *Parasitology*, *145*(6), 713–723. https://doi.org/10.1017/S0031182017000725

Gilbert, C., & Cordaux, R. (2017). Viruses as vectors of horizontal transfer of genetic material in eukaryotes. *Current Opinion in Virology*, *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. *Current Opinion in Genetics & Development*, *49*, 15–24. https://doi.org/10.1016/j.gde.2018.02.007

Gitau, C. W., Gundersen-Rindal, D., Pedroni, M., Mbugi, P. J., & Dupas, S. (2007). Differential expression of the CrV1 haemocyte inactivation-associated polydnavirus gene in the African maize stem borer Busseola fusca (Fuller) parasitized by two biotypes of the endoparasitoid Cotesia sesamiae (Cameron). *Journal of Insect Physiology*, *53*(7), 676–684. https://doi.org/10.1016/j.jinsphys.2007.04.008

Gruber, A., Stettler, P., Heiniger, P., Schümperli, D., & Lanzrein, B. (1996). Polydnavirus DNA of the braconid wasp Chelonus inanitus is integrated in the wasp's genome and excised only in later pupal and adult stages of the female. *The Journal of General Virology*, *77 (Pt 11)*, 2873–2879. https://doi.org/10.1099/0022-1317-77-11-2873

Guan, D., McCarthy, S. A., Wood, J., Howe, K., Wang, Y., & Durbin, R. (2020). Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics*, *36*(9), 2896–2898. https://doi.org/10.1093/bioinformatics/btaa025

Harwood, S. H., McElfresh, J. S., Nguyen, A., Conlan, C. A., & Beckage, N. E. (1998). Production of Early Expressed Parasitism-Specific Proteins in Alternate Sphingid Hosts of the Braconid WaspCotesia congregata. *Journal of Invertebrate Pathology*, *71*(3), 271–279. https://doi.org/10.1006/jipa.1997.4745

Herniou, E. A., Huguet, E., Thézé, J., Bézier, A., Periquet, G., & Drezen, J.-M. (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

Heringer P, Kuhn GCS. (2022). Multiple horizontal transfers of a Helitron transposon associated with a parasitoid wasp. *Mob DNA*. 13(1):20. doi: 10.1186/s13100-022-00278-y.

Kaiser, L., Fernandez-Triana, J., Capdevielle-Dulac, C., Chantre, C., Bodet, M., Kaoula, F., Benoist, R., Calatayud, P.-A., Dupas, S., Herniou, E. A., Jeannette, R., Obonyo, J., Silvain, J.-F., & Ru, B. L. (2017). Systematics and biology of Cotesia typhae sp. N.

(Hymenoptera, Braconidae, Microgastrinae), a potential biological control agent against the noctuid Mediterranean corn borer, Sesamia nonagrioides. *ZooKeys*, *682*, 105–136. https://doi.org/10.3897/zookeys.682.13016

Kaiser, L., Ru, B. P. L., Kaoula, F., Paillusson, C., Capdevielle-Dulac, C., Obonyo, J. O., Herniou, E. A., Jancek, S., Branca, A., Calatayud, P.-A., Silvain, J.-F., & Dupas, S. (2015). Ongoing ecological speciation in Cotesia sesamiae, a biological control agent of cereal stem borers. *Evolutionary Applications*, *8*(8), 807–820. https://doi.org/10.1111/eva.12260

Kankonda, O. M., Akaibe, B. D., Sylvain, N. M., & Ru, B.-P. L. (2018). Response of maize stemborers and associated parasitoids to the spread of grasses in the rainforest zone of Kisangani, DR Congo: Effect on stemborers biological control. *Agricultural and Forest Entomology*, *20*(2), 150–161. https://doi.org/10.1111/afe.12238

Keeling, P. J., & Palmer, J. D. (2008). Horizontal gene transfer in eukaryotic evolution. *Nature Reviews. Genetics*, *9*(8), 605–618. https://doi.org/10.1038/nrg2386

Li Y., Liu Z., Liu C., Shi Z., Pang L., Chen C., Chen Y., Pan R., Zhou W., Chen X.X., Rokas A., Huang J., Shen X.X. (2022). HGT is widespread in insects and contributes to male courtship in lepidopterans. *Cell*. 185(16):2975-2987.e10. 10.1016/j.cell.2022.06.014

Louda, S. M., Pemberton, R. W., Johnson, M. T., & Follett, P. A. (2003). Nontarget effects--the Achilles' heel of biological control? Retrospective analyses to reduce risk associated with biocontrol introductions. *Annual Review of Entomology*, *48*, 365–396. https://doi.org/10.1146/annurev.ento.48.060402.102800

Louis, F., Bézier, A., Periquet, G., Ferras, C., Drezen, J.-M., & Dupuy, C. (2013). The Bracovirus Genome of the Parasitoid Wasp Cotesia congregata Is Amplified within 13 Replication Units, Including Sequences Not Packaged in the Particles. *Journal of Virology*, *87*(17), 9649–9660. https://doi.org/10.1128/JVI.00886-13

Mahmoud, A. M. A., De Luna-Santillana, E. J., Guo, X., Reyes-Villanueva, F., & Rodríguez-Pérez, M. A. (2012). Development of the braconid wasp Cotesia flavipes in two Crambids, Diatraea saccharalis and Eoreuma loftini: Evidence of host developmental disruption. *Journal of Asia-Pacific Entomology*, *15*(1), 63–68. https://doi.org/10.1016/j.aspen.2011.07.007

Moyal, P., Tokro, P., Bayram, A., Savopoulou-Soultani, M., Conti, E., Eizaguirre, M., Le Rü, B., Avand-Faghih, A., Frérot, B., & Andreadis, S. (2011). Origin and taxonomic status of the Palearctic population of the stem borer Sesamia nonagrioides (Lefèbvre) (Lepidoptera: Noctuidae). *Biological Journal of the Linnean Society*, *103*(4), 904–922. https://doi.org/10.1111/j.1095-8312.2011.01666.x

Muller, H., Chebbi, M. A., Bouzar, C., Périquet, G., Fortuna, T., Calatayud, P.-A., Le Ru, B., Obonyo, J., Kaiser, L., Drezen, J.-M., Huguet, E., & Gilbert, C. (2021). Genome-Wide Patterns of Bracovirus Chromosomal Integration into Multiple Host Tissues during Parasitism. *Journal of Virology*, *95*(22), e00684-21. https://doi.org/10.1128/JVI.00684-21

Muller, H., Ogereau, D., Da-Lage, J.-L., Capdevielle, C., Pollet, N., Fortuna, T., Jeannette, R., Kaiser, L., & Gilbert, C. (2021). Draft nuclear genome and complete mitogenome of the Mediterranean corn borer, Sesamia nonagrioides, a major pest of maize. *G3 (Bethesda, Md.)*. https://doi.org/10.1093/g3journal/jkab155

Ono, R., Yasuhiko, Y., Aisaki, K., Kitajima, S., Kanno, J., & Hirabayashi, Y. (2019). Exosome-mediated horizontal gene transfer occurs in double-strand break repair during genome editing. *Communications Biology*, *2*(1), 1–8. https://doi.org/10.1038/s42003-019-0300-2

Pasquier-Barre, F., Dupuy, C., Huguet, E., Monteiro, F., Moreau, A., Poirié, M., & Drezen, J.-M. (2002). Polydnavirus replication: The EP1 segment of the parasitoid wasp Cotesia congregata is amplified within a larger precursor molecule. *Journal of General Virology*, *83*(8), 2035–2045. https://doi.org/10.1099/0022-1317-83-8-2035

Peccoud, J., Lequime, S., Moltini-Conclois, I., Giraud, I., Lambrechts, L., & Gilbert, C. (2018). A Survey of Virus Recombination Uncovers Canonical Features of Artificial Chimeras Generated During Deep Sequencing Library Preparation. *G3: Genes, Genomes, Genetics*, *8*(4), 1129–1138. https://doi.org/10.1534/g3.117.300468

Peccoud, J., Loiseau, V., Cordaux, R., & Gilbert, C. (2017). Massive horizontal transfer of transposable elements in insects. *Proceedings of the National Academy of Sciences*, *114*(18), 4721–4726. https://doi.org/10.1073/pnas.1621178114

Polaszek, A., & Walker, A. K. (1992). The Cotesia flavipes species-complex: Parasitoids of cereal stem borers in the tropics. *Proc. 4th Eur. Workshop Insect Parasitoids, F. Bin (Ed.). Perugia 1991. Redia 74,3 Appendix*, 335–341.

Sands, D. P. A., & Driesche, R. G. V. (2000). *Evaluating the Host Range of Agents for Biological Control of Arthropods: Rationale, Methodology and Interpretation*. 15.

Savary, S., Beckage, N., Tan, F., Periquet, G., & Drezen, J. M. (1997). Excision of the polydnavirus chromosomal integrated EP1 sequence of the parasitoid wasp Cotesia congregata (Braconidae, Microgastinae) at potential recombinase binding sites. *Journal of General Virology*, *78*(12), 3125–3134. https://doi.org/10.1099/0022-1317-78-12-3125

Schneider, S. E., & Thomas, J. H. (2014). Accidental Genetic Engineers: Horizontal Sequence Transfer from Parasitoid Wasps to Their Lepidopteran Hosts. *PLOS ONE*, *9*(10), e109446. https://doi.org/10.1371/journal.pone.0109446

Simion, P., Narayan, J., Houtain, A., Derzelle, A., Baudry, L., Nicolas, E., Arora, R., Cariou, M., Cruaud, C., Gaudray, F. R., Gilbert, C., Guiglielmoni, N., Hespeels, B., Kozlowski, D. K. L., Labadie, K., Limasset, A., Llirós, M., Marbouty, M., Terwagne, M., … Van Doninck, K. (2021). Chromosome-level genome assembly reveals homologous chromosomes and recombination in asexual rotifer Adineta vaga. *Science Advances*, *7*(41), eabg4216. https://doi.org/10.1126/sciadv.abg4216

Strand, M. R., & Burke, G. R. (2014). Polydnaviruses: Nature's Genetic Engineers. *Annual Review of Virology*, *1*(1), 333–354. https://doi.org/10.1146/annurev-virology-031413-085451

Teder, T., & Tammaru, T. (2002). Cascading effects of variation in plant vigour on the relative performance of insect herbivores and their parasitoids. *Ecological Entomology*, *27*(1), 94–104. https://doi.org/10.1046/j.0307-6946.2001.00381.x

Tettamanti, G., & Casartelli, M. (2019). Cell death during complete metamorphosis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *374*(1783), 20190065. https://doi.org/10.1098/rstb.2019.0065

Tewksbury, L., Casagrande, R., Blossey, B., Häfliger, P., & Schwarzländer, M. (2002). Potential for Biological Control of Phragmites australis in North America. *Biological*

Control, *23*(2), 191–212. https://doi.org/10.1006/bcon.2001.0994

Van Etten, J., & Bhattacharya, D. (2020). Horizontal Gene Transfer in Eukaryotes: Not if, but How Much? *Trends in Genetics*, *36*(12), 915–925. https://doi.org/10.1016/j.tig.2020.08.006

Wallau GL, Capy P, Loreto E, Le Rouzic A, Hua-Van A. (2016). VHICA, a New Method to Discriminate between Vertical and Horizontal Transposon Transfer: Application to the Mariner Family within Drosophila. *Mol Biol Evol*. 33(4):1094-109. doi: 10.1093/molbev/msv341. Epub 2015 Dec 18.

Wang, Z., Ye, X., Zhou, Y., Wu, X., Hu, R., Zhu, J., Chen, T., Huguet, E., Shi, M., Drezen, J.-M., Huang, J., & Chen, X. (2021). Bracoviruses recruit host integrases for their integration into caterpillar's genome. *PLOS Genetics*, *17*(9), e1009751. https://doi.org/10.1371/journal.pgen.1009751

Wybouw, N., Pauchet, Y., Heckel, D. G., & Van Leeuwen, T. (2016). Horizontal Gene Transfer Contributes to the Evolution of Arthropod Herbivory. *Genome Biology and Evolution*, *8*(6), 1785–1801. https://doi.org/10.1093/gbe/evw119

Wyder, S., Blank, F., & Lanzrein, B. (2003). Fate of polydnavirus DNA of the egg–larval parasitoid Chelonus inanitus in the host Spodoptera littoralis. *Journal of Insect Physiology*, *49*(5), 491–500. https://doi.org/10.1016/S0022-1910(03)00056-8

Xia, J., Guo, Z., Yang, Z., Han, H., Wang, S., Xu, H., Yang, X., Yang, F., Wu, Q., Xie, W., Zhou, X., Dermauw, W., Turlings, T. C. J., & Zhang, Y. (2021). Whitefly hijacks a plant detoxification gene that neutralizes plant toxins. *Cell*, *184*(7), 1693-1705.e17. https://doi.org/10.1016/j.cell.2021.02.014

Zhang, H.-H., Peccoud, J., Xu, M.-R.-X., Zhang, X.-G., & Gilbert, C. (2020). Horizontal transfer and evolution of transposable elements in vertebrates. *Nature Communications*, *11*(1), 1362. https://doi.org/10.1038/s41467-020-15149-4

11 - Article n°5: Massive Somatic and Germline Chromosomal Integrations of Polydnaviruses in Lepidopterans

My contribution to this article was smaller than to the other articles of this manuscript. The first part of this work, the analyses on the somatic insertions of the ichnovirus of *Hyposoter didymator* in *Spodoptera frugiperda* was achieved by Camille Heisserer, a master student whom I co-supervised. She used the pipeline I developed in Article n°3. In this study, we found that this Ichneumonidae wasp integrates its ichnovirus circles with mechanisms which are very similar to those we reported for CtBV, a bracovirus, despite the independent origins of these two genus of polydnaviruses. This rises the question of the evolutionary mechanisms that led to such a similarity.

The second part, the search of HT from bracoviruses and ichnoviruses into lepidopteran genomes, was achieved by Clément Gilbert, and I participated in the discussions and in the design of some figures. Here, we showed that HIM-mediated integrations of polydnaviruses in the germline of Lepidoptra is widespread.

Massive Somatic and Germline Chromosomal Integrations of Polydnaviruses in Lepidopterans

Camille Heisserer,[†,1,2] Héloïse Muller,[†,1] Véronique Jouan,[3] Karine Musset,[2] Georges Periquet,[2] Jean-Michel Drezen ,[2] Anne-Nathalie Volkoff ,[3] and Clément Gilbert *[,1]

[1]Université Paris-Saclay, CNRS, IRD, UMR Évolution, Génomes, Comportement et Écologie, Gif-sur-Yvette, France

[2]UMR 7261 CNRS, Institut de Recherche sur la Biologie de l'Insecte, Faculté des Sciences et Techniques, Université de Tours, Tours, France

[3]DGIMI, INRAE, University of Montpellier, Montpellier, France

[†]These authors contributed equally to this work.

***Corresponding author:** E-mail: clement.gilbert1@universite-paris-saclay.fr.

Associate editor: Mary O'Connell

Abstract

Increasing numbers of horizontal transfer (HT) of genes and transposable elements are reported in insects. Yet the mechanisms underlying these transfers remain unknown. Here we first quantify and characterize the patterns of chromosomal integration of the polydnavirus (PDV) encoded by the Campopleginae *Hyposoter didymator* parasitoid wasp (HdIV) in somatic cells of parasitized fall armyworm (*Spodoptera frugiperda*). PDVs are domesticated viruses injected by wasps together with their eggs into their hosts in order to facilitate the development of wasp larvae. We found that six HdIV DNA circles integrate into the genome of host somatic cells. Each host haploid genome suffers between 23 and 40 integration events (IEs) on average 72 h post-parasitism. Almost all IEs are mediated by DNA double-strand breaks occurring in the host integration motif (HIM) of HdIV circles. We show that despite their independent evolutionary origins, PDV from both Campopleginae and Braconidae wasps use remarkably similar mechanisms for chromosomal integration. Next, our similarity search performed on 775 genomes reveals that PDVs of both Campopleginae and Braconidae wasps have recurrently colonized the germline of dozens of lepidopteran species through the same mechanisms they use to integrate into somatic host chromosomes during parasitism. We found evidence of HIM-mediated HT of PDV DNA circles in no less than 124 species belonging to 15 lepidopteran families. Thus, this mechanism underlies a major route of HT of genetic material from wasps to lepidopterans with likely important consequences on lepidopterans.

Key words: **polydnavirus, horizontal transfer, host–parasite relationships, insects, parasitoid wasps, evolutionary genomics.**

Introduction

Parasitoid wasps are a paraphyletic group of Hymenoptera that are classified as ectoparasites or endoparasite insects, depending on whether they develop on or within an arthropod host, respectively (Beckage and Drezen 2012). To ensure the developmental success and survival of their eggs and larvae within hosts, many parasitoid wasps use viral particles that are akin to gene or protein-delivery agents (Herniou et al. 2013). These particles are injected together with wasp's eggs in parasitized hosts, of which they alter the physiology, thus enabling the development of wasp larvae within the host body (Beckage and Gelman 2004). The structural components of these viral particles are encoded by genes originating from molecular domestication of viral genomes that were integrated into the genome of wasp's ancestors (fig. 1).

Such domestication events occurred multiple times independently in different hymenopteran lineages, yielding viral particles that differ in terms of structure, function, and content (Volkoff et al. 2010; Béliveau et al. 2015; Pichon et al. 2015; Burke 2019; Di Giovanni et al. 2020; Burke et al. 2021; Mao et al. 2022). Given their viral origin, the particles packaging DNAs encoded by parasitoid wasps have been referred to as viruses and classified in their own viral family, the polydnaviridae, or PDVs (Stoltz et al. 1984; Herniou et al. 2013). The two currently recognized genera of PDVs are the bracoviruses (BVs) and the ichnoviruses (IVs), respectively carried by wasps of the Braconidae and Ichneumonidae families that parasitize mainly lepidopteran hosts (Bézier et al. 2009; Béliveau et al. 2015; Burke et al. 2018; Volkoff and Cusson 2020). BVs result from the domestication of a nudivirus integrated in a braconid ancestor ~100 Ma (Bézier et al. 2009), and are found today in all wasps belonging to the "Microgastroid complex" estimated to contain about 50,000 species (Murphy et al. 2008). IVs derive from an unknown viral

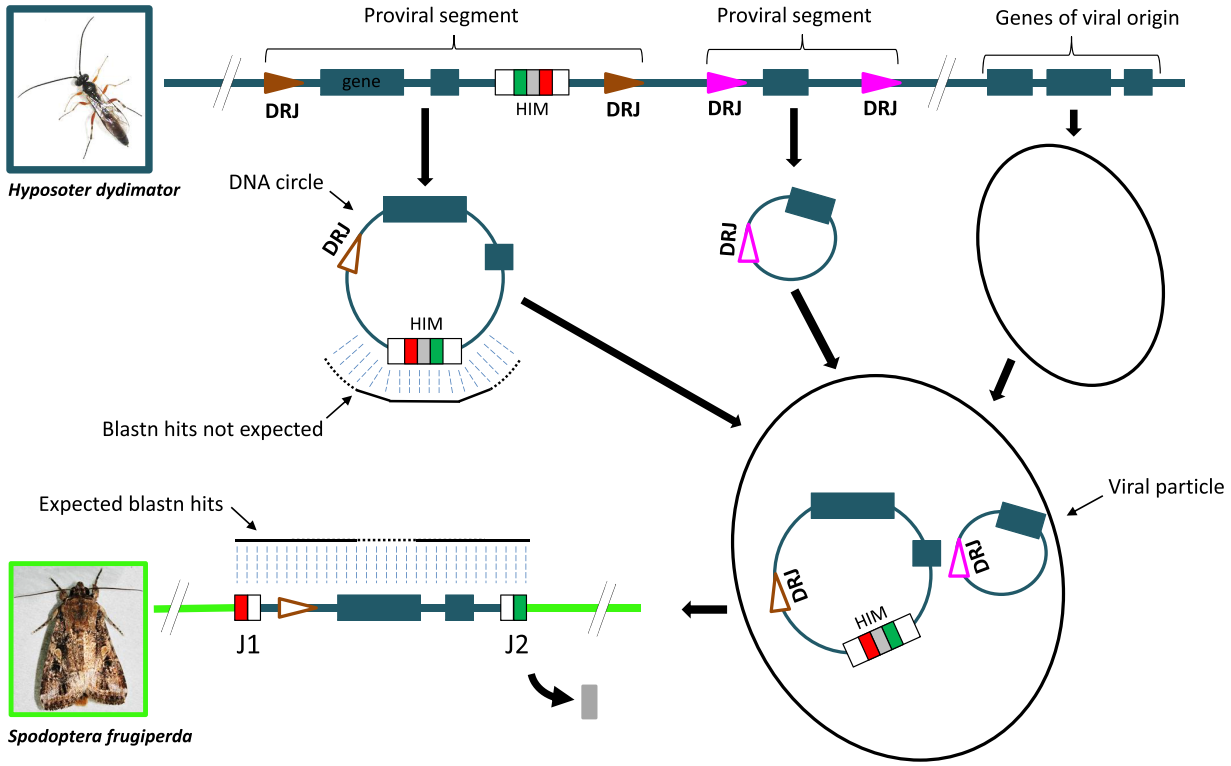


Fig. 1. Genome structure and chromosomal integration of PDVs. The genome and genes of the wasp (here *H. didymator*, Ichneumonidae) are shown in blue whereas the lepidopteran host (here *S. frugiperda*) is shown in green. The viral particle proteins of PDVs are encoded by genes of viral origin that have been coopted by parasitoid wasps. DNA circles packaged into viral particles originate from proviral segments, which are amplified and circularized in the calyx of female wasps, likely through recombination at the DRJ motifs (filled triangles). Circularization leads to the formation of a single recombined DRJ, or DRJ circle (Gauthier et al. 2021) made of a portion of the 5′ DRJ and a portion of the 3′ DRJ (empty triangles). Contrary to BVs in which DRJ motifs are similar between proviral segments, DRJ motifs differ between IVs segments, hence they are depicted with different colors (pink and brown). While ovipositing into larvae of their lepidopteran hosts, female wasps inject large amounts of PDV viral particles, which enter host somatic cells and deliver DNA circles to host cell nuclei. DNA circles possessing a HIM then undergo chromosomal integration catalyzed by wasp and host integrases (Wang et al. 2021). Integration involves a double-strand break within the HIM and linearized circles newly integrated into host DNA are bordered by junction 1 (J1 in red) and junction 2 (J2 in green) motifs. The short intervening region between J1 and J2 (in grey; about 50 bp in most circles) is lost during the process. We predicted that if HT of PDV sequences occur recurrently in lepidopteran hosts through HIM-mediated germline genome integration, few or no PDV sequence integrated in lepidopteran genomes should contain the intervening region. Thus, we should recover few or no blastn hit covering this region. By contrast, most or all blastn hits covering the HIM region should end at the J1 or J2 motif, which lie at the predicted junction with host DNA. The picture of *H. didymator* and *S. frugiperda* were taken by Marie Frayssinet, INRAE DGIMI lab.

family (Volkoff et al. 2010) and are present in two related Ichneumonidae subfamilies containing about 4,000 species. It was questioned whether their distribution in Ichneumonidae is due to one or two independent integration events (Béliveau et al. 2015) but recent phylogenetic analyses favor the second hypothesis (Santos et al. 2022).

A remarkable feature of BVs and IVs is that they both package circular DNA molecules (Volkoff and Huguet 2021; fig. 1). These DNA circles arise from amplification and excision of so-called "proviral segments" located in the wasp genome. The segments contain genes of different origins that encode virulence factors and are expressed in the parasitized host throughout the development of the wasp larvae. In braconid wasp genomes, most BV segments are arranged in tandem in large loci. For example, *Cotesia congregata* harbors 35 proviral segments, 17 of which are clustered in an ~2-Mb macrolocus located on chromosome 5 (Gauthier et al. 2021). Each of the segments is flanked by direct repeat junctions (DRJs) (filled triangles in fig. 1), sequence motifs that all contain an AGCT tetramer motif at which circularization through site-specific recombination occurs (Drezen et al. 1997; Desjardins et al. 2008; Burke et al. 2015). Ichneumonid IV segments are generally more numerous than BV segments and are highly dispersed in wasp genomes. For example, *Hyposoter didymator* harbors 57 proviral segments that are all separated by megabase-long portions of wasp genome (Legeai et al. 2020). Like BV segments, IV segments are flanked by DRJs but contrary to BVs, which segments share similar extremities, the sequence of IV DRJs are specific of each segment. Circularization is thought to occur through homologous recombination between left and right DRJs, with breakpoints located at varying positions along the DRJs (Legeai et al. 2020). For both BVs and IVs, circularization leads to the

formation of a single DRJ composed of a portion of the 5′ DRJ and a portion of the 3′ DRJ of proviral segments (empty triangles in fig. 1).

Although the different genomic organization of BV and IV proviral segments imply differences in the mechanisms regulating their amplification and circularization, DNA circles are produced in cells of the calyx, a special region of the female gonad, for both PDV types. They are then packaged into viral particles, released in the oviduct lumen and injected into the host during oviposition. Early cell line-based studies suggested that once in host cells, some DNA circles could persist as chromosomally integrated forms (fig. 1) (McKelvey et al. 1996; Volkoff et al. 2001; Gundersen-Rindal and Lynn 2003). This was confirmed using targeted Sanger and Illumina sequencing by showing that at least a subset of BV circles integrates into host hemocytes, and that integration likely involves an enzymatically regulated process through recognition of a specific motif of the DNA circles called host integration motif (HIM) (Beck et al. 2011; Chevignon et al. 2018). In agreement with this, a recent study found by functional analysis that in some lepidopteran hosts, a fraction of DNA circle integration events are catalyzed by a series of viral integrases, whereas host integrases are also involved in the integration of some circles (Wang et al. 2021a). Using bulk Illumina sequencing of parasitized Mediterranean corn borer larvae (*Sesamia nonagrioides*, Lepidoptera, Noctuidae) we showed that chromosomal integration of BV circles from *Cotesia typhae* occurs systematically in multiple host tissues and that only BV DNA circles containing HIM motifs integrate at measurable levels. Our quantitative approach also allowed us to estimate that each host haploid genome suffers between 12 and 85 BV DNA circle integration events depending on the tissue. Despite this high level of integration activity, we found that non-integrated circles also persist as much as integrated circles in parasitized hosts during at least half the development of wasp larvae (Muller et al. 2021). Recent bulk Illumina sequencing of parasitized diamondback moths (*Plutella xylostella*, Lepidoptera, Plutellidae) larvae also revealed that IV DNA circles from *Diadegma semiclausum* (DsIV) undergo massive integration into host hemocyte genomes during parasitism (Wang et al. 2021b). Four DNA circles were found to integrate via a HIM motif, as observed for BV, but contrary to BV DNA circles, IV circles devoid of HIM were also found to integrate at substantial levels through DNA double-strand breaks occurring at varying locations along the circles (Wang et al. 2021b). The finding that both IV and BV DNA circles integrate via a shared HIM-mediated mechanism despite their independent origin is remarkable and several scenarios have been proposed to explain the possible evolutionary ties between these two PDV lineages (Muller et al. 2021; Wang et al. 2021b).

In the present study, we first apply the approach developed in Muller et al. (2021) on the *C. typhae*/*S. nonagrioides* system to comprehensively characterize and quantify DNA circle integration of yet another PDV, the *Hyposoter didymator* IV (HdIV), in two tissues of parasitized fall armyworms (*Spodoptera frugiperda*, Lepidoptera, Noctuidae). *Hyposoter didymator* is a Campopleginae wasp (Ichneumonidae) showing a marked host preference for *Helicoverpa armigera* in the wild, but it is able to parasitize a large variety of other noctuid moths, including *S. frugiperda* (Frayssinet et al. 2019). Our results further underline the striking similarities in integration patterns between IVs and BVs, despite that these domesticated endogenous viruses are derived from viruses belonging to different families. We then assess whether HIM-mediated IV and BV chromosomal integration occurred in the germline of their hosts in the past by screening the genome of 775 lepidopteran species. We found sequences from IV and BV DNA circles in no less than 124 species belonging to 15 different lepidopteran families, suggesting that HIM-mediated wasp-to-host horizontal transfer (HT) of PDV DNA circles occurred recurrently during the evolutionary history of lepidopterans.

Results

Quantifying *S. frugiperda* and *H. didymator* Genomic Material

In order to characterize genome-wide patterns of HdIV integration during parasitism of the fall armyworm (*S. frugiperda*), we Illumina-sequenced whole DNA extracted from fat bodies and hemolymph of parasitized larvae. We obtained from 321 to 403 million trimmed reads depending on the tissue and time post-parasitism (p.p.) (hemolymph 24 and 72 h [H24 and H72], fat body 24 and 72 h [FB24h and FB72h]) (table 1).

The majority of these reads (from 77% to 81%) aligned to the *S. frugiperda* genome (384 Mb), which was covered almost entirely (98%) in all samples. The average sequencing depth on the *S. frugiperda* genome varied from 101× for the H72h sample to 132× for the FB24h sample (fig. 2). By contrast, only 0.3–15% of the 226-Mb *H. didymator* genome was covered depending on the sample, at very low average sequencing depths (1–2×), most likely corresponding to the DNA of *H. didymator* eggs. By contrast, the average sequencing depth of the *H. didymator* wasp regions annotated as ichnovirus proviral segments by Legeai et al. (2020) was much higher than the rest of the wasp's genome: 914× and 631× in the hemolymph (24 and 72 h p.p.) and 1,220× and 386× for the fat body (24 and 72 h p.p.) (fig. 2). This higher depth is consistent with the presence of many integrated and/or non-integrated forms of HdIV circles in the caterpillar after parasitization, since large amounts of virus particles are introduced with the parasite egg.

Quantification and Characterization of Integrated HdIV into *S. frugiperda* Somatic Genomes

To quantify chromosomal integration of HdIV circles into parasitized *S. frugiperda* cells, we searched for chimeric reads containing HdIV–*S. frugiperda* junctions (see

Table 1. Summary of the Number of Reads and Chimeric Reads Obtained from *S. frugiperda* Larvae Parasitized by *H. didymator*.

	Accession Number (SRA)	Total Number of Trimmed Reads	Reads Mapped on Caterpillar Genome	Number of Chimeric Reads	RPM on Caterpillar Genome	Number of IEs
H24h	SRX15279414	371 695 498	320 038 242 (86%)	2,861	9	2,449
H72h	SRX15279415	321 231 671	263 127 889 (82%)	4,903	19	4,398
FB24h	SRX15279416	403 147 682	346 759 613 (86%)	2,516	7	2,065
FB72h	SRX15279417	347 218 922	299 552 940 (86%)	3,325	11	2,908

Note. H, hemolymph; FB, fat body. Numbers between brackets are percentages of total sequencing reads mapped on the two reference genomes (that of *S. frugiperda* and that of *H. didymator*). IEs: Integrations events (host–wasp junctions with the same coordinate in the host genome but covered by more than one chimeric read are counted as one integration event). RPM: "number of chimeric reads per million reads mapping" on the caterpillar genome.

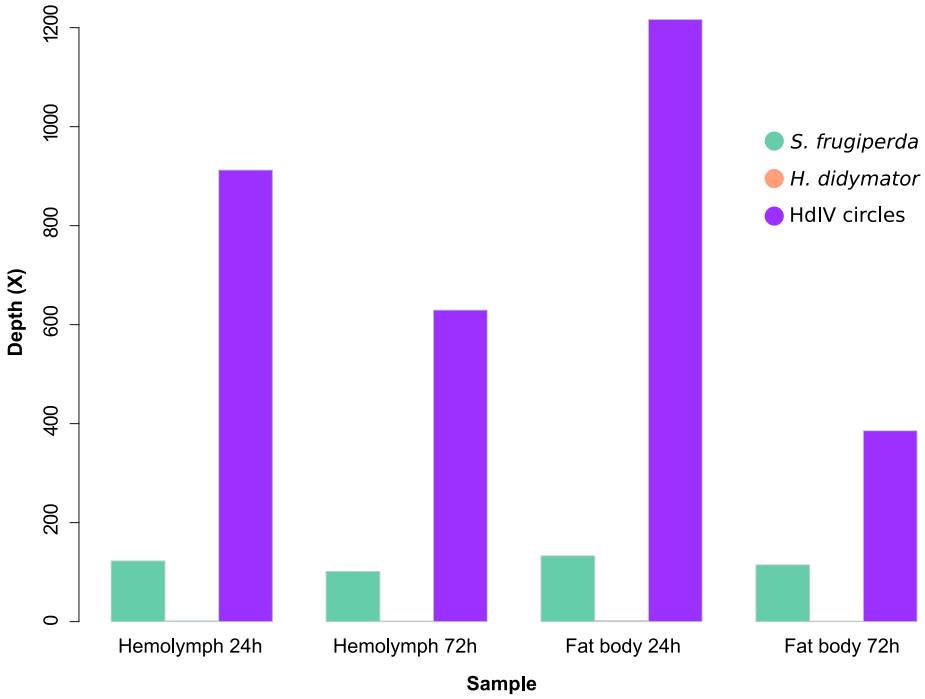


Fig. 2. Mean sequencing depth over parasitized host (*S. frugiperda*) and parasitoid wasp (*H. didymator*) genomes.

Materials and Methods). Depending on the sample, we detected from 2,516 to 4,903 chimeric reads (table 1). Even after discarding PCR duplicates, a same integration event (IE), as defined by a given set of coordinates in the HdIV and *S. frugiperda* genome, might be covered by more than one chimeric read. This may occur when a given IE is duplicated through cell division. We thus estimated the number of independent IEs by counting as one event all reads with identical or nearly identical coordinates for both genomes. We found that the vast majority of IEs were covered by only one read and in total, we counted between 2,065 and 4,398 IEs (table 1; fig. 3). The number of HdIV junctions covered by more than one reads went from 1 to 121 depending on the tissue (highest number of reads covering a junction = 3).

We found that seven out of the 57 HdIV circles had more than ten chimeric reads mapping to them in all four DNA samples, strongly suggesting these seven segments underwent chromosomal integration. The seven segments form two groups depending on whether they have generated more or less than one IE per million reads mapped on the host (*S. frugiperda*) genome (respectively, group 1: Hd12, Hd16 and Hd27 and group 2: Hd13, Hd44.2, Hd47 and Hd53) (fig. 3). Comparing normalized IEs per segment revealed a higher number of IEs in the hemolymph than in the fat body and a higher number of IEs at 72 h than at 24 h p.p. These trends hold for all seven integrated segments and are in agreement with earlier findings on BVs (Beck et al. 2007; Muller et al. 2021).

Next, we followed the rationale exposed in Muller et al. (2021) and assessed how many IEs, on average, occurred per sequenced haploid genome. To do so, we divided the relative number of IEs (IEs per million reads mapped to the host genome) for each integrated circle by the read length and we multiplied it by the *S. frugiperda* genome size. We counted a total of 18, 40, 13, and 23 IEs per cell, on average, in the Hemocyte 24 and 72 h, and the Fat body 24 and 72 h samples, respectively.

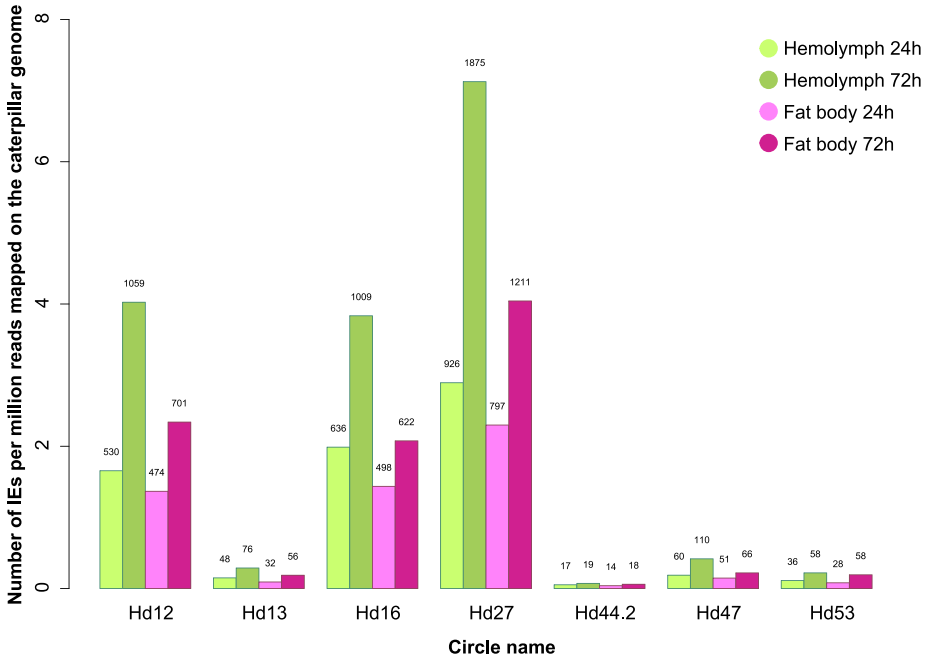


Fig. 3. Number of HdIV chromosomal IEs per HdIV circle in parasitized *S. frugiperda* larvae. Integration events (host–wasp junctions with the same coordinate in the host genome but covered by more than one chimeric read are counted as one IE).

Mechanisms Underlying Integration of HdIV Circles

To assess whether HdIV circle integration involves recombination within HIM as observed for BVs, and/or other mechanisms, we generated plots of chimeric read depths along all seven circles found integrated into *S. frugiperda* genomes (fig. 4; supplementary fig. S1, Supplementary Material online). The vast majority of chimeric reads (96.3–100% depending on the circle) map to a narrow region (in white in fig. 4) and most of the positions along each HdIV circle are not mapped by any chimeric reads. Depending on the circle, the length of the mapped region is between 92 and 120 bp. We arbitrarily delineated these specific regions in the HdIV circles following Muller et al. (2021), by identifying the two positions along the circles with the highest number of chimeric reads and by selecting 30 bp upstream of the left-most one and 30 bp downstream of the right-most one. We then verified that these regions align to the recently described HIM motifs in various IVs (Wang et al. 2021b) (fig. 4). Zooming on the HIM also confirmed the presence of two peaks of chimeric reads, as observed in BVs (Muller et al. 2021) and corresponding to the J1 and J2 motifs, at which double strand breaks most often occur during linearization and integration of the circles (Chevignon et al. 2018) (fig. 4; supplementary fig. S1, Supplementary Material online). Thus, our results show that as for BVs, most if not all chromosomal integrations of HdIV circles are mediated by ichnovirus HIMs.

To further characterize the mechanism involved in the integration within the HIM regions and in particular whether homology between viral and host sequences could be involved in the integration process, we studied the distribution of microhomology lengths detected at the junction between the HdIV and the *S. frugiperda* genome. We only investigated chimeric reads mapping to the HIM because the remaining chimeric reads represent only a very small fraction of all chimeric reads (0.8%). As previously done in our study on *C. typhae* BV (Muller et al. 2021), we studied separately the major part of chimeric reads which map to the J1 and J2 motifs from those mapping outside of these motifs (but still within the HIM). We artificially generated chimeric reads following the method described in Peccoud et al. (2018) to compare the number of observed microhomology lengths with the number of expected microhomology lengths if junctions occurred randomly between HdIV HIMs and the *S. frugiperda* genome. For chimeric reads mapping to the J1 or J2 motifs, the analysis revealed a strong excess of 2- and 3-bp microhomologies and a slight excess of 4-bp microhomologies (fig. 5). The pattern was similar for chimeric reads mapping outside J1 and J2, though the excess of 3-bp microhomologies is less marked and that of 4-bp microhomologies no longer present. Interestingly, we observe a strong depletion of 0- and 1-bp microhomologies compared to the pattern expected by chance. This indicates that junction involving blunt-ended host-wasp sequences or 1-bp microhomology between them almost never occur.

Persistence of Integrated versus Non-integrated Circles

We evaluated the quantity of integrated versus non-integrated HdIV circles during parasitism, as approximated by sequencing depth. We found that except for circle Hd27

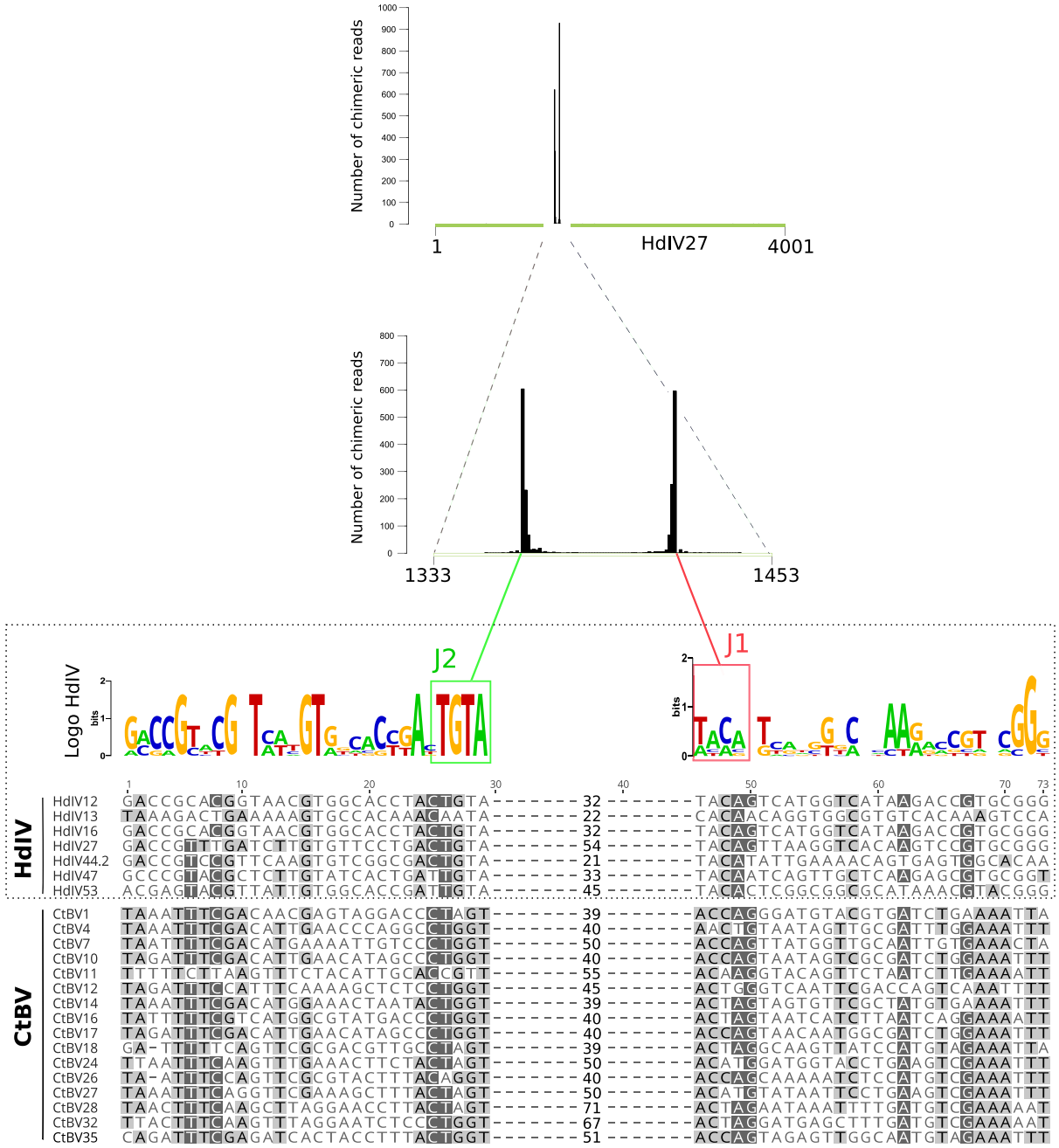


FIG. 4. Polydnavirus DNA circles undergo HIM-mediated chromosomal integration. Top graph: map of chimeric reads along *Hyposoter didymator* ichnovirus circle 27 (Hd27) (green) in hemocytes showing that almost all chimeric reads map to the HIM (white). The circle is oriented from the 5′ DRJ to the 3′ DRJ. Below is a magnified view containing the 120-bp HIM, showing the two regions with many chimeric reads, called J2 (left) and J1 (right), corresponding both to the borders of Hd27 sequences integrated in parasitized host DNA. Below is shown the sequence logo of J2 and J1 from HdIV circles only (not including CtBV circles, as indicated by the dashed rectangle) generated with weblogo.berkeley.edu, and the alignment of the HIMs of the seven HdIV circles that integrated into the *S. frugiperda* genome. An alignment of the CtBV circles described in Muller et al. (2021) is also shown below to allow comparison between BV and IV HIM. Grey shading indicates the level of sequence conservation. Numbers in the middle of the alignment indicate the number of nucleotides that are present between the J1 and J2 motifs and that are lost upon integration of PDV circles in host genomes. This region is not conserved between circles and was thus removed to facilitate the reading of the figure.

which showed a very high sequencing depth (normalized depth > 15× in three samples), the ranges of sequencing depths of integrated HdIV circles were similar to that of non-integrated HdIV circles (fig. 6). For example, in hemolymph 24 h p.p., integrated circles (in blue in fig. 6) were sequenced at depths varying between 243× and 3,400×

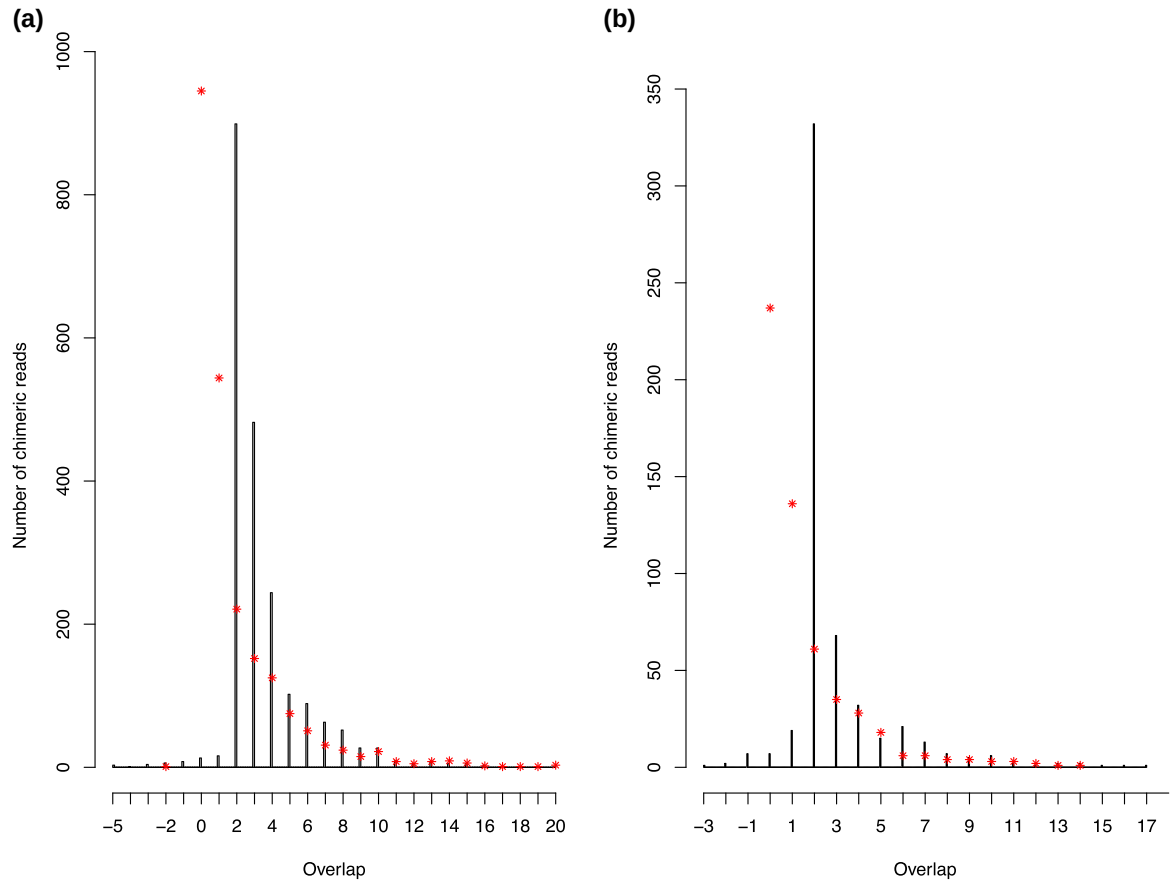


Fig. 5. Distribution of microhomology lengths at wasp–host junctions in chimeric reads. Black bars correspond to the numbers of observed chimeric reads for each microhomology length. Red asterisks correspond to the expected numbers of chimeric reads for each microhomology length. (*a*) Distribution of microhomology lengths for HdIV–host junctions mapped in J1 or J2. (*b*) Distribution of microhomology lengths for HdIV–host junctions mapped within HIM but outside J1 or J2.

whereas non-integrated circles (in red) were sequenced at depths varying between 93× and 4,164×. At the later time-point during parasitism (72 h p.p.), the overall quantity of circles (integrated and non-integrated) decreases, with, for example, about three times less circles in the fat body at 72 h p.p. than at 24 h p.p. (fig. 2). This decrease holds true even accounting for the fact that overall read count was lower at 72 h p.p. than at 24 h p.p. (table 1). Interestingly, this decrease is overall stronger for non-integrated circles than for integrated ones.

This is well-illustrated by the fact that sequencing depth over Hd16 is lower than that of some non-integrated circles at 24 h but it becomes higher than non-integrated circles at 72 h p.p., a trend holding in both hemolymph and fat body (fig. 6). The capacity to integrate for a circle may thus increase its ability to persist in large amount during parasitism.

Distribution of Wasp Circles Throughout the Genome of *S. frugiperda*

We investigated whether the integrations of *H. didymator* viral circles occur randomly along the caterpillar genome or whether there is an integration bias. To achieve this, we have split the *S. frugiperda* genome into 100,000-bp windows and assessed whether some windows were subject to more integrations than expected by chance. We found that the numbers of HdIV integrations per window did not follow a Poisson distribution with *P*-values <0.001 for all samples indicating that these integrations are not randomly distributed along the *S. frugiperda* genome. Under the Poisson distribution, we do not expect any windows with more than six or seven IEs depending on the samples. However, for example, we detected in the hemolymph samples a window with 14 IEs (H24h) and another with 18 (H72h) IEs. These results suggest that segments do not integrate completely randomly in the caterpillar genome, where IEs seem to be slightly concentrated in specific genomic regions.

HIM-mediated Integrations of Polydnavirus DNA Circles in Multiple Lepidopteran Genomes

This analysis together with that of Wang et al. (2021b) on DsIV, that of Wang et al. (2021a) on CvBV, and our earlier study on CtBV (Muller et al. 2021) shows that PDV circles

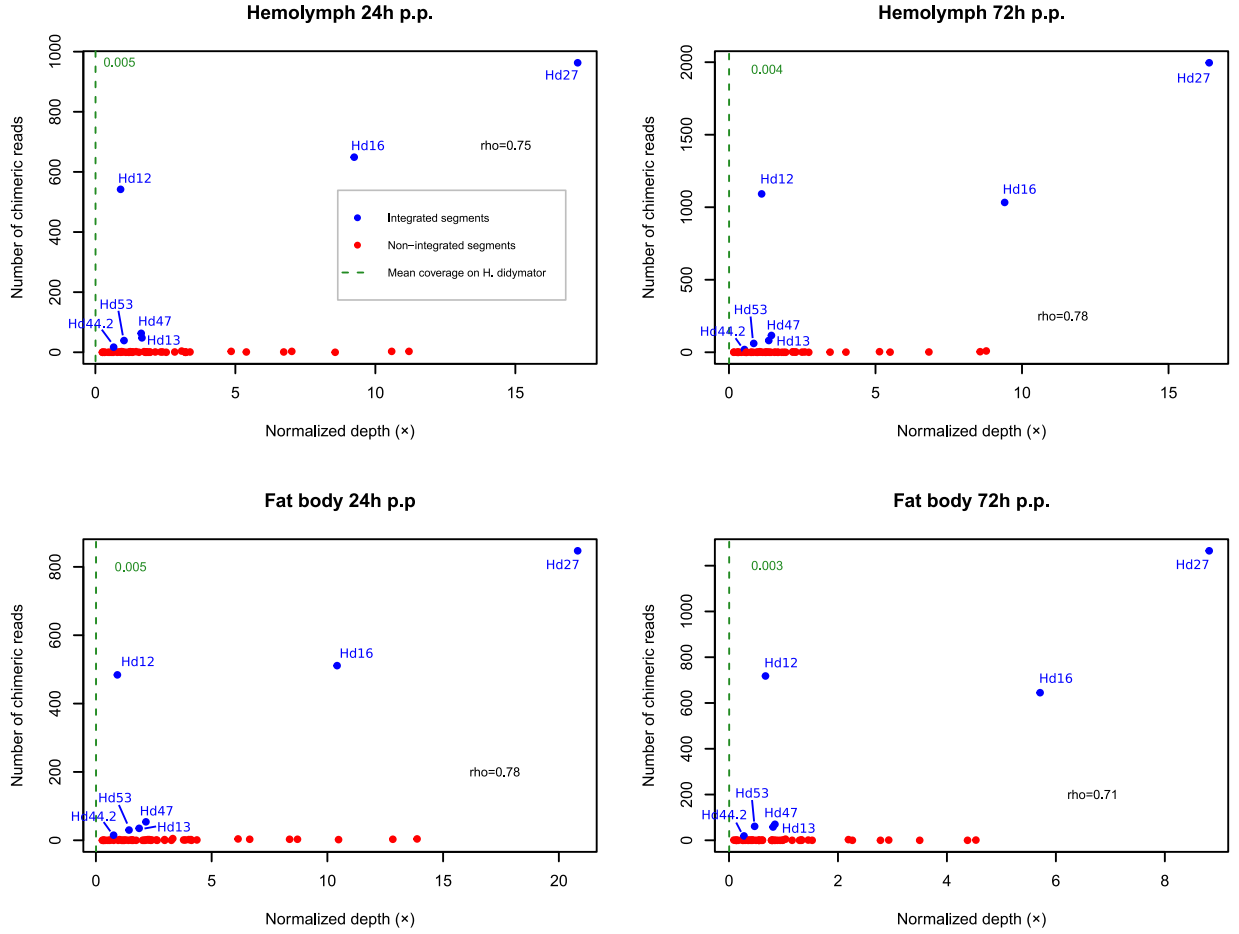


FIG. 6. Plot of the sequencing depth versus the number of chimeric reads for each of the HdIV circle in hemolyphm 24 and 72 h post parasitism (p.p.), and in fat body 24 and 72 h p.p. Blue dots represent circles that do integrate into the *S. frugiperda* genome, and red dots represent circles that do not integrate. The identification numbers of the segments are shown near each blue dot. The vertical green dashed line shows the average sequencing depth over the *H. didymator* genome (the precise value is indicated at the top of the line). The Spearman rho value indicates the correlation between sequencing depth and the number of IEs for circles that do integrate into the *S. frugiperda* genome. The *x*-axis shows normalized depth, that is sequencing depth per million reads.

undergo massive HIM-mediated chromosomal integration in host tissues during parasitism. We thus sought to assess whether HT of PDV circles occurred from wasp to lepidopterans through HIM-mediated chromosomal integration in the germline of lepidopteran hosts, followed by vertical transmission in host populations. For this, we used HIM-containing HdIV and CtBV circles to perform blastn similarity searches on 775 lepidopteran genomes available in Genbank as of December 2021. Our search yielded 4,648 and 49,079 lepidopteran sequences longer than 300 bp that showed similarity to HIM-containing HdIV and CtBV circles, respectively, with an e-value lower than 0.0001. We reasoned that if some of these sequences were integrated through double-strand breaks within HIM, then they were bordered by the J1 and J2 motifs (fig. 1) and may have since conserved these extremities, or at least one of them (J1 or J2). Moreover, the region found between J1 and J2 in the circular PDV molecule before integration (in grey in fig. 1), which is lost during integration, should not be retrieved during the analyses (see above and Chevignon et al. (2018) and Muller et al. (2021)). In agreement with these expectations, no less than 2,213 of the 4,648 IV sequences were found in lepidopteran genomes starting or ending within the HIM, and none of them contained the entire region lying between the J1 and J2 motifs. Regarding the 49,079 blastn hits involving CtBV, we found 174 sequences starting or ending within the HIM and only three containing the entire region lying between the J1 and J2 motifs. Our blastn searches did not allow us to directly retrieve full length PDV circles bordered by J1 on one side and J2 on the other side. Importantly however, we manually curated several lepidopteran PDV sequences and were able to reconstruct several examples of full-length integrated circles bordered by J1 and J2 and containing a single recombined DRJ sequence (supplementary Dataset 1, Supplementary Material online). We found that all lepidopteran PDV sequences were highly rearranged compared with our query

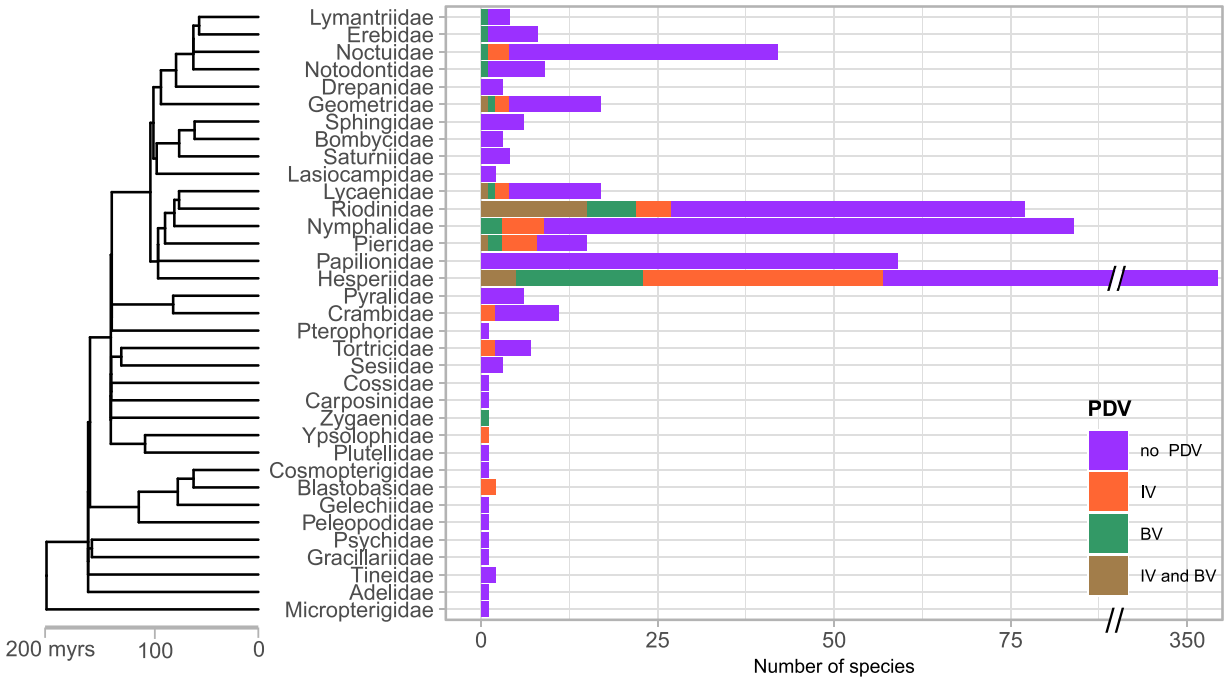


Fig. 7. Numbers of genomes per lepidopteran families in which J1- or J2-bordered BV and IV fragments were found. Details on the numbers and type of PDV circles in each lepidopteran species can be found in figure 8 and supplementary tables S3 and S4, Supplementary Material online. The tree and divergence times were recovered from timetree.org (Kumar et al. 2017).

wasp PDV sequences. This may be because integrations are ancient and were followed by rearrangements and/or degradation of the circles. Another reason behind the fragmentary nature of the blastn hits may be that the donor wasp species was distantly related to *H. didymator* or *C. typhae*. Although complete, circles from such distant species may be too divergent to align over their entire length with no interruption to those of *H. didymator* or *C. typhae*. In the following, we refer to fragments of IV or BV sequences as retrieved by our blastn search, that is fragments longer than 300 bp and bordered either by the J1 or J2 motif, to describe patterns of HIM-mediated integration of PDV circles into lepidopteran genomes.

Numbers and Distribution of Polydnavirus DNA Circles in the Lepidopteran Tree

IV fragments bordered by the J1 or J2 motif were found for 4 out of the 7 HIM-containing HdIV circles in a total of 87 lepidopteran species (out of 775) that belong to 11 out of the 35 lepidopteran families included in our search (supplementary table S1, Supplementary Material online; fig. 7). BV fragments bordered by the J1 or J2 motif were found for 14 out of 16 HIM-containing CtBV circles in a total of 60 lepidopteran species belonging to 11 families, including 7 (Geometridae, Hesperiidae, Lycaenidae, Noctuidae, Nymphalidae, Pieridae, and Riodinidae), which also had IV fragments integrated in their genome (supplementary table S2, Supplementary Material online; fig. 7). Altogether, PDV fragments (IV + BV) were retrieved in a total of 124 species with 23 having both IV and BV in their genome. These 124 species are from 15 lepidopteran families, which belong to eight superfamilies (Noctuoidea, Geometroidea, Papilionoidea, Pyraloidea, Tortricoidea, Zygaenoidea, Yponomeutoidea, Gelechioidea) that diverged more than 100 Ma, and cover a large diversity of lepidopterans (fig. 7) (Kawahara et al. 2019).

Lepidopteran IV fragments retained for this study are between 300 and 4,110 bp (median = 621 bp) and show between 65% and 96% nucleotide identity to HdIV circles (median = 78.5%). BV fragments are between 304 and 2860 bp (median = 515 bp) and show between 65% and 88% nucleotide identity to CtBV circles (median = 77%). All PDV fragments bordered by the J1 or J2 motifs as well as their alignment coordinates and percent identity to CtBV or HdIV are provided in supplementary tables S1 and S2, Supplementary Material online.

Numbers of PDV fragments are heterogenous both in terms of circles and lepidopteran species. In terms of lepidopteran species, numbers of IV fragments are generally relatively low, with 52 species having only one or two fragments and 76 out of the 87 species having less than 10 (supplementary table S3, Supplementary Material online). However, three species have high (81 in *Blastobasis adustella* and 80 in *B. lacticollela*, Blastobasidae) to very high (1,756 in the wainscot hooktip *Ypsolopha scabrella*, Ypsolophidae) numbers of IV fragments (supplementary table S3, Supplementary Material online, fig. 8). Numbers of BV fragments are also generally relatively low, with 48 out of the 60 species having only one or two J1- or

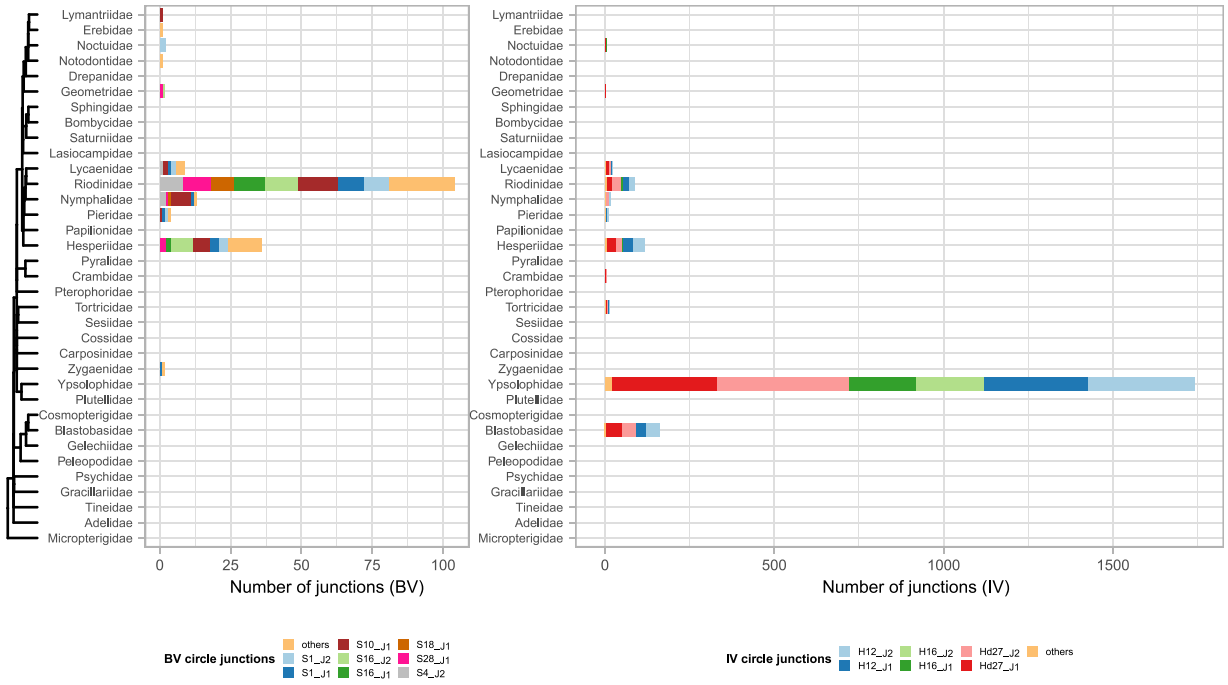


Fig. 8. Numbers of J1- or J2-bordered BV and IV fragments per lepidopteran families. Details on the numbers and type of PDV circles in each lepidopteran species can be found in supplementary tables S3 and S4, Supplementary Material online. Junctions' types for which less than 8 (BV) or 10 (IV) fragments were found in all families were grouped together in the "others" category.

J2-bordered BV fragments (supplementary table S4, Supplementary Material online). Four species have ten or more (up to 17) BV fragments (*Calephelis nemesis*, *C. perditalis*, *Apodemia duryi*, and *Boloria selene*) (supplementary table S4, Supplementary Material online, fig. 8). The number of both IV and BV fragments per lepidopteran family was positively correlated with the number of genomes surveyed per family (both Pearson's $r = 0.98$; P-values < 0.000001).

In terms of circles, we found a much higher number of IV fragments corresponding to Hd12/16 (634 J1 and 595 J2) and Hd27 (498 J1 and 422 J2) than to Hd47 (28 J1 and 35 J2) and Hd44.2 (1 J2) (fig. 8; supplementary table S3, Supplementary Material online). This mirrors the pattern observed for somatic integrations, with Hd12, Hd16, and Hd27 being highly integrated and Hd47 being the mostly highly integrated among lowly integrated circles (fig. 3). Numbers of IV J1 and J2 junctions are similar, which is consistent with the fact that during HIM-mediated integration, each integrated IV circle is linearized through double-strand break within the HIM and thus ends flanked by the J1 motif on one side and J2 motif on the other side (fig. 1). In terms of BV circles, the number of fragments found in lepidopteran genomes did not match with the relative abundance of integrated circles in parasitized hosts. For example, although we found the highest number of BV fragments for CtBV circles 16 (21 J2 and 13 J2) and 10 (31 J2) (supplementary table S4, Supplementary Material online, fig. 8), these two circles are among the least integrated circles in somatic genomes of parasitized *S. nonagrioides* individuals (fig. 6 in Muller et al. 2021). CtBV circle 1, which is by far the most highly integrated circle in parasitized host tissues is among the most integrated circles in lepidopteran genomes but the number of C1 fragments (16 J1 and 17 J2) are close to those of several other circles (supplementary table S4, Supplementary Material online, fig. 8). Contrasting with IV, for which similar amounts of J1-bordered and J2-bordered fragments were recovered for each circle in lepidopteran genomes, we often found different numbers of J1 and J2 BV fragments. In fact, we found only J1-bordered fragments (no J2-bordered fragments) for nine BV circles (supplementary table S4, Supplementary Material online, fig. 8). We believe that this disequilibrium is unlikely to have biological underpinnings because when we lowered the length threshold used to filter blastn hits to 200 bp, that is when we retain all BV-like sequences longer than 200 bp instead of 300 bp, the number of J1- and J2-bordered BV fragments are more similar for some circles (not shown). Yet we chose to keep the 300 bp threshold to retain blastn hits in order to maximize specificity and to ensure recovering large enough PDV fragments to conduct phylogenetic analyses.

Sequencing Depth Supports Germline Integration of Polydnavirus DNA Circles

Several features of the lepidopteran PDV fragments strongly suggest that they do not result from contamination or from

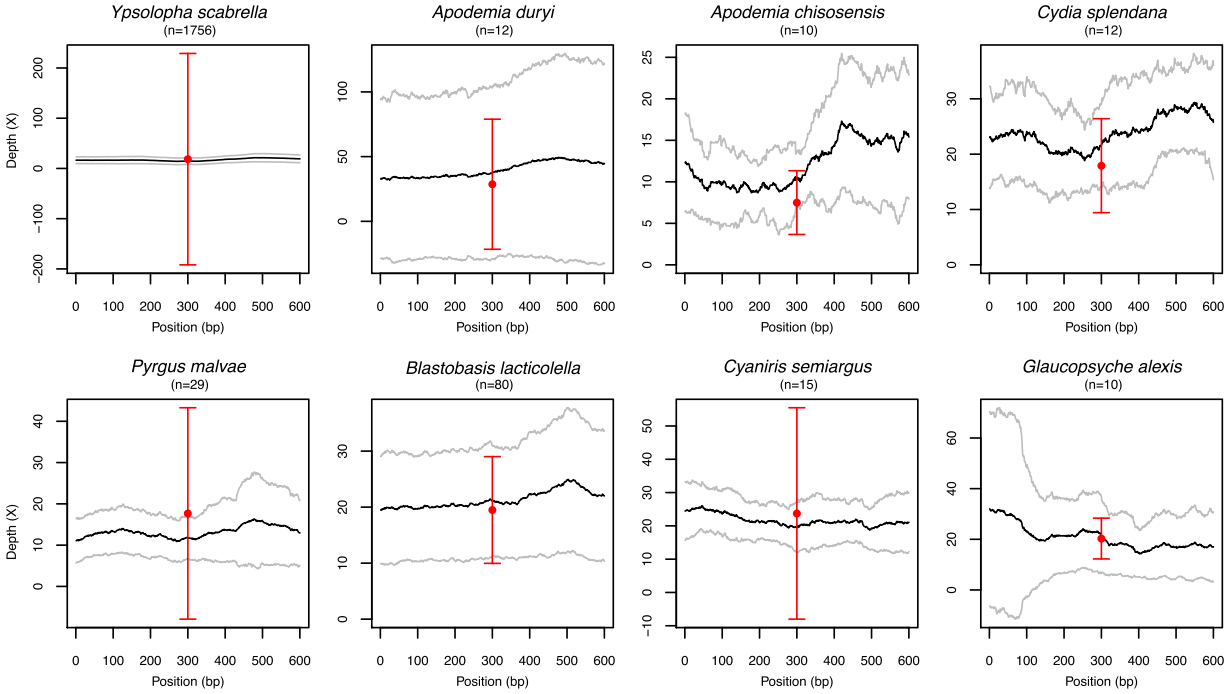


Fig. 9. Sequencing depth and numbers of chimeric reads support HIM-mediated chromosomal integration of IV fragments in the germline of lepidopteran species. Eight lepidopteran species in which 10 or more J1- or J2-bordered IV fragments were found were selected to compute average sequencing depth over all junctions. Number of junctions for each species are indicated between brackets. The *x*-axis indicates the genome position along the junction, the IV fragments being located between position 1 and 300 bp and the flanking lepidopteran genome regions being between 301 and 600 bp. The black line indicates average sequencing depth over 300 bp upstream and downstream of the junctions, with grey lines indicating standard deviation. Red dots indicate average numbers of chimeric reads supporting junctions in each species. Standard deviation is also shown.

chromosomal integrations in somatic genomes of parasitized individuals. Polydnavirus DNA circles are known to occur in three forms (fig. 1): 1) proviral segments flanked by DRJs in the genome of wasps, in which J1 and J2 motifs are next to each other, separated by a short sequence (about 50 bp) in the HIM, 2) circular sequences in the wasp ovaries and in host larvae, containing a single, recombined DRJ, and in which J1 and J2 motifs are also next to each other in the HIM, and 3) sequences bordered by J1 and J2 motifs integrated in parasitized hosts' somatic genomes. The fact that PDV fragments we report here in lepidopteran genomes are bordered at one of their extremities by the J1 or J2 motif, unlike in the wasp genome in which these motifs lie next to each other, argues against the possibility that these fragments result from contamination by wasp DNA that could have occurred during DNA extraction and/or sequencing. However, these fragments could correspond to somatic integrations that could have occurred in the individuals that were used for whole genome sequencing (many lepidopteran genomes are obtained from a single individual). This would imply that these individuals were parasitized by a wasp prior to sequencing. An important difference between somatic and germline integrations of PDV circles is that although the former is present only in a subset of somatic cells, the later should be present in all cells of an individual that would have received them from its parents. In agreement with the presence of PDV IEs in a subset of somatic cells during parasitism, 98% of the CtBV IEs we identified in parasitized *S. nonagrioides* larvae were covered by only one chimeric read (Muller et al. 2021). The pattern is similar here for HdIV, with on average 96.6% of HdIV IEs covered by only one read across samples. In contrast with somatic IEs, germline integrations should be covered by multiple reads because they are expected to be present in all cells of the sequenced lepidopteran individual. Furthermore, sequencing depth over the integration should be similar between the PDV and lepidopteran flanking genomic region. To verify this, we recovered raw Illumina reads for the eight lepidopteran species having ten or more IV fragments, and for the six species having five or more BV IEs, and we computed average sequencing depth and number of chimeric reads for all IV or BV junctions in each species. In agreement with our predictions, we found that the average number of chimeric reads per species was always higher than one (figs. 9 and 10). In fact, it was higher than five for five out of six species harboring BV fragments (fig. 10) and higher than ten for seven out of eight species harboring IV fragments (fig. 9). Importantly, the average number of chimeric reads per IE was always close to the average sequencing depth at the PDV integration point. Thus, variations in numbers of chimeric reads between species have no biological underpinnings but are due to variation in sequencing depth. Furthermore, we found that for most species, average

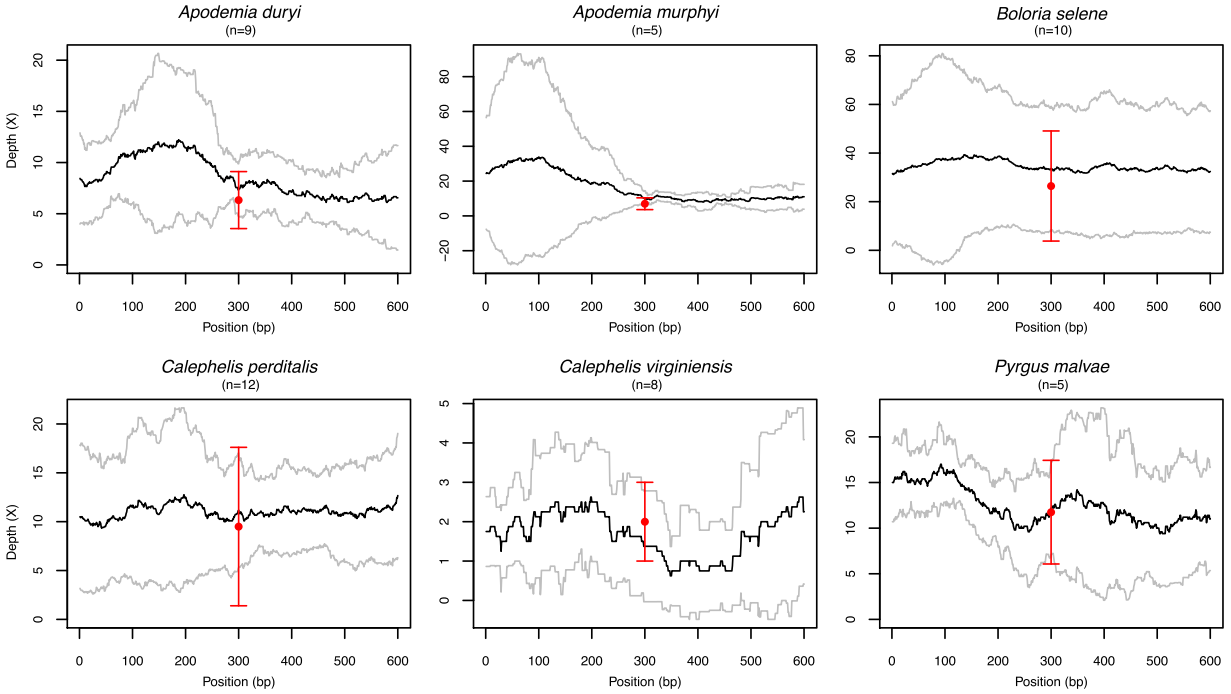


Fig. 10. Sequencing depth and numbers of chimeric reads support HIM-mediated chromosomal integration of BV fragments in the germline of lepidopteran species. Six lepidopteran species in which five or more J1- or J2-bordered BV fragments were found were selected to compute average sequencing depth over all junctions. Number of junctions for each species are indicated between brackets. The *x*-axis indicates the genome position along the junction, the BV fragments being located between position 1 and 300 bp and the flanking lepidopteran genome regions being between 301 and 600 bp. The black line indicates average sequencing depth over 300 bp upstream and downstream of the junctions, with grey lines indicating standard deviation. Red dots indicate average numbers of chimeric reads supporting junctions in each species. Standard deviation is also shown.

sequencing depth per species was homogeneous across PDV/lepidopteran junctions (figs. 9 and 10).

For two species (*Apodemia duryi* and *A. murphyi*), we noticed a particularly high increase in average sequencing depth over BV circles compared to the flanking lepidopteran genomic sequence (fig. 10), which could be interpreted as resulting from the presence of unintegrated BV circles in these species. However, reads produced by unintegrated circles are expected to cover such circles relatively homogeneously, which should here translate into a sharp increase in depth on the circle side starting from the very junction. Yet, here the increase in depth is progressive. We note that the quality of the assembly of the two species is low as both have N50 lower than 500 bp. It is thus possible that the increase in sequencing depth over the circle side in *Apodemia* species may be caused by the fact that several integrated circles have not been included in the assembly. Furthermore, the average number of chimeric reads in the two species (6.3 and 7.0) is consistent with the presence of most BV IEs being present in all cells. Although we are unable to provide a definitive explanation for the higher depth on BV circles compared to flanks for these two species, our observations are not consistent with the presence of unintegrated circles. Overall, we contend that these results indicate that most J1/J2-bordered PDV fragments we found in lepidopteran genomes result from germline integration that were then transmitted vertically in host populations.

An independent confirmation of this reasoning was obtained by assessing experimentally the presence of J1 and J2 extremities of insertions related to the HdIV12 circle and to the *C. congregata* BV circles 1 and 17 in different individuals of two very common butterflies: the small white and the green vein white (*Pieris rapae* and *Pieris napi*). Twelve individuals of each species were collected from three different locations in France. We obtained specific bands by PCR amplification of the junctions between flanking and viral sequences for all individuals (fig. 11).

Sequencing of the PCR products confirmed that J1 and J2 constituted actually the extremities of the four insertions identified from *P. rapae* and *P. napi* genomes, which thus appear to be fixed in the population and not a particular feature of the individuals used for genome sequencing. All sequences are provided in supplementary Dataset 2, Supplementary Material online. The *P. napi* HdIV 12 J2 sequence corresponds to the blastn hit number 167 in supplementary table S1, Supplementary Material online. The other five verified junctions were not recovered in the bioinformatic search because this experimental study was performed independently by three of us (G.P., J.M.D., and K.M.) in 2018. The study differed from the bioinformatic survey performed in 2022 in that it used CcBV as a query bracovirus instead of CtBV and it included two genomes of *P. napi* instead of one (see Materials and Methods and supplementary Dataset 2, Supplementary Material online).

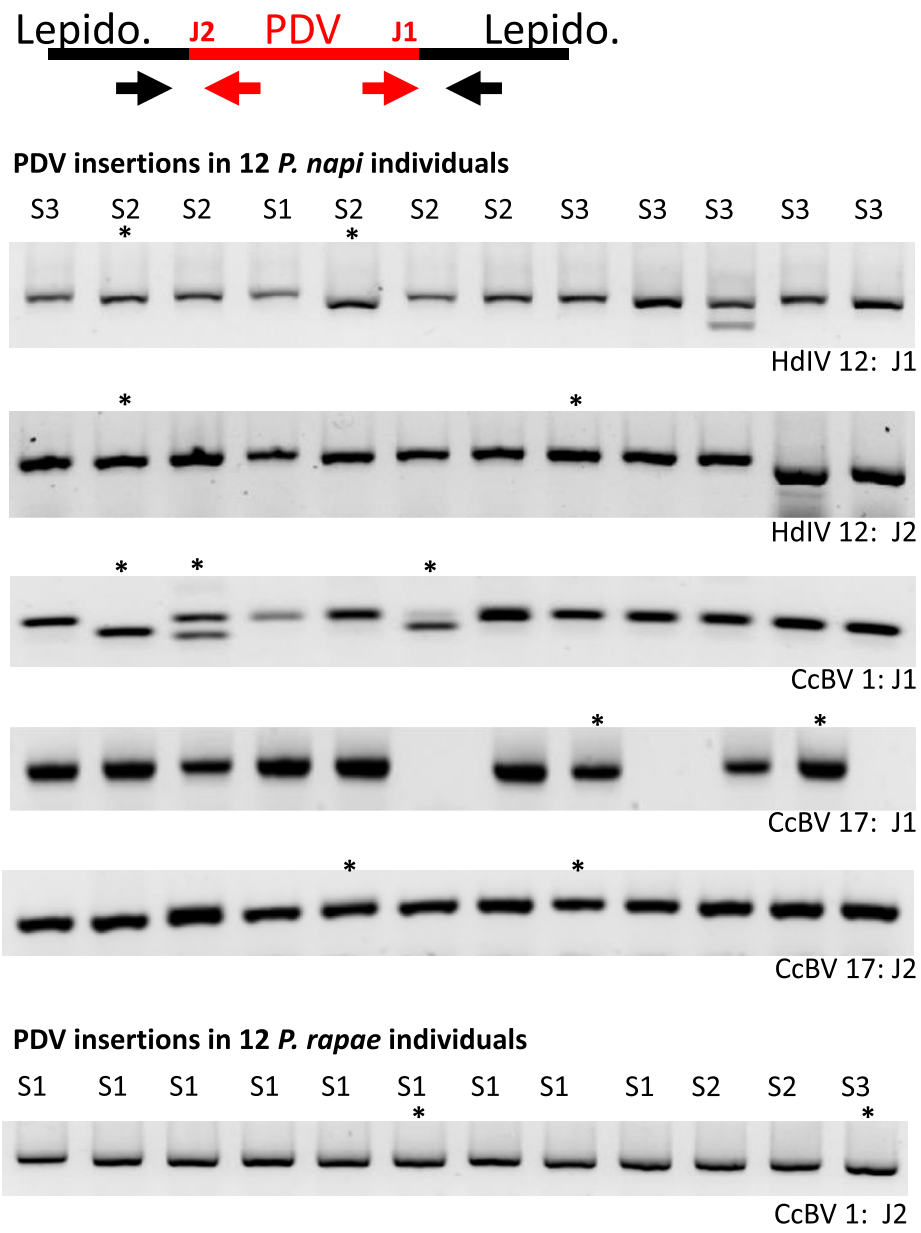


Fig. 11. Experimental verification of PDV–lepidopteran junctions in populations of two butterfly species (*Pieris napi* and *P. rapae*). The top illustration shows a schematic HIM-mediated insertion of a PDV circle bordered by the J1 and J2 motifs, as well as the position of the PCR primers (arrows at the bottom) designed on each side of the junction. Below are pictures of electrophoresis gels showing presence or absence of PCR bands obtained with primers targeting five PDV–*P. napi* junctions and one PDV–*P. rapae* junction in 12 individuals of each species sampled in three different geographical locations (S1, S2, S3, see Materials and Methods). Asterisks indicate PCR products that were Sanger-sequenced. The sequences are provided in supplementary Dataset 2, Supplementary Material online. For each verified junction, the type of PDV circle is indicated at the bottom right of the gel. HdIV, *Hyposoter didymator* ichnovirus; CcBV, *C. congregata* BV. The genomic coordinate of each junction is provided in supplementary Dataset 2, Supplementary Material online. Note that polymorphism was observed in *P. napi* for CcBV1 J1 containing PCR products (two individuals having two bands) and in CcBV17 J1 (amplification was reproducibly not obtained from the DNA of three individuals) probably reflecting a still ongoing erosion of these regions. CcBV1 insertions in *P. napi* and *P. rapae* have different flanking regions (see supplementary Dataset 2, Supplementary Material online).

Integration Dynamics of Polydnavirus Fragments During Lepidopteran Evolution

Some PDV circle integrations are likely to be ancient and thus shared with other species as found for a 5-million years old bracovirus integrations shared by the monarch and related species of the Danaina subtribe (Gasmi et al. 2015). To gain insights into the timing of different IEs of PDV fragments in lepidopteran germlines, we first searched for PDV fragments that are orthologous between species, that is fragments located at the same position in

Table 2. Characteristics of J1- or J2-bordered PDV Fragments Shared at Orthologous Loci Between Two or More Lepidopteran Species.

PDV Type	Number[a]	Junction	Species 1	Species 2	%ID	Ali. Length[b]	PDV Size[c]	Flank Size
IV	$1^{(A)}$	Hd27_J2	*A. duryi*	*A. virgulti*	97.652	1,533	546	987
IV	$5^{(A)}$	Hd27_J2	*A. mejicanus*	*A. virgulti*	99.547	1,546	546	1000
IV	$7^{(A)}$	Hd27_J2	*A. mormo*	*A. virgulti*	95.111	1,493	553	940
IV	$2^{(B)}$	Hd27_J1	*A. duryi*	*A. virgulti*	99.367	1,421	421	1,000
IV	$4^{(B)}$	Hd27_J1	*A. mejicanus*	*A. duryi*	99.766	856	421	435
IV	$6^{(B)}$	Hd27_J1	*A. mormo*	*A. duryi*	95.148	742	421	321
IV	3	H12_J1	*A. duryi*	*A. virgulti*	91.442	1,297	564	733
IV	19	Hd27_J2	*C. nemesis*	*C. perditalis*	90.94	1,181	465	716
IV	20	H16_J2	*C. nemesis*	*C. virginiensis*	82.149	1,238	490	748
IV	23	Hd27_J1	*Doberes anticus*	*Telegonus cellus*	95.878	1,407	407	1,000
IV	24	Hd27_J1	*Ostrinia furnacalis*	*O. nubilalis*	99.142	1,049	433	616
IV	25	H12_J2	*P. macdunnoughi*	*P. napi*	98.541	754	1745	298
BV	2	S10_J1	*C. nemesis*	*C. perditalis*	93.57	1,377	479	898
BV	3	S1_J1	*C. nemesis*	*C. perditalis*	92.62	1,628	629	999
BV	4	S1_J2	*C. nemesis*	*C. perditalis*	93.14	1,890	901	989
BV	6	S16_J2	*E. lupina*	*E. tegula*	94.18	969	429	547
BV	7	S7_J1	*O. hopfferi*	*O. stangelandi*	99.07	1,288	487	1,000

[a]Orthologous J1- or J2-bordered PDV fragments are listed by pair of species, number refer to their number label in sequence alignments provided in supplementary Datasets 3 and 4, Supplementary Material online. Letters between brackets indicate pairs of species sharing the same orthologous fragment (e.g., PDV fragments 1, 5, and 7 are shared at the same orthologous locus in *Apodemia duryi*, *A. virgulti*, *A. mejicanus* and *A. mormo*).
[b]Ali. size indicate the length of the PDV + flank alignment provided in supplementary Datasets 3 and 4, Supplementary Material online.
[c]PDV size indicate the size of the largest PDV fragment among a group of orthologous fragments.

the genome of two or more species. Our approach allowed us to identify five BV fragments and eight IV fragments shared by two or more species at the same genomic position (table 2).

One IV orthologous fragment, found in *Calephelis* metalmarks may be as old as the monarch integration as it shows only 82.1% nucleotide identity between *C. nemesis* and *C. virginiensis* and the split between the two species has been dated at 4.9 Ma (table 2, supplementary Dataset 3, Supplementary Material online) (Cong et al. 2017). By contrast, all other cases involve species for which divergence time is unknown but is likely recent because the species are congeneric and/or nucleotide identity between orthologous fragments is high (from 90.9% to 99.3%; mean = 96.6%). Regarding IV orthologous fragments, two are shared between four very closely related species of metalmark butterflies (genus *Apodemia*, Riodiniidae) (Zhang et al. 2019), one of them is shared between the green-veined white (*Pieris napi*, Pieridae) and its close relative *P. macdunnoughi* (Chew and Watt 2006) and another one is shared between the European corn borer (*Ostrinia nubilalis*, Crambidae) and the Asian corn borer (*O. furnacalis*), which diverged recently (table 2; supplementary Dataset 3, Supplementary Material online) (Bourguet et al. 2014). In terms of BV orthologous fragments, three are shared between *C. nemesis* and *C. perditalis* metalmark butterflies, one is shared between two species of firetips (*Oxynetra hopfferi* and *O. stangelandi*) and one is shared between *Emesis lupina* and *E. tegula* (table 2; supplementary Dataset 4, Supplementary Material online) (Cong et al. 2017).

In addition to orthologous fragments shared between species, we found several paralogous IV fragments sharing the same immediate flanking genome region (supplementary table S5 and supplementary Dataset 5, Supplementary Material online). Numbers of such paralogous fragments vary from two in the Kamehameha butterfly *Vanessa tameamea* (Nymphalidae) to 83 in the wainscot hooktip *Y. scabrella* (Ypsolophidae), with a maximum of three sequences in a given paralogous group (supplementary table S5, Supplementary Material online). The finding of paralogous sequences sharing flanks indicates that some IV fragments were duplicated after integration. Most duplications are likely recent as identity levels within these paralogous groups are generally high (from 82.7% to 100%; mean = 97.2%; median = 99.2%).

We next generated multiple alignments of PDV fragments inserted in lepidopteran genomes and PDV circles from various wasps, that we submitted to phylogenetic analyses. It is important to note that these analyses were not performed to infer the direction of HT of PDV fragments (supplementary Text 1, Supplementary Material online). The direction can indeed be unambiguously inferred to be from wasp-to-lepidopteran based on the wasp origin of PDVs, the presence of J1 and/or J2 motif at the extremities of PDV fragments and the presence of recombined DRJs as explained in the preceding sections (supplementary Text 1, Supplementary Material online). As expected in these trees, PDV insertions found to be orthologous between two or more lepidopteran species group together. This is the case, for example, of the four Hd27 J2 fragments flanked by the same genomic region in four *Apodemia* species (table 2; fig. 12). By contrast, when multiple PDV fragments were found in genomes of a given lepidopteran family, genus, or species, they generally do not cluster together in a monophyletic group (fig. 12; supplementary fig. S2, Supplementary Material online). For example, Hd27 J2 fragments found in various Hesperiidae species (dark green in fig. 12) fall in at least

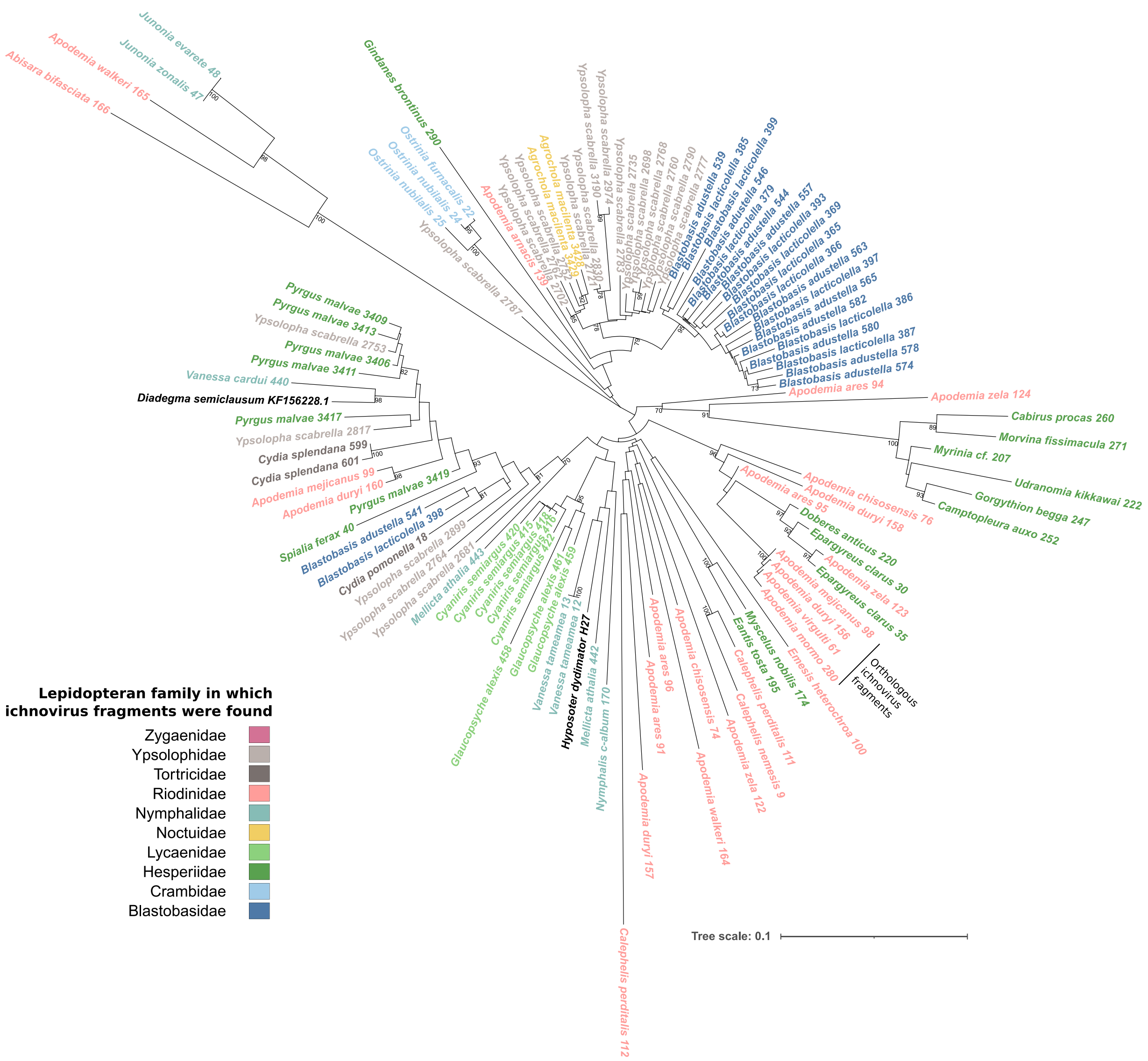


Fig. 12. Unrooted phylogeny of IV fragments homologous to the J2 junction of the HdIV27 circle found in lepidopteran genomes. All fragments similar to the J2 extremity of the HdIV27 circle found in lepidopteran genomes were aligned together with HdIV27 and a homologous circle from *D. semiclausum* (both are in black). Prior to alignment, the numerous fragments found in *Y. scabrella*, *B. adustella* and *Blastobasis lacticolella* were clusterized at 95% nucleotide identity threshold. The name of the species in which PDV fragments were found are colored according to lepidopteran family and the number associated to each species name is a unique identifier that allows retrieving the sequence in supplementary table S1, Supplementary Material online (column "Number"). Numbers on branches are bootstrap values higher than 70%. The HdIV27-like orthologous fragments found in *Apodemia* (table 2) are indicated. The tree was built using the Maximum Composite Likelihood model of substitution available in MEGA 11 (Tamura et al. 2021) with 100 bootstrap replicates.

six different clusters intermingled with fragments found in other lepidopteran families. Similarly, Hd27 J2 fragments from various species of the *Apodemia* genus, as well as those from the species *Y. scabrella* are polyphyletic, forming multiple clusters in the tree (fig. 12).

Furthermore, the few PDV sequences available only from different contemporary wasps are also generally scattered in the trees. This is the case, for example, of the Hd27 J2 sequence from the wasp *D. semiclausum*, which falls more closely related to lepidopteran Hd27 J2 than to the Hd27 J2 sequence of *H. didymator* (fig. 12). Similarly, H12/H16 J1 sequences from *H. fugitivus* and *D. semiclausum* are scattered in the phylogeny, falling closer to lepidopteran H12/H16 J1 fragments than to H12 and H16 J1 sequences of *H. didymator* (supplementary fig. S2, Supplementary Material online). Altogether, these results suggest that lepidopteran species have suffered IV integrations from multiple, relatively distantly related donor wasps (and at different times of the evolutionary history of wasp species). We predict that repeating such an analysis when PDVs from more Campopleginae and Braconidae wasps (which contain

thousands of species) are available will allow inferring the wasp source of all or most PDV insertions found in lepidopteran genomes.

Discussion

Preferential Integration of HdIV in Hemolymph Cells of *S. frugiperda*

In this study, we have shown that seven out of the 57 IV circles of the Ichneumonidae wasp *H. didymator* undergo chromosomal integration into two tissues (hemolymph and fat body) of parasitized fall armyworms (*S. frugiperda*). Three of the circles integrate massively into these somatic cells and we estimate that on average, each host haploid genome suffers between 13 and 40 IEs 72 h p.p. These numbers are similar to those reported earlier for *C. typhae* bracovirus (CtBV) which typically undergoes 12–85 IEs per host haploid genome on average, depending on the tissue (Muller et al. 2021). As observed for CtBV, HdIV integrations are more numerous in the hemolymph than in the fat body. This is also in line with earlier observations made for *Microplitis demolitor* BV (Beck et al. 2007). Given that hemocytes are individual cells that circulate through the entire body of caterpillars, they may be more accessible to PDVs than cells of other tissues, such as the fat body, that are akin to cell aggregates in which only external layers may be immediately reachable by PDVs. Alternatively, or in addition, PDVs may have evolved to preferentially target hemocytes, the main mediators of insect immune response (Lavine and Strand 2002; Jiang et al. 2010).

What Role for PDV Circle Integration in Parasitism Success?

A question emerging from earlier studies of PDVs relates to the role of circle integration during parasitism (Beck et al. 2011; Benoist et al. 2017; Chevignon et al. 2018; Muller et al. 2021). In our study of CtBV, we found that similar quantities of integrated circles (i.e., circles possessing a HIM motif) and non-integrated circles (no HIM motif) can persist over at least 7 days after oviposition, that is about half the time needed for *C. typhae* larvae to emerge from their host (Muller et al. 2021). Thus, it seemed that integration did not play so much of a role in ensuring persistence of BV circles in the context of this earlier study. However, host tissues parasitized by *C. typhae* were sequenced at only one time point in Muller et al. (2021), such that we could not assess how the overall quantity of integrated versus non-integrated circles evolves during parasitism. Here, we show that the number of IEs increases with time after oviposition. Although we did not perform replicates to test whether this increase is statistically supported, we observe that it holds both for the hemolymph and fat body, which may be considered pseudo-replicates. We deduce that this increase in IEs through time is due to continuous integration of circles from 24 to 72 h p.p., rather than to caterpillar cell replication, since we did not detect more IEs covered by multiple reads (as a result of cell division) at 72 h p.p. than at 24 h p.p. Mirroring this increase in integrated circles, we found that the sequencing depth over HdIV, and thus the overall quantity of IV circles, decreases through time (fig. 2). The decrease is particularly striking in the fat body, whereby HdIV sequencing depth is divided by three between 24 and 72 h p.p. (fig. 2). The combined observation of an increase in integrated IV circles and a decrease in overall circle quantity implies that the quantity of non-integrated circles decreases quite sharply through time after oviposition, and faster than that of integrated circles. In this context, integration may be seen as a way for the PDVs to persist in larger amounts in host cells throughout the entire duration of wasp progeny development.

Remarkable Similarity in Ichnovirus and Bracovirus Chromosomal Integration Patterns

This study is the second one investigating integration of IV DNA circles using bulk Illumina sequencing of parasitized host tissues. Like in Wang et al. (2021b), we show that the majority of IEs occur through double-strand breaks located in the HIM. We recently investigated chromosomal integration of a bracovirus (CtBV) during parasitism using the same approach as here and found that CtBV circles also almost exclusively integrated through double-strand breaks occurring within the HIM (Muller et al. 2021). This confirms the central role of these motifs in mediating circle integration in host somatic cells and ensuring parasitism success. Applying the same method to characterize circle integration of an IV and a BV (Muller et al. 2021) allows us to directly compare the two systems. Patterns of chromosomal integrations turn out to be highly similar between HdIV and CtBV, in terms of nature—all IEs are mediated by a HIM motif having a similar structure with conserved inverted repeats separated by a short stretch of sequence not conserved—, tissue tropism (hemocytes as preferential target), and numbers of IEs per host cell. This similarity suggests the mechanisms involved in integration could be highly related, which is unexpected given that IV and BV have an independent origin, deriving from domestication of viruses belonging to different viral families (Gimenez et al. 2020; Gauthier et al. 2021; Gilbert and Belliardo 2022).

The mechanisms underlying PDV circle integration are not fully understood. In the case of BV, three candidate wasp genes of nudiviral origin coding for Tyrosine recombinases (*vlf-1* and *int-2 homologues*) have been proposed to operate together ensuring the integration of HIM-containing circles (Chevignon et al. 2018) and were shown by functional assays to be involved in the integration of some circles of *Cotesia vestalis* bracovirus (Wang et al. 2021a). Moreover, there is evidence showing that some BV circles also rely on different lepidopteran retroviral integrases for chromosomal integration (Wang et al. 2021a). In the case of IV, circle integration has not been studied at the functional level yet, but the viral machinery

does not include a gene encoding a protein with a known integrase domain (Legeai et al. 2020), which suggests that host integrases could be involved. Indeed, site-specific recombinases/integrases such as retroviral integrases and transposases are abundant in eukaryote genomes. Another possibility is that among the genes of unknown function deriving from the IV ancestor, there is a recombinase from a still uncharacterized family that has not yet been recognized as such. It is noteworthy that the proteins encoded by the conserved gene set of ichnoviruses do not contain any recognizable conserved domain (Volkoff et al. 2010). The striking similarities between BV and IV circles integration mechanisms could be explained by the fact that they were shaped by similar structural constraints, despite their independent origin. In fact, the J1 and J2 motifs are much alike inverted terminal repeats of DNA transposons (ITRs). However, unlike ITRs which lie at the extremities of DNA transposon copies, these inverted repeats are located next to each other, only separated by the short stretch of nucleotides deleted during integration. Once a transposase binds to the ITRs, the formation of the transpososome relies on properties of the DNA separating the ITRs, including a minimum length (Hickman et al. 2018) (supplementary fig. S3, Supplementary Material online). Similarly, to form an integration complex starting from a PDV circle, a minimum sequence length between the ITRs is most probably required, which would explain the presence of the intervening DNA between the J1 and J2 motifs (supplementary fig. S3, Supplementary Material online). The fact that the IV and BV HIM only show very low similarity in primary sequence may be due to the involvement of recombinases belonging to distantly related families. Thus, the action of similar constraints could be sufficient to explain the similarity of the HIM structure of BV and IVs. Confirmation of this hypothesis requires to better characterize the molecular mechanisms involved in BV and IV circles integration. Another possibility would be that recombinases of the same family, which remain to be identified, may act as major players in circle integration of both BVs and IVs. In this case, HT events of HIM sites might be invoked to explain that the two PDVs genera share the same mechanism despite their different origin. The similarity in J1 and J2 motifs as well as in other positions of the HIM between BVs and IVs may be viewed as supporting this scenario (fig. 4). HT events have been proposed as an explanation of the common phylogenetic signature shared by the virulence factors V-ANKs of BV and IVs (Falabella et al. 2007; Ceirqueira de Araujo et al. forthcoming).

Finally, though as for HdIV, Wang et al. (2021b) found a majority of DsIV IEs occurring at HIM in *P. xylostella* hemocytes, they also identified large numbers of IEs involving random positions along both HIM-bearing circles and circles devoid of HIM. This led them to conclude that two distinct mechanisms were involved in DsIV chromosomal integration. Here the number of IE falling outside HIM or in circle devoid of HIM is so low that the biological significance of these integrations in terms of parasitism success in the *H. didymator*/*S. frugiperda* system is questionable. Interestingly, whatever the system (CtBV, DsIV, HdIV), there is an enrichment for microhomologies between the host genome and IV circles at the virus–host junction site, suggesting that host DNA repair mechanisms may be involved in at least a subset of IEs (Muller et al. 2021; Wang et al. 2021b). In fact, the fraction of junctions involving microhomology is higher for IV than for BV as contrary to BV, for IVs we observe almost no junction involving blunt-ended wasp and host sequences or 1-bp microhomology between them.

Widespread Germline Infiltrations of Polydnaviruses in Multiple Lepidopteran Families

Earlier studies uncovered a number of BV circle fragments in the genome of various species of lepidopterans and proposed that these sequences were horizontally transferred from wasp through HIM-mediated integration in the germline genome of lepidopterans (Schneider and Thomas 2014; Gasmi et al. 2015; Di Lelio et al. 2019). However, none of the horizontally transferred BV sequences reported in these studies contained the J1 or J2 motif of the HIM. This could be explained either because the transfers were ancient and the integrated PDV sequences (including the HIM) were degraded, or because these transfers occurred through a mechanism not involving a double-strand break within the HIM. The finding that CcBV (Chevignon et al. 2018), DsIV (Wang et al. 2021b), CvBV (Wang et al. 2021a), CtBV (Muller et al. 2021), and HdIV (this study) undergo massive HIM-mediated chromosomal integration in host somatic cells during parasitism, together with the recent availability of hundreds of lepidopteran whole genome sequences led us to assess the extent to which this mechanism fostered wasp-to-lepidopteran HT during evolution. We uncovered dozens of J1- or J2-bordered IV and BV fragments in a total of 124 lepidopteran species belonging to 15 families. In addition to show that sequencing depth was homogenous over lepidopteran/PDV circle junctions for most of these fragments, we also found that they were supported by multiple chimeric reads, contrasting with somatic PDV circle insertions that are almost all covered by only one chimeric read. We also found that many of the PDV fragments integrated into lepidopteran genomes contained a single, recombined DRJ motif (supplementary tables S1 and S2; supplementary Dataset 1, Supplementary Material online; fig. 1), consistent with the transfer of wasp PDV circles that were then linearized and integrated into host genomes. Furthermore, we PCR-validated typical extremities of HIM-mediated junctions of two IV and two BV integrations in multiple individuals of two species of *Pieris* butterflies and uncovered multiple integrations shared at orthologous loci between two to four lepidopteran species. Altogether these results indicate that HIM-mediated germline infiltration of PDV circles is widespread in lepidopterans, the main hosts of braconid and ichneumonid parasitoid wasps.

Limitations of the Current Study and Perspectives

The number of J1- or J2-bordered PDV circle fragments we report here is necessarily underestimated because we only used HdIV and CtBV circles, for which HIM were annotated using the same approach (Muller et al. 2021; this study), as queries to perform our similarity searches. In addition, we used relatively conservative criteria to retain blast hits for downstream analysis (at least 300-bp long and *e*-value < 0.0001) and many genomes included in our search are of poor quality (only 210 out of 775 genomes included have a N50 > 10 kb). However, our phylogenetic analyses reveal widespread polyphyly of PDV fragments at the lepidopteran family, genus, and even species level, indicating that a given lepidopteran species or multiple species within a family or a genus have received PDV fragments from various wasp donor species. We found that the number of PDV fragments uncovered per lepidopteran family was positively correlated with the number of genomes available in each family, with no apparent biological factor further explaining the distribution of these fragments in lepidoptera. It is, however, noteworthy that some species have remarkably high numbers of PDV fragments (e.g., 81 and 80 PDV fragments in 2 *Blastobasis* species and 1,756 PDV fragments in *Y. scabrella*) that were acquired from multiple donor wasp species (fig. 12; supplementary fig. S2, Supplementary Material online). We are unaware of how frequently these species are targeted by parasitoid wasps in the wild. By contrast, all lepidopteran species that are known hosts of CtBV (host: *S. nonagrioides*), CcBV (host: *Manduca sexta*), CvBV (host: *P. xylostella*), HdIV (host: *H. armigera*) and DsIV (host: *P. xylostella*) are devoid of HIM-mediated PDV insertions. It will be interesting to assess what factors may foster HIM-mediated germline integration of PDV circles in different hosts or non-hosts lepidopteran species using a dedicated and comprehensive sampling of parasitoid wasp and lepidopteran species.

This study is only a first step in the large-scale characterization of wasp-to-host HT of PDV circles. We chose to only focus on J1- or J2-bordered PDV fragments because these fragments contain the molecular signature typical of well-characterized chromosomal integrations occurring during parasitism. We could thus formulate explicit predictions regarding expected patterns of sequencing depth and number of chimeric reads. However, our similarity search uncovered many more PDV-like sequences not bordered by the J1 or J2 motifs of the HIM in lepidopteran genomes. For example, our initial blastn search yielded 2,214 hits longer than 300 bp with an *e*-value < 0.0001 to HdIV circles devoid of HIM. This number was even much higher ($n = 19{,}558$) when using CtBV devoid of HIM as queries. Though these sequences likely underwent some form of HT involving lepidopterans and parasitoid wasps, another full dedicated study will be necessary to determine and quantify which scenarios best explain these transfers. For example, it is possible that some of these sequences originate from donor wasp species in which these PDV circles bear a HIM motif but that this motif is not present in the homologous PDV circles of the wasps we used as queries to perform our searches (*H. didymator* or *C. typhae*). Alternatively, the fact that nucleocapsids of PDVs are able to enter the nuclei may also favor integration of circles by a more general mechanism such as DNA repair. Although we did not observe a significant level of circle integration not involving HIM sites during parasitism, such rare events could occur in the germline at the time scale of evolution. It is also possible that many of these sequences correspond to transposable elements present in HdIV and CtBV (Dupuy et al. 2011) and that these TEs were horizontally transferred either together with PDV circles or independently, as proposed by Heringer and Kuhn (2022). Another limitation of this study is that it is focused only on lepidopterans, the major hosts of PDV-encoding braconid and campoplegine parasitoid wasps (Gauld 1988). Yet, some wasps within the two families are known to parasitize non-lepidopteran hosts (Robin et al. 2019) and it will be interesting to extend the search for HIM-mediated HT of PDV circles to a larger diversity of hosts. Extending the search to species not known to be hosts of parasitoid wasps may also reveal unexpected wasp–host interactions.

A Major Route of Horizontal Transfer of Genetic Material Among Insects

HT of genetic material is widespread and a major force shaping prokaryote evolution (Soucy et al. 2015). In eukaryotes, the importance of HT is increasingly recognized (Husnik and McCutcheon 2018; Sibbald et al. 2020; Van Etten and Bhattacharya 2020), but the extent to which it influenced genome evolution remains debated, especially in animals and other multicellular taxa (Martin 2017; Salzberg 2017). Several studies reported spectacular individual cases of HT of genes in various animals, often with functional evidence supporting an important role played by horizontally transferred genes in the recipient species (Moran and Jarvik 2010; Acuna et al. 2012; Wybouw et al. 2014, 2016; Danchin et al. 2016; Leclercq et al. 2016; Gasmi et al. 2021; Xia et al. 2021; Cummings et al. 2022). Large-scale analyses of HT of genes and TEs in insects tend to show that these transfers occurred recurrently in most insect lineages and likely had important consequences on insect genome evolution (Peccoud et al. 2017; Li et al. 2022). So far however, the mechanisms through which these transfers occurred remain unclear and no mechanism dedicated to HT is known in animals. Together with previous studies on CcBV (Chevignon et al. 2018), DsIV (Wang et al. 2021b), CvBV (Wang et al. 2021a), and CtBV (Muller et al. 2021), our work on HdIV shows that parasitoid wasps use massive HIM-mediated HT of PDV circles to hijack host somatic cells, ensuring parasitism success. We also show that though not dedicated to germline integration, this mechanism fostered many accidental HT of PDV sequences in the germline of a large number of lepidopteran hosts that survived to

parasitism (or to the injection of PDV only) and transmitted these sequences to their offspring. HIM-mediated HT of PDV sequences may thus be viewed as a major route of HT in insects. The extent to which it may have facilitated the HT of non-PDV sequences integrated into PDV DNA circles or copackaged with PDV circles into PDV viral particles remains to be assessed. Two studies have provided functional evidence that a PDV gene (*Sl* gasmin), acquired by a noctuid moth through HT, plays an essential role in the antibacterial immune response of the moth (Gasmi et al. 2015; Di Lelio et al. 2019). Our work suggests that many PDV genes have been acquired by lepidopterans through HIM-mediated HT. It opens new avenues to further quantify this phenomenon and the impact it had on insect evolution.

Materials and Methods

Insects Rearing, Parasitization and DNA Sequencing

The *S. frugiperda* laboratory colony is maintained on a semi-synthetic maize diet under stable conditions (24 ± 2 °C; 75–65% relative humidity; 16 h light:8 h dark photoperiod). The wasp *H. didymator* laboratory colony is reared on *S. frugiperda* at 26 ± 2 °C with a 16 h light:8 h dark photoperiod. Fourth instar *S. frugiperda* larvae were each parasitized by exposing them to one *H. didymator* female wasp until one oviposition event was observed. Two tissues were sampled: the hemolymph, which is the most targeted by BVs (Muller et al. 2021), and the fat body, because it is easy to isolate large quantity of this tissue, in turn facilitating DNA extractions. Hemolymph and fat body of 20 parasitized caterpillars were collected after 24 h or 72 h after oviposition (i.e., post-parasitism). Hemolymph was collected from the caterpillar proleg and stocked into 1.5 ml of ATL buffer from the Qiagen DNeasy Blood & Tissue Kit. The fat body was collected, washed with PBS buffer and also stored into 1.5 ml of ATL buffer after caterpillars were dissected and the digestive tract removed. The genomic DNA from the four caterpillar pools was extracted using the DNeasy Blood & Tissue Kit. DNA of the four samples from the parasitized caterpillars *S. frugiperda* was quantified on a Qubit apparatus and was sent to Novogene for Illumina sequencing in a 2 × 150 bp paired-end mode.

Reference Genomes Used to Characterize Somatic Insertions of *H. didymator* Ichnovirus

The genome of *S. frugiperda* (corn strain) was sequenced by Gimenez et al. (2020) using the long-read PacBio RSII (Pacific Biosciences) technology and assembled with Platanus. It is available on the BIPAA platform (https://bipaa.genouest.org/is/; Rennes, France) and under accession number PRJNA662887 in NCBI. This genome is 384.46 Mb, it has an N50 (i.e., minimum contig length that covers 50% of the genome) of 13.15 Mb and it is composed of 125 contigs.

The genome of *H. didymator* was sequenced by Legeai et al. (2020) using the short-read Illumina HiSeq technology and assembled with Platanus. It is available in the NCBI database under accession number PRJNA 589497. The genome is 226.9 Mb, it has an N50 of 3.3 Mb and it is composed of 131,161 sequences. HdIV regions were annotated by Legeai et al. (2020) using the genome annotation editor Apollo browser. Fifty-seven viral segments localized in 66 viral loci were annotated in 32 scaffolds.

Measuring Sequencing Depth on *H. didymator* and *S. frugiperda* Genomes

The four datasets of Illumina reads we obtained (two parasitized caterpillar tissues at two timepoints post-parasitism; SRA accession numbers provided in table 1) were quality-trimmed with Trimmomatic v.0.38 (LEADING:20 TRAILING:20 SLIDINGWINDOW:4:15 MINLEN:36 options) (Bolger et al. 2014) and the quality was evaluated with FastQC v.0.11.8 (Wingett and Andrews 2018). To get statistics on sequencing depth, we aligned the four datasets of trimmed paired-end reads using Bowtie2 v.2.3.4.1 (Langmead and Salzberg 2012) in end-to-end mode on both *S. frugiperda* and *H. didymator* reference genomes. We converted the output SAM files into sorted BAM files with samtools v.1.9. The sequencing depth for each of our samples was estimated with bedtools genomecov v2.26.0 (Quinlan and Hall 2010) on both reference genomes (-ibam and -d options). To get the sequencing depth of *H. didymator* proviral segments, we used bedtools coverage v2.26.0 (-d option) and provided as input a bed file containing the proviral segment coordinates and the bam file of the Illumina reads aligned on the *H. didymator* genome.

Characterizing Integrations of Wasp Ichnovirus DNA Circles in Somatic Host Genomes

To identify integrations of HdIV DNA circles in the *S. frugiperda* somatic genomes, we searched for chimeric reads, that is reads for which a portion aligns exclusively on HdIV and another portion aligns exclusively on the *S. frugiperda* genome. Reverse and forward raw fastq files were converted into fasta files. The resulting fasta files were aligned on the wasp genome with blastn version 2.6.0 (-task megablast -max_target_seqs 1 -outfmt 6). Reads aligning on the wasp genome were extracted from the fasta file using seqtk subseq (https://github.com/lh3/seqtk) and aligned with blastn (same options as above) on the *S. frugiperda* genome.

To further find chimeric reads most likely to result from the integration of HdIV into *S. frugiperda* genome, we used the approach described in Muller et al. (2021). Briefly, we used the tabulated outputs of the two successive blastn searches as entries of an R pipeline initially written to identify artificial chimeras generated during deep sequencing library preparation (Peccoud et al. 2018). This script identifies reads as chimeric reads if 1) at least 90% of the read aligns on one or the other genome, 2) at least 16 bp align *only* on one of the two genomes, 3) with a maximal of 20 bp overlap between the two read portions aligning on

a different genome, and 4) a maximal of random 5 bp insert at the chimeric junction. Then, we estimated the number of independent IEs by counting as one event all reads with identical or nearly identical coordinates in the two genomes. To account for possible sequencing errors and/or alignment differences between reads covering the same HdIV—*S. frugiperda* junction (resulting from the fact that the length of the chimeric read region corresponding to each genome varies between reads), we allowed the coordinate of chimeric reads to vary by 5-bp in the two genomes, that is all reads aligning at the same position ±5 bp in the two genomes were considered as reflecting the same integration events.

To be able to compare the numbers of integration events between samples, we normalized values by the number of sequenced reads mapping to the *S. frugiperda* genome, as in Muller et al. (2021). We calculated a scaling factor by dividing the number of reads mapped on the *S. frugiperda* genome by 1,000,000. Then, to obtain the number of IEs per million reads mapping on *S. frugiperda* for each HdIV circle, we divided the absolute number of IEs by this scaling factor. We used samtools view (version 1.9, option-c-F 4) to determine the number of reads mapped on the *S. frugiperda* genome.

Searching for Polydnavirus DNA Circles in Lepidopteran Genomes

To search for HIM-mediated HT of PDV circles in lepidopteran genomes available in Genbank, we used the seven HIM-containing HdIV circles and the 16 HIM-containing *C. typhae* BV (CtBV) circles previously described in Muller et al. (2021) as queries to perform similarity searches using blastn (-task blastn) on lepidopteran genomes. All HdIV and CtBV DNA circle sequences are provided in supplementary Dataset 6, Supplementary Material online and the HIM coordinates within these circles are provided in supplementary Dataset 7, Supplementary Material online. The sequences of each HIM motifs of HdIV and CtBV circles are also provided as a multiple alignment in supplementary Dataset 8, Supplementary Material online as well as in figure 4. A total of 844 lepidopteran genomes were available in Genbank as of December 2021. When more than one genome per species was available, we only retained the largest one, resulting in 775 genomes that were submitted to our search. Accession numbers, size, and N50 of these genomes are provided in supplementary table S6, Supplementary Material online. Blastn outputs were filtered in R (R Core Team) based on hit size, *e*-value, and coordinates. Some PDV circles are similar to each other. For example, Hd12 and Hd16 are similar to each other over most of their length. To avoid counting multiple times the same lepidopteran genome region as resulting from HIM-mediated HT, alignment coordinates of blastn hits were thus merged using bedtools v2.26.0 merge (-s -c 4 -o distinct) (Quinlan and Hall 2010). Merging was performed independently for IV-like and BV-like sequences. The script written for this part of the study is available on GitHub: https://github.com/HeloiseMuller/Chimera/tree/master/scriptsArticles/Heisserer2022. DRJ, recombined DRJ (also called "DRJ circle") and HIM motifs were searched manually or by blastn in lepidopteran PDV insertions using DRJ and HIM motifs previously annotated in CcBV and HdIV as queries (Legeai et al. 2020; Gauthier et al. 2021).

Sequencing Depth Surrounding Junctions Between PDV DNA Circles and Lepidopteran Genomes

To compute sequencing depths on junctions between PDV DNA circles and lepidopteran genomes, raw Illumina reads were downloaded from Genbank for nine species in which more than four junctions were found. Accession numbers of these reads are provided in supplementary table S7, Supplementary Material online. Reads were mapped on the genome of the nine species using bowtie2 (default options) (Langmead and Salzberg 2012). Read depth over 300 bp upstream and downstream of each junction was obtained using bedtools v2.26.0 coverage (Quinlan and Hall 2010).

Search for Orthologous PDV Fragments in Lepidopteran Genomes

We extracted all PDV fragments ending in the HIM together with 1,000 bp flanking the J1 or J2 motif of the HIM. We filtered out fragments showing similarity to PDV sequences over >100 bp in their flank, which may result from integration of PDV circles next to (or within) each other. We then used blastn to align all PDV fragments plus flanks on themselves. Self blastn hits as well as hits involving sequences corresponding to different circles were filtered out. We retained blastn hits involving the PDV fragments plus at least 200 bp of flanking regions in both sequences as candidate orthologous sequences. These candidate orthologous sequences were then all submitted to manual inspection to only retain PDV sequences integrated at the very same position in two species.

PCR Verifications

To experimentally verified that PDV are present in natural populations of butterflies, 12 individuals of *Pieris napi* and 12 individuals of *P. rapae* were collected in 2018 in three sites in France: S1, Pénerf (Morbihan) 47.511463, −2,622753; S2, Vernou-sur Brenne (Indre-et-Loire) 47.415418 0.858978; S3, Tours (Indre-et-Loire) 47.355069 0,702746. PDV junctions were identified in the genome of *P. napi* and *P. rapae* using similarity searches (blastn). This study was performed in 2018, before the large-scale bioinformatic study reported in this article. It was performed on two different assemblies of the *P. napi* genome (CAJQFU010000000 and DWAF00000000) whereas the large-scale bioinformatic study only included one (DWAF00000000). Although ichnovirus circles used as queries for the blastn searches were those of *H. didymator* (the same as those used in the large-scale bioinformatic study), bracovirus circles were those of *C. congregata* BV (instead of CtBV in the large-scale bioinformatic study).

This explains why only one junction (out of six) verified by PCR was also found in the large-scale bioinformatic search. PCR primers were designed on each side of each PDV–butterfly junction. DNA was extracted from all 24 butterfly and a standard PCR protocol was used to screen for the presence/absence of six junctions in these individuals. Junctions found in at least two individuals were Sanger-sequenced and are provided in supplementary Dataset 2, Supplementary Material online, together with the genomic coordinates of each junction in the genomes of *P. napi* and *P. rapae*.

Phylogenetic Analyses

Sequences homologous to PDV DNA circles and starting or ending within HIM were extracted from lepidopteran genomes using seqtk subseq (https://github.com/lh3/seqtk) and are provided in supplementary tables S1 and S2, Supplementary Material online. These sequences were then aligned, together with homologous DNA circles from various parasitoid wasps (*D. semiclausum*, *H. didymator*, and *H. fugitivus* for IV and *Cotesia sesamiae*, *C. vestalis* and *C. congregata* for BV) using MUSCLE (Edgar 2004). Given the large number of IV junctions found in some species, we clustered these sequences for all species in which more than five sequences were found for a given junction using Trimal (-maxidentity 0.95) (Capella-Gutierrez et al. 2009). Alignments were then trimmed with Trimal (-automated1) and submitted to Neighbor-joining analysis in MEGA 11 (Maximum Composite Likelihood model of substitution, uniform rates among sites and lineages, pairwise deletion) (Tamura et al. 2021). Bootstrap values were obtained through 100 replicates of the analysis as implemented in the MEGA software.

Supplementary Material

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

Acknowledgment

This work was supported by Agence Nationale de la Recherche, project ANR-18-CE02-0021-01 TranspHorizon.

Data Availability

All data produced during this study are available in public repositories. Genbank accession numbers of raw reads produced from *S. frugiperda* larvae parasitized by *H. didymator* are provided in table 1. Scripts are provided in GitHub: https://github.com/HeloiseMuller/Chimera/tree/master/scriptsArticles/Heisserer2022. All wasp PDV circles and all lepidopteran PDV fragments bordered by HIM are provided in supplementary Datasets 6, Supplementary Material online and supplementary tables S1 and S2, Supplementary Material online. Accession numbers of lepidopteran genomes surveyed in this study are provided in supplementary table S7, Supplementary Material online.

References

Acuna R, Padilla BE, Florez-Ramos CP, Rubio JD, Herrera JC, Benavides P, Lee SJ, Yeats TH, Egan AN, Doyle JJ, *et al.* 2012. Adaptive horizontal transfer of a bacterial gene to an invasive insect pest of coffee. *Proc Natl Acad Sci U S A.* **109**:4197–4202.

Beck MH, Inman RB, Strand MR. 2007. Microplitis demolitor bracovirus genome segments vary in abundance and are individually packaged in virions. *Virology* **359**:179–189.

Beck MH, Zhang S, Bitra K, Burke GR, Strand MR. 2011. The encapsidated genome of Microplitis demolitor bracovirus integrates into the host pseudoplusia includens. *J Virol.* **85**:11685–11696.

Beckage NE, Drezen J-M. 2012. *Parasitoid viruses: symbionts and pathogens*. 1st ed. London; New York: Elsevier/Academic Press.

Beckage NE, Gelman DB. 2004. W ASP PARASITOID DISRUPTION OF HOST DEVELOPMENT: implications for new biologically based strategies for insect control. *Annu Rev Entomol.* **49**:299–330.

Béliveau C, Cohen A, Stewart D, Periquet G, Djoumad A, Kuhn L, Stoltz D, Boyle B, Volkoff A-N, Herniou EA, *et al.* 2015. Genomic and proteomic analyses indicate that Banchine and Campoplegine polydnaviruses have similar, if not identical, viral ancestors. *J Virol.* **89**:8909–8921.

Benoist R, Chantre C, Capdevielle-Dulac C, Bodet M, Mougel F, Calatayud PA, Dupas S, Huguet E, Jeannette R, Obonyo J, *et al.* 2017. Relationship between oviposition, virulence gene expression and parasitism success in Cotesia typhae nov. sp. parasitoid strains. *Genetica* **145**:469–479.

Bézier A, Herbinière J, Lanzrein B, Drezen J-M. 2009. Polydnavirus hidden face: the genes producing virus particles of parasitic wasps. *J Invertebr Pathol.* **101**:194–203.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**:2114–2120.

Bourguet D, Ponsard S, Streiff R, Meusnier S, Audiot P, Li J, Wang Z-Y. 2014. 'Becoming a species by becoming a pest' or how two maize pests of the genus *Ostrinia* possibly evolved through parallel ecological speciation events. *Mol Ecol.* **23**:325–342.

Burke GR. 2019. Common themes in three independently derived endogenous nudivirus elements in parasitoid wasps. *Curr Opin Insect Sci.* **32**:28–35.

Burke GR, Hines HM, Sharanowski BJ. 2021. The presence of ancient core genes reveals endogenization from diverse viral ancestors in parasitoid wasps. *Genome Biol Evol.* **13**(7):evab105.

Burke GR, Simmonds TJ, Sharanowski BJ, Geib SM. 2018. Rapid viral symbiogenesis via changes in parasitoid wasp genome architecture. *Mol Biol Evol.* **35**:2463–2474.

Burke GR, Simmonds TJ, Thomas SA, Strand MR. 2015. Microplitis demolitor bracovirus proviral loci and clustered replication genes exhibit distinct DNA amplification patterns during replication. *J Virol.* **89**:9511–9523.

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**:1972–1973.

Chevignon G, Periquet G, Gyapay G, Vega-Czarny N, Musset K, Drezen J-M, Huguet E. 2018. Cotesia congregata bracovirus circles encoding *PTP* and *Ankyrin* genes integrate into the DNA of parasitized *Manduca sexta* hemocytes. *J Virol.* **92**(15):e00438–18. Available from: http://jvi.asm.org/lookup/doi/10.1128/JVI.00438-18

Chew FS, Watt WB. 2006. The green-veined white (Pieris napi L.), its Pierine relatives, and the systematics dilemmas of divergent character sets (Lepidoptera, Pieridae): DIVERGENT CHARACTERS AND DNA PHYLOGENY. *Biol J Linn Soc.* **88**:413–435.

Cong Q, Shen J, Li W, Borek D, Otwinowski Z, Grishin NV. 2017. The first complete genomes of Metalmarks and the classification of butterfly families. *Genomics* **109**:485–493.

Cummings TFM, Gori K, Sanchez-Pulido L, Gavriilidis G, Moi D, Wilson AR, Murchison E, Dessimoz C, Ponting CP,

Christophorou MA. 2022. Citrullination was introduced into animals by horizontal gene transfer from Cyanobacteria. *Mol Biol Evol*. **39**:msab317.
Danchin EG, Guzeeva EA, Mantelin S, Berepiki A, Jones JT. 2016. Horizontal gene transfer from bacteria has enabled the plant-parasitic nematode globodera pallida to feed on host-derived sucrose. *Mol Biol Evol*. **33**:1571–1579.
Desjardins CA, Gundersen-Rindal DE, Hostetler JB, Tallon LJ, Fadrosh DW, Fuester RW, Pedroni MJ, Haas BJ, Schatz MC, Jones KM, *et al.* 2008. Comparative genomics of mutualistic viruses of Glyptapanteles parasitic wasps. *Genome Biol*. **9**:R183.
Di Giovanni D, Lepetit D, Guinet B, Bennetot B, Boulesteix M, Couté Y, Bouchez O, Ravallec M, Varaldi J. 2020. A behavior-manipulating virus relative as a source of adaptive genes for *Drosophila* parasitoids. *Mol Biol Evol*. **37**:2791–2807.
Di Lelio I, Illiano A, Astarita F, Gianfranceschi L, Horner D, Varricchio P, Amoresano A, Pucci P, Pennacchio F, Caccia S. 2019. Evolution of an insect immune barrier through horizontal gene transfer mediated by a parasitic wasp. *PLoS Genet*. **15**:e1007998.
Drezen JM, Periquet G, Savary S, Tan F, Beckage N. 1997. Excision of the polydnavirus chromosomal integrated EP1 sequence of the parasitoid wasp Cotesia congregata (Braconidae, Microgastinae) at potential recombinase binding sites. *J Gen Virol*. **78**:3125–3134.
Dupuy C, Periquet G, Serbielle C, Bézier A, Louis F, Drezen J-M. 2011. Transfer of a chromosomal maverick to endogenous bracovirus in a parasitoid wasp. *Genetica* **139**:489–496.
Edgar RC. 2004. [No title found]. *BMC Bioinformatics* **5**:113.
Falabella P, Varricchio P, Provost B, Espagne E, Ferrarese R, Grimaldi A, de Eguileor M, Fimiani G, Ursini MV, Malva C, *et al.* 2007. Characterization of the IκB-like gene family in polydnaviruses associated with wasps belonging to different Braconid subfamilies. *J Gen Virol*. **88**:92–104.
Frayssinet M, Audiot P, Cusumano A, Pichon A, Malm LE, Jouan V, Vabre M, Malavieille S, Delalande M, Vargas-Osuna E, *et al.* 2019. Western European populations of the ichneumonid wasp hyposoter didymator belong to a single taxon. *Front Ecol Evol*. **7**:20.
Gasmi L, Boulain H, Gauthier J, Hua-Van A, Musset K, Jakubowska AK, Aury JM, Volkoff AN, Huguet E, Herrero S, *et al.* 2015. Recurrent domestication by lepidoptera of genes from their parasites mediated by bracoviruses. *PLoS Genet*. **11**:e1005470.
Gasmi L, Sieminska E, Okuno S, Ohta R, Coutu C, Vatanparast M, Harris S, Baldwin D, Hegedus DD, Theilmann DA, *et al.* 2021. Horizontally transmitted parasitoid killing factor shapes insect defense to parasitoids. *Science* **373**:535–541.
Gauld ID. 1988. Evolutionary patterns of host utilization by ichneumonoid parasitoids (Hymenoptera: Ichneumonidae and Braconidae). *Biol J Linn Soc*. **35**:351–377.
Gauthier J, Boulain H, van Vugt JJFA, Baudry L, Persyn E, Aury J-M, Noel B, Bretaudeau A, Legeai F, Warris S, *et al.* 2021. Chromosomal scale assembly of parasitic wasp genome reveals symbiotic virus colonization. *Commun Biol*. **4**:104.
Gilbert C, Belliardo C. 2022. The diversity of endogenous viral elements in insects. *Curr Opin Insect Sci*. **49**:48–55.
Gimenez S, Abdelgaffar H, Goff GL, Hilliou F, Blanco CA, Hänniger S, Bretaudeau A, Legeai F, Nègre N, Jurat-Fuentes JL, *et al.* 2020. Adaptation by copy number variation increases insecticide resistance in the fall armyworm. *Commun Biol*. **3**:664.
Gundersen-Rindal DE, Lynn DE. 2003. Polydnavirus integration in lepidopteran host cells in vitro. *J Insect Physiol*. **49**:453–462.
Heringer P, Kuhn GCS. 2022. Multiple horizontal transfers of a Helitron transposon associated with a parasitoid wasp. *Mob DNA*. **13**:20.
Herniou EA, Huguet E, Theze J, Bezier A, Periquet G, Drezen JM. 2013. When parasitic wasps hijacked viruses: genomic and functional evolution of polydnaviruses. *Philos Trans R Soc Lond B Biol Sci*. **368**:20130051.
Hickman AB, Voth AR, Ewis H, Li X, Craig NL, Dyda F. 2018. Structural insights into the mechanism of double strand break formation by Hermes, a hAT family eukaryotic DNA transposase. *Nucleic Acids Res*. **46**(19):10286–10301. Available from: https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gky838/5103951
Husnik F, McCutcheon JP. 2018. Functional horizontal gene transfer from bacteria to eukaryotes. *Nat Rev Microbiol*. **16**:67–79.
Jiang H, Vilcinskas A, Kanost MR. 2010. Immunity in lepidopteran insects. In: Söderhäll K, editor. *Invertebrate immunity*. Vol. 708. Advances in Experimental Medicine and Biology. Boston, MA: Springer. p. 181–204. Available from: http://link.springer.com/10.1007/978-1-4419-8059-5_10
Kawahara AY, Plotkin D, Espeland M, Meusemann K, Toussaint EFA, Donath A, Gimnich F, Frandsen PB, Zwick A, dos Reis M, *et al.* 2019. Phylogenomics reveals the evolutionary timing and pattern of butterflies and moths. *Proc Natl Acad Sci U S A*. **116**:22657–22663.
Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: a resource for timelines, timetrees, and divergence times. *Mol Biol Evol*. **34**:1812–1819.
Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. **9**:357–359.
Lavine MD, Strand MR. 2002. Insect hemocytes and their role in immunity. *Insect Biochem Mol Biol*. **32**:1295–1309.
Leclercq S, Thézé J, Chebbi MA, Giraud I, Moumen B, Ernenwein L, Grève P, Gilbert C, Cordaux R. 2016. Birth of a W sex chromosome by horizontal transfer of *Wolbachia* bacterial symbiont genome. *Proc Natl Acad Sci U S A*. **113**:15036–15041.
Legeai F, Santos BF, Robin S, Bretaudeau A, Dikow RB, Lemaitre C, Jouan V, Ravallec M, Drezen J-M, Tagu D, *et al.* 2020. Genomic architecture of endogenous ichnoviruses reveals distinct evolutionary pathways leading to virus domestication in parasitic wasps. *BMC Biol*. **18**:89.
Li Y, Liu Z, Liu C, Shi Z, Pang L, Chen C, Chen Y, Pan R, Zhou W, Chen X, *et al.* 2022. HGT is widespread in insects and contributes to male courtship in lepidopterans. *Cell* **185**(16):2975–2987.
Mao M, Strand MR, Burke GR. 2022. The complete genome of *Chelonus insularis* reveals dynamic arrangement of genome components in parasitoid wasps that produce bracoviruses. *J Virol*. **96**:e01573-21.
Martin WF. 2017. Too much eukaryote LGT. *BioEssays* **39**:1700115.
McKelvey TA, Lynn DE, Gundersen-Rindal D, Guzo D, Stoltz DA, Guthrie KP, Taylor PB, Dougherty EM. 1996. Transformation of gypsy moth (Lymantria dispar) cell lines by infection with glyptapanteles indiensis polydnavirus. *Biochem Biophys Res Commun*. **225**:764–770.
Moran NA, Jarvik T. 2010. Lateral transfer of genes from fungi underlies carotenoid production in aphids. *Science* **328**:624–627.
Muller H, Chebbi MA, Bouzar C, Périquet G, Fortuna T, Calatayud P-A, Le Ru B, Obonyo J, Kaiser L, Drezen J-M, *et al.* 2021. Genome-wide patterns of bracovirus chromosomal integration into multiple host tissues during parasitism. *J Virol*. **95**:e00684-21.
Murphy N, Banks JC, Whitfield JB, Austin AD. 2008. Phylogeny of the parasitic microgastroid subfamilies (Hymenoptera: Braconidae) based on sequence data from seven genes, with an improved time estimate of the origin of the lineage. *Mol Phylogenet Evol*. **47**:378–395.
Peccoud J, Lequime S, Moltini-Conclois I, Giraud I, Lambrechts L, Gilbert C. 2018. A survey of virus recombination uncovers canonical features of artificial chimeras generated during deep sequencing library preparation. *G3* **8**:1129–1138.
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.
Pichon A, Bézier A, Urbach S, Aury J-M, Jouan V, Ravallec M, Guy J, Cousserans F, Thézé J, Gauthier J, *et al.* 2015. Recurrent DNA virus

domestication leading to different parasite virulence strategies. *Sci Adv*. **1**:e1501150.
Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**:841–842.
R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.
Robin S, Ravallec M, Frayssinet M, Whitfield J, Jouan V, Legeai F, Volkoff A-N. 2019. Evidence for an ichnovirus machinery in parasitoids of coleopteran larvae. *Virus Res*. **263**:189–206.
Salzberg SL. 2017. Horizontal gene transfer is not a hallmark of the human genome. *Genome Biol*. **18**:85.
Santos BF, Klopfstein S, Whitfield JB, Sharanowski BJ. 2022. Many evolutionary roads led to virus domestication in ichneumonoid parasitoid wasps. *Curr Opin Insect Sci*. **50**:100861.
Schneider SE, Thomas JH. 2014. Accidental genetic engineers: horizontal sequence transfer from parasitoid wasps to their lepidopteran hosts. *PLoS ONE* **9**:e109446.
Sibbald SJ, Eme L, Archibald JM, Roger AJ. 2020. Lateral gene transfer mechanisms and pan-genomes in eukaryotes. *Trends Parasitol*. **36**:927–941.
Soucy SM, Huang J, Gogarten JP. 2015. Horizontal gene transfer: building the web of life. *Nat Rev: Genet*. **16**:472–482.
Stoltz DB, Krell P, Summers MD, Vinson B. 1984. Polydnaviridae—a proposed family of insect viruses with segmented, double-stranded, circular DNA genomes. *Intervirology* **21**:1–4.
Tamura K, Stecher G, Kumar S. 2021. MEGA11: molecular evolutionary genetics analysis version 11.battistuzzi FU, editor. *Mol Biol Evol*. **38**:3022–3027.
Van Etten J, Bhattacharya D. 2020. Horizontal gene transfer in eukaryotes: not if, but how much? *Trends Genet*. **36**:915–925.
Volkoff A-N, Cusson M. 2020. The unconventional viruses of ichneumonid parasitoid wasps. *Viruses* **12**:1170.
Volkoff A-N, Huguet E. 2021. Polydnaviruses (polydnaviridae), editors. *Encyclopedia of virology*: San Diego: Elsevier. p. 849–857. Available from: https://linkinghub.elsevier.com/retrieve/pii/B9780128096338215562
Volkoff A-N, Jouan V, Urbach S, Samain S, Bergoin M, Wincker P, Demettre E, Cousserans F, Provost B, Coulibaly F, *et al.* 2010. Analysis of virion structural components reveals vestiges of the ancestral ichnovirus genome. *PLoS Pathog*. **6**:e1000923.
Volkoff A-N, Rocher J, Cérutti P, Ohresser MCP, d'Aubenton-Carafa Y, Devauchelle G, Duonor-Cérutti M. 2001. Persistent expression of a newly characterized Hyposoter didymator polydnavirus gene in long-term infected lepidopteran cell lines. *J Gen Virol*. **82**:963–969.
Wang Z, Ye X, Zhou Y, Wu X, Hu R, Zhu J, Chen T, Huguet E, Shi M, Drezen J-M, *et al.* 2021a. Bracoviruses recruit host integrases for their integration into caterpillar's genome. *PLoS Genet*. **17**:e1009751.
Wang Z-H, Zhou Y, Yang J, Ye X, Shi M, Huang J, Chen X. 2021b. Genome-wide profiling of diadegma semiclausum ichnovirus integration in parasitized plutella xylostella hemocytes identifies host integration motifs and insertion sites. *Front Microbiol*. **11**:608346.
Wingett SW, Andrews S. 2018. FastQ Screen: a tool for multi-genome mapping and quality control. *F1000Res*. **7**:1338.
Wybouw N, Dermauw W, Tirry L, Stevens C, Grbic M, Feyereisen R, Van Leeuwen T. 2014. A gene horizontally transferred from bacteria protects arthropods from host plant cyanide poisoning. *Elife* **3**:e02365.
Wybouw N, Pauchet Y, Heckel DG, Van Leeuwen T. 2016. Horizontal gene transfer contributes to the evolution of arthropod herbivory. *Genome Biol Evol*. **8**:1785–1801.
Xia J, Guo Z, Yang Z, Han H, Wang S, Xu H, Yang X, Yang F, Wu Q, Xie W, *et al.* 2021. Whitefly hijacks a plant detoxification gene that neutralizes plant toxins. *Cell*. **184**:1693–1705.e17.
Zhang J, Shen J, Cong Q, Grishin NV. 2019. Genomic analysis of the tribe emesidini (lepidoptera: riodinidae). *Zootaxa*. **4668**:475–488.

Part III

Factors promoting horizontal transfers

12 - Introduction

In this last part, we are investigating whether some factors could promote HT. Several factors were proposed, and we chose to test four of them: the aquatic habitat, the mode of fertilization, the geographical proximity, and the phylogenetic proximity.

We chose to investigate the aquatic habitat and the mode of fertilization following the publication of Zhang *et al.* (2020) in which they looked for HTT among 307 vertebrates. In this study they recovered 975 HTT events, yet 93.7% of these events involve teleost fishes. Following this result, we were wondering whether this over-representation of HT in teleost fishes is particular to this taxa, or whether it could be explained by an abiotic factor. The two main ecological differences between teleost fishes and the other vertebrates included in their study we could think of are the aquatic habitat and the mode of fertilization, two factors already hypothesized to play a role in HT (see chapter 4), although none of them were investigated yet. This is why we designed a first study in order to test these two factors (chapter 13).

In addition, Peccoud *et al.* (2017) recovered 2248 HTT events among 197 insects, and they found a positive correlation between the number of HTT events and both the phylogenetic proximity and the geographical proximity. Their dataset was composed of genomes from NCBI so they did not know the exact localities of their samples, yet they were able to assign a native biogeographic realm to 179 of their samples. Thus, the correlation they showed between the number of HTT events and the geographical proximity was found at large-scale. Here, we would like to know whether this correlation also exists at a smaller scale. For this, we designed a second study in which we produced our own dataset, sampling insects ourselves in the field (chapter 14).

The two studies I am going to present in this part were still in progress when writing this manuscript. I chose to include them despite the lack of final results because I spent as much time on these studies as on the ones of the previous parts, and because while working on these studies I acquired different bioinformatic skills than in the two other parts of my PhD. This is why I thought it was important to present this ongoing work in order to fully present my PhD work, and I am hoping it will also rise some additional interesting discussions during my defense.

The first study I am going to present, the chapter 13, benefited a lot from interactions with Sylvain Charlat (LBBE, Lyon) and Jean Peccoud (EBI, Poitiers). In fact, the entire design of the sampling was the topic of many (still ongoing) discussions with Jean and Sylvain and the pipeline used to detect HTT was heavily inspired from that designed by Jean and Clement in Zhang *et al.* (2020). I did many tests, trying to improve the pipeline at each step, and adapting it to my own study. The workflow of the pipeline I ended up with is illustrated in figure 12.1, and I will be using the same one in the second study (chapter 14).

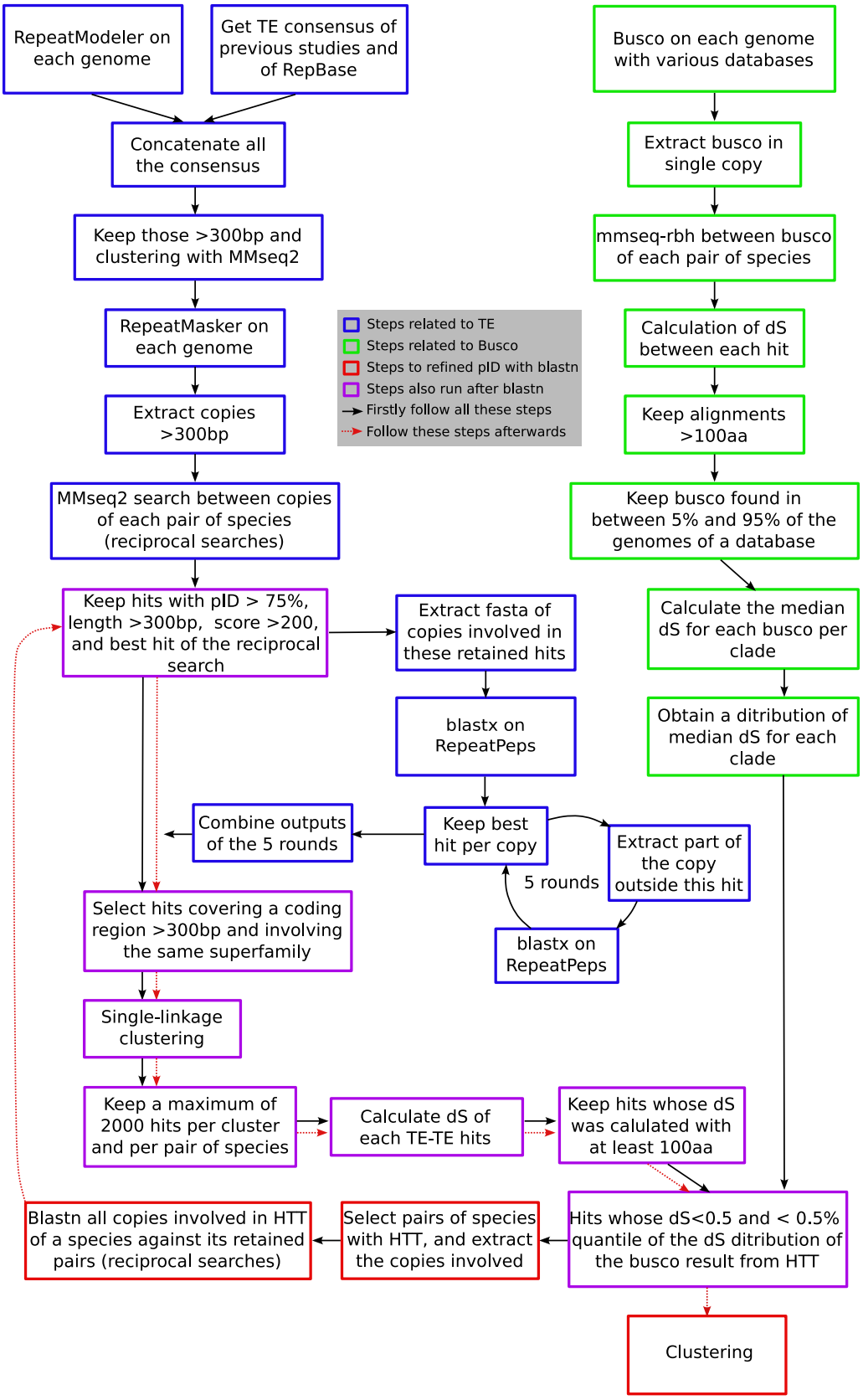


Figure 12.1: **Overview of the pipeline used to recover HTT in the two large-scale studies**. aa: amino acid; bp: base pairs; dS: synonymous distance; HTT: horizontal transfer of transposable element; pID: percentage of identity; TE: transposable elements

13 - The aquatic habitat and the mode of fertilization as factors promoting HTT

13.1 . Designing the dataset

Although it was shown that teleost fishes have more HTT than other vertebrates (Zhang *et al.* 2020), it could be due to another factor. Teleost fishes share a common ancestor and thus have many traits in common. To test whether it is the aquatic habitat (and/or the external fertilization) that promote HTT, we had to look for HTT in various and independent lineages. Although ideally it would be best to use all the genomes available on NCBI (8269 assemblies, representing 4516 species at the date of January 2022), it is computationally not possible. To choose the genomes that would be included in the study, we had to do some researches on the ecology of animals, looking for interesting taxa, which are the taxa with both aquatic and terrestrial species, and with genomes of both habitats available on NCBI. Here, taxa are defined arbitrarily, at different levels of taxonomy. To maximize the number of transitions (from terrestrial to aquatic or from aquatic to terrestrial), we chose to not focus on insects like during the rest of my PhD, but to extend my work to all other animals. In total, we included 6 taxa in which aquatic species are fully aquatic and that also have terrestrial species: Boreoeutheria, Afrotheria, Squamata, Gastropoda, Annelida, and Hemiptera. Hemiptera is the only taxa of insects for which genomes of aquatic species were available: two genomes of water striders which live at the surface of the water. In addition, we chose to include six other taxa of insects whose "aquatic" species are aquatic only at the larval stage: Palaeoptera, Amphiesmenoptera, Neuropterida, Syrphoidea, Ephydroidea, and Nematocera (the 3 last ones are Diptera). We also included three taxa of vertebrates whose "aquatic" species are not fully aquatic: Amphibia (most are aquatic at the larval stage only, while some are fully aquatic or fully terrestrial), Archosauria (crocodiles are more or less aquatic and they will be compared to terrestrial birds), and Testudines (even the marine turtles lay their eggs on the shore, whereas others are semi-aquatic and others are terrestrial). We also included 2 taxa composed of fully aquatic species only: the Actinopterygii and the Coelacanthi. These two taxa of aquatic vertebrates will be compared to the already included terrestrial Tetrapoda (Amphibia, Boreoeutheria, Afrotheria, Squamata, Archosauria, Testudines). We chose these two aquatic taxa, instead of Chondrichthyes (sharks) or Agnatha for example, because they are more closely related to Tetrapoda. Actinopterygii will also work as a positive control since we expect to recover many HTT in this taxa. To summarize, we identified 20 interesting taxa to test the factor of the habitat.

Among these 20 taxa, only few are also interesting to test the mode of fertilization. Actinopterygii is the most interesting taxa since it was estimated that viviparity (when the development of the embryo takes place inside the body of the parent, usually the female) originated 12 independent times in this taxa (Blackburn 2005). Yet, only six of these 12 transitions are represented by genomic data on NCBI. Another transition took place in Coelacanthi and another transition in the ancestor of Amniota (Tetrapoda others than Amphibia). So in addition of comparing the number of HTT in Actinopterygii that use external fertilization *versus* the ones that use internal fertilization, we will also compare the number of HTT in Actinopterygii that use external fertilization *versus* the aquatic Amniota (all with internal fertilization) and the Coelacanthi (which is aquatic and use internal fertilization). A 4th

interesting taxa to test this factor is the Amphibia: although all the terrestrial Amphibia use internal fertilization, some aquatic Amphibia also use internal fertilization while others use external fertilization. Two other taxa that could have been interesting are the Chelicerata and the Annelida. However, Chelicerata is not as convenient to test this factor since the species that use internal fertilization are all terrestrial and the ones that use external fertilization are all aquatic. So in this taxa, there is a dependence of the two factors we want to test. Regarding Annelida, they have both modes of fertilization, but which species uses which mode is not well documented, so we are not going to be able to investigate this factor in that taxa. Because we identified just four interesting taxa to test this second factor, our statistical power on this factor will not be as strong as on the type of habitat.

Once all these taxa of interest were identified, we had to select genomes. For this, we downloaded the statistics of all the assemblies of animals available on NCBI at the date of January 2022 (8269 assemblies). We estimated that we should be able to handle a maximum of 300 genomes with the resources of our laboratory. Having 20 taxa, we decided to include a maximum of 20 genomes per taxa, 10 of each habitat. Because most of the HT we would recover between two very related genomes would be shared, it would not be so informative to include these related genomes in our dataset. This is why we firstly automatically selected one assembly per genus, picking the one with the best N50. Then, we picked 10 genomes per taxa and per habitat, favoring the best N50, a consistent genome size, and we maximized the taxonomical diversity. We reached 20 genomes only for Boreoeutheria, all the other taxa having either a limited number of terrestrial or aquatic species available. For the limiting groups, we took all the genomes available, no matter the quality. Regarding Actinopterigii and Amphibia, we also took the mode of fertilization into account during the sampling. In the case of Actinopterigii, which are all aquatic, we decided to sample 14 genomes: five with internal fertilization and nine with external fertilization. In the case of Amphibia, we picked all the available genomes of species using internal fertilization (four) and six genomes of species using external fertilization. Doing so, we sampled a total of 247 genomes (Table 13.1). These 247 genomes represent a minimum of 21 transitions of habitat (either from aquatic to terrestrial or from terrestrial to aquatic) and a minimum of 6 transitions of mode of fertilization.

GCA_017591415.1	*Amia calva*	GCA_019972215.1	*Plecoglossus altivelis*	GCA_013265735.3	*Oncorhynchus mykiss*
GCA_014773175.1	*Lucifuga dentata*	GCA_904066995.1	*Poecilia reticulata*	GCA_014839685.1	*Anableps anableps*
GCA_014905685.2	*Nematolebias whitei*	GCA_015220745.1	*Sebastes umbrosus*	GCA_910589615.1	*Taurulus bubalis*
GCA_903798195.1	*Danio kyathit*	GCA_013347855.1	*Anguilla anguilla*	GCA_018136845.1	*Heterotis niloticus*
GCA_017654505.1	*Polyodon spathula*	GCA_900747795.4	*Erpetoichthys calabaricus*	GCA_901765095.2	*Microcaecilia unicolor*
GCA_902459505.2	*Geotrypetes seraphini*	GCA_901001135.2	*Rhinatrema bivittatum*	GCA_018994145.1	*Leptobrachium ailaonicum*
GCA_011038615.1	*Limnodynastes dumerilii*	GCA_905171765.1	*Bufo bufo*	GCA_019512145.1	*Engystomops pustulosus*
GCA_019857665.1	*Eleutherodactylus coqui*	GCA_905171775.1	*Rana temporaria*	GCA_019447015.1	*Hymenochirus boettgeri*
GCA_014898055.1	*Talpa occidentalis*	GCA_018350175.1	*Felis catus*	GCA_016077325.2	*Equus asinus*
GCA_002837175.2	*Physeter catodon*	GCA_004363515.1	*Inia geoffrensis*	GCA_000442215.1	*Lipotes vexillifer*
GCA_005190385.2	*Monodon monoceros*	GCA_003031525.2	*Neophocaena asiaeorientalis*	GCA_011762595.1	*Tursiops truncatus*
GCA_004027085.1	*Mesoplodon bidens*	GCA_004363435.1	*Platanista minor*	GCA_004363455.1	*Eubalaena japonica*
GCA_004329385.1	*Megaptera novaeangliae*	GCA_008782695.1	*Muntiacus muntjak*	GCA_001292865.1	*Sus scrofa*
GCA_014176215.1	*Rousettus aegyptiacus*	GCA_014108415.1	*Molossus molossus*	GCA_013371645.1	*Oryctolagus cuniculus*
GCA_016881025.1	*Ictidomys tridecemlineatus*	GCA_000222185.1	*Macaca fascicularis*	GCA_004026845.1	*Heterohyrax brucei*
GCA_004026925.2	*Procavia capensis*	GCA_000001905.1	*Loxodonta africana*	GCA_014332765.1	*Elephas maximus*
GCA_000243295.1	*Trichechus manatus*	GCA_013391785.1	*Hydrodamalis gigas*	GCA_015147995.1	*Dugong dugon*
GCA_000296735.1	*Chrysochloris asiatica*	GCA_004026705.1	*Microgale talazaci*	GCA_000313985.2	*Echinops telfairi*
GCA_000299155.1	*Elephantulus edwardii*	GCA_000298275.1	*Orycteropus afer*	GCA_021028975.1	*Sphaerodactylus townsendi*
GCA_009733165.1	*Naja naja*	GCA_900608585.1	*Pseudonaja textilis*	GCA_004319985.1	*Emydocephalus ijimae*
GCA_019473425.1	*Hydrophis cyanocinctus*	GCA_019677565.1	*Pituophis catenifer*	GCA_009769535.1	*Thamnophis elegans*
GCA_003400415.2	*Crotalus viridis*	GCA_004798865.1	*Varanus komodoensis*	GCA_019175285.1	*Sceloporus undulatus*
GCA_014337955.1	*Aspidoscelis marmoratus*	GCA_011800845.1	*Zootoca vivipara*	GCA_001723895.1	*Crocodylus porosus*
GCA_001723915.1	*Gavialis gangeticus*	GCA_001541155.1	*Alligator mississippiensis*	GCA_016128335.1	*Dromaius novaehollandiae*
GCA_012275295.1	*Melopsittacus undulatus*	GCA_015227805.3	*Hirundo rustica*	GCA_017639655.1	*Falco naumanni*
GCA_015227895.2	*Colaptes auratus*	GCA_009819595.1	*Merops nubicus*	GCA_018139145.1	*Gymnogyps californianus*
GCA_020740795.1	*Apus apus*	GCA_901699155.2	*Streptopelia turtur*	GCA_016699485.1	*Gallus gallus*
GCA_007922225.1	*Emydura subglobosa*	GCA_007922195.1	*Podocnemis expansa*	GCA_019425775.1	*Rafetus swinhoei*
GCA_007922185.1	*Carettochelys insculpta*	GCA_007922305.1	*Dermatemys mawii*	GCA_007922165.1	*Chelydra serpentina*
GCA_009764565.4	*Dermochelys coriacea*	GCA_015237465.2	*Chelonia mydas*	GCA_003597395.1	*Chelonoidis abingdonii*
GCA_007399415.1	*Gopherus evgoodei*	GCA_000241765.5	*Chrysemys picta*	GCA_001728815.2	*Malaclemys terrapin*
GCA_000225785.1	*Latimeria chalumnae*	GCA_016618385.1	*Nymphon striatum*	GCA_014155125.1	*Tachypleus gigas*
GCA_011833715.1	*Carcinoscorpius rotundicauda*	GCA_000517525.1	*Limulus polyphemus*	GCA_006491805.2	*Dysdera silvatica*
GCA_907164885.1	*Dolomedes plantarius*	GCA_015342795.1	*Argiope bruennichi*	GCA_000365465.3	*Parasteatoda tepidariorum*
GCA_013339695.1	*Rhipicephalus sanguineus*	GCA_002443255.1	*Varroa destructor*	GCA_000239435.1	*Tetranychus urticae*
GCA_015350385.1	*Aculops lycopersici*	GCA_002085665.2	*Dermatophagoides farinae*	GCA_020844145.1	*Sarcoptes scabiei*
GCA_010014785.1	*Anisolabis maritima*	GCA_019457785.1	*Vandiemenella viatica*	GCA_017312745.1	*Gryllus bimaculatus*
GCA_019974035.1	*Apteronemobius asahinai*	GCA_002313205.1	*Laupala kohalensis*	GCA_002928295.1	*Timema cristinae*
GCA_002778355.1	*Clitarchus hookeri*	GCA_000762945.2	*Blattella germanica*	GCA_002891405.2	*Cryptotermes secundus*
GCA_013340265.1	*Coptotermes formosanus*	GCA_001676475.1	*Isoperla grammatica*	GCA_907164805.1	*Brachyptera putata*
GCA_921293315.1	*Nemurella pictetii*	GCA_003287335.1	*Lednia tumana*	GCA_001676325.1	*Amphinemura sulcicollis*
GCA_014529405.1	*Neoneuromus ignobilis*	GCA_020423425.1	*Chrysopa pallens*	GCA_905475395.1	*Chrysoperla carnea*
GCA_001017535.1	*Tipula oleracea*	GCA_001014845.1	*Mochlonyx cinctipes*	GCA_001014815.1	*Chaoborus trivitattus*
GCA_013758885.1	*Anopheles albimanus*	GCA_015732765.1	*Culex quinquefasciatus*	GCA_000004015.3	*Aedes aegypti*
GCA_018397935.1	*Propsilocerus akamusi*	GCA_900005825.1	*Clunio marinus*	GCA_018290095.1	*Polypedilum vanderplanki*
GCA_902825295.1	*Chironomus riparius*	GCA_000265325.1	*Lutzomyia longipalpis*	GCA_000262795.1	*Phlebotomus papatasi*
GCA_001014335.1	*Coboldia fuscipes*	GCA_910594885.1	*Bibio marci*	GCA_010015015.1	*Bolitophila cinerea*
GCA_014529535.1	*Bradysia coprophila*	GCA_011634745.1	*Catotricha subobsoleta*	GCA_021018905.1	*Sitodiplosis mosellana*
GCA_000149185.1	*Mayetiola destructor*	GCA_905231855.1	*Eristalis tenax*	GCA_905187475.1	*Syritta pipiens*
GCA_907269105.1	*Volucella inanis*	GCA_916610125.1	*Cheilosia vulpina*	GCA_917880715.1	*Criorhina berberina*
GCA_905220385.1	*Xylota sylvarum*	GCA_916050605.1	*Platycheirus albimanus*	GCA_920937365.1	*Sphaerophoria rueppellii*
GCA_905146935.1	*Scaeva pyrastri*	GCA_910595825.1	*Xanthogramma pedissequum*	GCA_911387755.1	*Chrysotoxum bicinctum*
GCA_001014675.1	*Ephydra gracilis*	GCA_001015075.1	*Cirrula hians*	GCA_001014415.1	*Phortica variegata*
GCA_018903435.1	*Leucophenga varia*	GCA_018150985.1	*Chymomyza costata*	GCA_003285725.2	*Scaptodrosophila lebanonensis*
GCA_008121215.1	*Drosophila athabasca*	GCA_018904275.1	*Lordiphosa clarofinis*	GCA_018901835.1	*Scaptomyza graminum*
GCA_018903675.1	*Zaprionus capensis*	GCA_020383195.1	*Neomicropteryx cornuta*	GCA_910592155.1	*Ypsolopha scabrella*
GCA_905163555.1	*Notocelia uddmanniana*	GCA_918358865.1	*Melinaea marsaeus*	GCA_905163395.2	*Endotricha flammealis*
GCA_019059595.1	*Cnaphalocrocis exigua*	GCA_910589355.1	*Parapoynx stratiotata*	GCA_907165245.1	*Habrosyne pyritoides*
GCA_907269065.1	*Crocallis elinguaria*	GCA_916999025.1	*Orgyia antiqua*	GCA_019925095.1	*Dendrolimus kikuchii*
GCA_009617725.1	*Hydropsyche tenuis*	GCA_008973525.1	*Stenopsyche tienmushanensis*	GCA_009617715.1	*Plectrocnemia conspersa*
GCA_003347265.1	*Glossosoma conforme*	GCA_016648135.1	*Agrypnia vestita*	GCA_917563855.2	*Limnephilus lunatus*
GCA_016648045.1	*Hesperophylax magnus*	GCA_018340805.1	*Microvelia longipes*	GCA_001010745.2	*Gerris buenoi*
GCA_019843655.1	*Lethocerus indicus*	GCA_009739505.2	*Apolygus lucorum*	GCA_019009955.1	*Riptortus pedestris*
GCA_911387785.1	*Aelia acuminata*	GCA_014356525.1	*Nilaparvata lugens*	GCA_021130785.1	*Homalodisca vitripennis*
GCA_011764245.1	*Trialeurodes vaporariorum*	GCA_012654025.1	*Pachypsylla venusta*	GCA_009761765.1	*Phenacoccus solenopsis*
GCA_020882235.1	*Schizaphis graminum*	GCA_013282895.1	*Eriosoma lanigerum*	GCA_020796165.1	*Pantala flavescens*
GCA_000376725.2	*Ladona fulva*	GCA_011762765.1	*Rhinocypha anisoptera*	GCA_921293095.1	*Ischnura elegans*
GCA_002093875.1	*Calopteryx splendens*	GCA_000507165.2	*Ephemera danica*	GCA_001676355.1	*Baetis rhodani*
GCA_902829235.1	*Cloeon dipterum*	GCA_905338405.1	*Notodromas monacha*	GCA_017493165.1	*Daphnia pulicaria*
GCA_000591075.2	*Eurytemora affinis*	GCA_019096065.1	*Paracyclopina nana*	GCA_905330665.1	*Lepeophtheirus salmonis*
GCA_015104395.1	*Macrobrachium nipponense*	GCA_017591435.1	*Portunus trituberculatus*	GCA_020424385.2	*Procambarus clarkii*
GCA_023014485.1	*Bathynomus jamesi*	GCA_001587735.2	*Parhyale hawaiensis*	GCA_015478945.1	*Trachelipus rathkii*
GCA_004104545.1	*Armadillidium vulgare*	GCA_904063045.1	*Dimorphilus gyrociliatus*	GCA_903813345.1	*Owenia fusiformis*
GCA_000328365.1	*Capitella teleta*	GCA_019095985.1	*Streblospio benedicti*	GCA_001703475.1	*Hydroides elegans*
GCA_020002185.1	*Paraescarpia echinospica*	GCA_009193005.1	*Lamellibrachia luymesi*	GCA_905160935.1	*Enchytraeus crypticus*
GCA_000326865.1	*Helobdella robusta*	GCA_011800805.1	*Hirudo medicinalis*	GCA_015345955.1	*Poecilobdella manillensis*
GCA_900000155.1	*Eisenia fetida*	GCA_020284085.1	*Aporrectodea caliginosa*	GCA_020405105.1	*Eudrilus eugeniae*
GCA_900184025.1	*Amynthas corticis*	GCA_018105865.1	*Metaphire vulgaris*	GCA_016097555.1	*Gigantopelta aegis*
GCA_003343065.1	*Haliotis rufescens*	GCA_916613615.1	*Steromphala cineraria*	GCA_917208275.1	*Patella pellucida*
GCA_004794335.1	*Pomacea canaliculata*	GCA_018292915.1	*Batillaria attramentaria*	GCA_017654935.1	*Phymorhynchus buccinoides*
GCA_018857735.1	*Alviniconcha marisindica*	GCA_019648995.1	*Plakobranchus ocellatus*	GCA_014524965.1	*Biomphalaria glabrata*
GCA_009760885.1	*Achatina immaculata*	GCA_020796225.1	*Arion vulgaris*	GCA_014155875.1	*Cepaea nemoralis*
GCA_905116865.2	*Candidula unifasciata*				

Table 13.1: **The 247 assemblies of the dataset with their species names.** Species are in the same order (from left to right) as in figure 13.2, and the colors used for each taxa are the same ones.

We ran BUSCO v5.4 on these 247 genomes to get a better idea of their qualities, but also because we will need BUSCO genes for several other steps of the pipeline. We used the BUSCO dataset with the most precise lineage as possible. Looking at the BUSCO scores we replaced three assemblies. Overall, the BUSCO score of the 247 genomes are quite good, with a median of 90.9% of complete single BUSCO genes and an average of 84.9% (figure 13.1). 10 assemblies however have a score under 50%. We could not replace them because of a lack of genomes in their respective groups. Yet we chose to include them despite their low quality, because it is better to have additional genomes, even partial. Importantly, TE being in numerous copies, we predict that we will be able to annotate many TE even in a partial genome, as long as the genome is not too fragmented. One can never recover all HTT events anyway, not even in an assembly of good quality. It will be important though to take into consideration genome qualities when investigating the effect of our two factors on the number of HTT events. Yet, the number of low quality genomes that belong to terrestrial species and aquatic species is similar: four and six genomes with less than 50% of complete single BUSCO, respectively.

13.2 . Phylogeny

We imported the 247 species names in the TimeTree website (http://www.timetree.org/), that was able to automatically generate a dated tree with 186 of these species. We added the missing 61 species manually with the function bind.tip of the R package TreeTools. We placed the species according to the divergence time with their closest relative among the 186 species initially placed on the tree by TimeTree. We found the divergence time of most of the missing species in previous studies, however 31 polytomies remained. To solve them, we generated a *de nono* phylogeny independently for each subgroup of species with polytomy(ies), using 300 BUSCO genes in single copy. We chose the genes independently for each subgroup, favoring genes found in most genomes of the respective subgroups. We aligned these genes with MAFFT (–auto), and we trimmed the resulting alignments with trimAL (-strictplus). We then built a tree for each subgroup with iqtree2 (-m MFP -B 1000) without any constraints. We then dated this tree, giving some divergence times for pair of species that were available in TimeTree (options –date and –date-tip 0). This method allowed us to obtain the phylognetic tree shown in figure 13.2.

13.3 . Core genes dS distribution

We generated core-gene dS distributions for each clade. We refer as clade each node of the phylogenetic tree. The core-gene dS distribution of a clade represents the genome wide divergence associated with vertical transmission between two lineages (species at the left of the node vs species at the right of the node). Any dS of a TE-TE hit significantly under this distribution will be considered as resulting from a HT (see the right panel of figure 5.1). To build this core-gene dS distribution, we extracted all the single-copy BUSCO genes, on which we did similarity searches between pair of species. To reduce the workload, we did not consider pairs whose divergent time is less than 40 Myrs, and in the case of clades older than 250 Myrs, we used just one genome per subclade younger than 30 Myrs, the one with the highest number of annotated core genes. This led to 27,521 similarity searches, that we performed with the module easy-rbh of MMseq2 (Steinegger and Söding 2017).

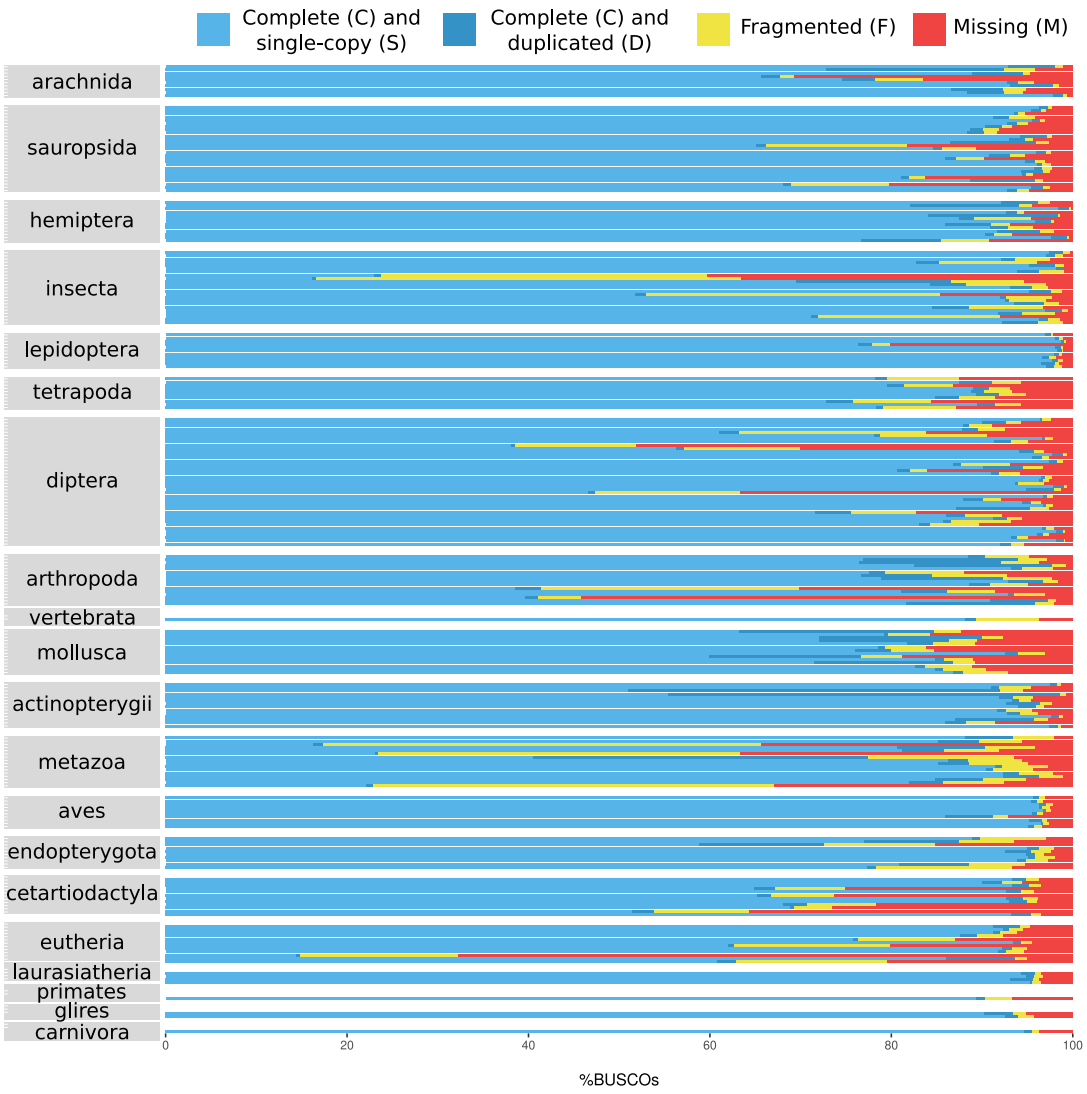


Figure 13.1: **BUSCO scores of the 247 genomes**. Each colored line represents the BUSCO score of one genome, which are grouped by database of BUSCO used for this analysis. The name of the databases are specified in the gray rectangles on the left

This module searches in the two directions of the two sets of proteins, and automatically returns the best reciprocal hit.

After extracting protein regions involved in hits, we aligned them with the Biostrings R package. We then translated these protein alignments to nucleotide alignments thanks to the corresponding nucleotide sequences. dS were then computed on each hit with Li's method implemented in the seqinr R package. Doing so, we obtained several dS values for a same gene in a same clade (because we compare several pairs of species in each clade). To obtain just one dS per gene (and per clade), Zhang *et al.* (2020) chose the dS calculated on the longest alignment for each BUSCO gene. However this dS is not necessary the most representative. This is why we chose to take all the dS values into account (or almost all), by calculating the median dS of each BUSCO gene (and for each clade). We still wanted to avoid abnormal values in this calculation, so we removed alignments smaller than 100aa, and we filtered out BUSCO genes that looked abnormal, *i.e.* we removed BUSCO genes that were

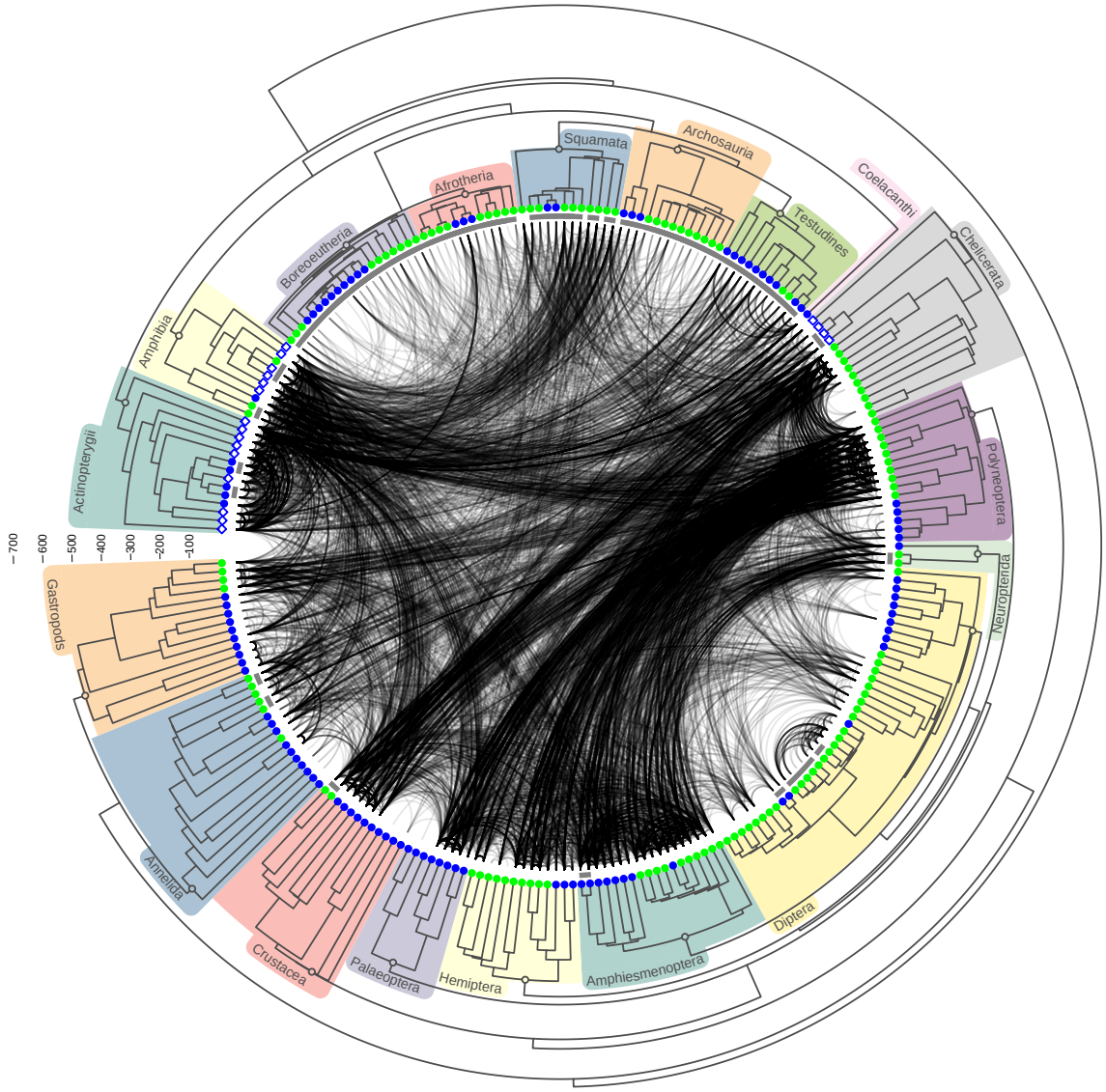


Figure 13.2: **The 11,573 independent HTT events across the phylogeny of the 247 species**. The scale of divergence time is in million years. Clade for which we did not look for HTT (because they diverged less than 40 Myrs ago or because they did not pass filter B) are grouped with a gray line. Aquatic species (fully or partial) are represented by blue dots, while terrestrial ones are in green. Empty dots represent species that use external fertilization, whereas those that use internal fertilization or those for which we do not have this information have filled dots. Only one hit per independent HTT events is represented on the tree, by a black line: the one with the best percentage of identity

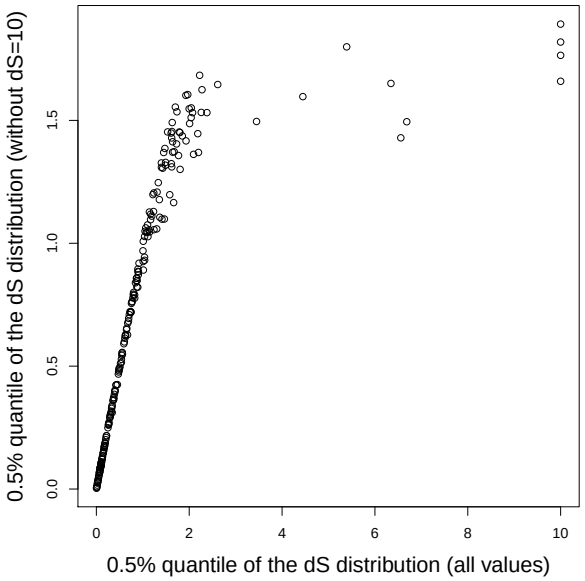


Figure 13.3: **Investigating the impact of dS values of 10 on the 0.5% quantile of the distribution of synonymous distances calculated on BUSCO genes for each pair of clades**. On the x axis, all dS values were used to calculate the median dS per BUSCO gene and per pair of clade, whereas on the y axis, dS values of 10 were discarded.

not often, or on the contrary very often, found in genomes for which we used the BUSCO database of interest (under the 5% or above the 95% quantile of the distribution). Once we calculated the median dS of each BUSCO gene (for each clade), we obtained a distribution of median dS (for each clade). Each distribution represents the rate of divergence between the respective two sister lineages since their last common ancestor. These distributions were composed of 91 values in minimum, with a maximum of 36,379 values. This number of values is higher to what is used in most studies, with for example a distribution based on 50 values in VHICA (Wallau *et al.* 2016). Some clades had a normal distribution, as expected, but many had a bi-modal distribution (a low distribution around two plus one around 6) in addition of a peak at 10. The peak is due to the fact that Li's method is not able to accurately calculate large dS values, and assigns the value of 10 in such cases. The bi-modal distribution is due to the fact that the same BUSCO gene can have a dS around 2 in a pair of species but of 10 in another pair of the same clade, which gives a median around 6. To make sure that those high values of 10 will not affect our analysis, we looked whether they have an impact on the low quantiles of the distributions of median dS. We focused on low quantiles because it is the only part of the distribution we will use. Indeed, a TE-TE hit will be considered as resulting of HT, only if (i) its dS is under a low quantile (the choice of the exact quantile is determined in section 13.6) of the median dS distribution of the BUSCO genes, and if (ii) its dS is under the absolute value of 0.5. It appears that taking into account these values of 10 or not does not change the values of low quantiles when the values are under 0.5 (example for the 0.5% quantile in figure 13.3). 0.5 being our absolute threshold, the cases above this threshold will not impact our study.

We expect to observe a correlation between the dS of BUSCO genes and the time of divergence. Here, we looked more specifically at the correlation between the low quantiles of the dS distribution (we tested the quantiles 0.1%, 0.5%, 1%, and 5%) and the time of divergence, since it is the only part of the distribution we are going to use. We obtained a correlation of 0.88 for intra-chordata clades,

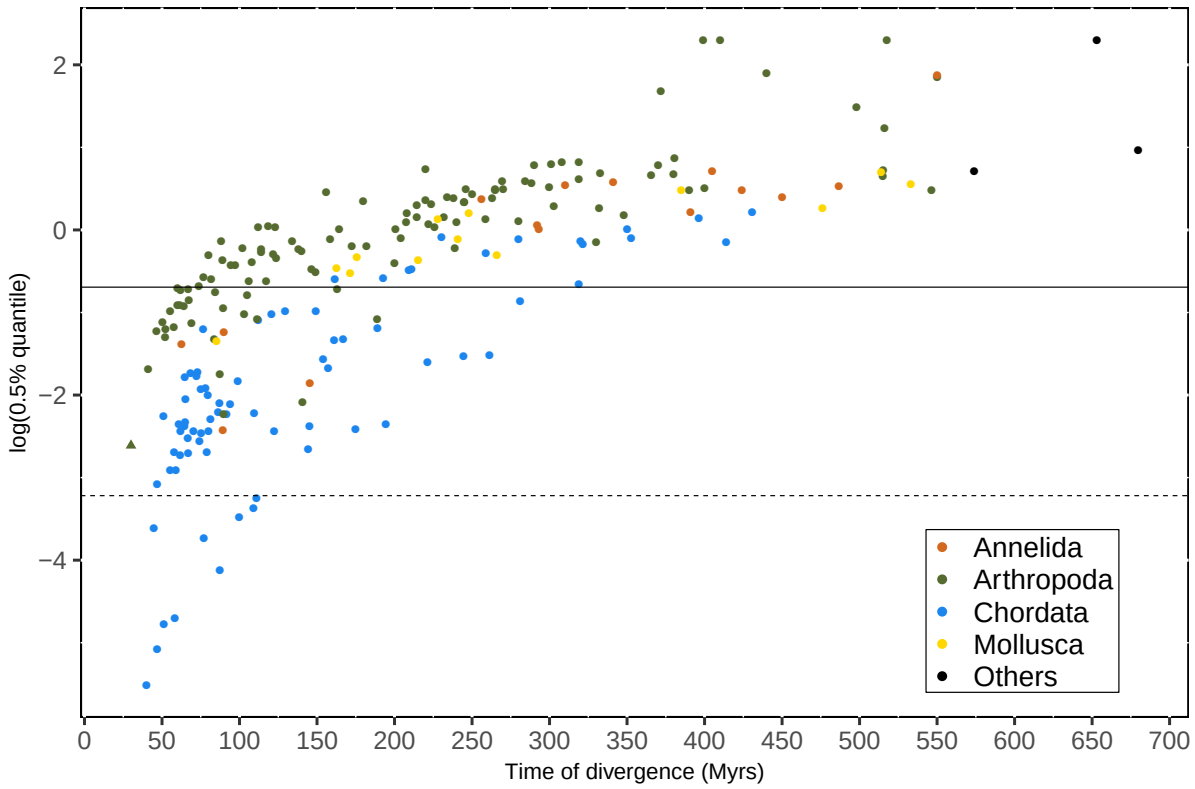


Figure 13.4: **The 0.5% quantile of the distribution of synonymous distances between BUSCO genes as a function of divergence time between clades**. The 0.5% quantile is in log scale in order to better visualize the outliers. The dash line shows the separation of the outliers (at 0.04). The full line shows our absolute threshold of 0.5. Colors indicate to what phylum each clade belongs. 'Others' indicates older clades that involve species of different phyla. The only triangle represents a clade for which we could not retrieve the time of divergence (we arbitrarily indicated 30 Myrs).

no matter the quantile, and a correlation from 0.81 to 0.71 for intra-arthropod clades for quantiles from 0.1% to 5%. In addition, we can observe that Chordata have the smallest values (figure 13.4), which is in line with their slower mutation rates (Allio *et al.* 2017; Buffalo 2021). We could identify 10 outliers, all being clades of Chordata. One outlier was a clade of Afrotheria, the 9 others of Reptilia. These clades have very low quantiles of dS, suggesting either particularly low mutation rates overall, or of just some genes evolving very slowly, or of the presence of horizontally transferred gene(s). The two last possibilities would cause a long left tail in the distribution of dS, decreasing the value of low quantiles. The fact that they are still outliers with the 5% quantile (not shown) suggests that these low values are not due to a long left tail of distribution, but rather to a distribution centered on particularly low values. Thus, these 10 clades seem to have a particularly low mutation rate overall. No matter the cause of these low values, it will impede our ability to recover HTT in these clades, leading to false negatives. Yet, we still should be able to recover many HTT events in these species, events that would have taken place between them and species of more distant clades.

In addition, we investigated whether any other clades might have BUSCO gene(s) inherited horizontally. In such a case, one can expect the low quantiles of the dS distribution to be more influenced by HT than higher quantiles. This is why we looked at the ratio of the 0.1% over the 5% quantiles. 9 clades had quite a high ratio (between 3 and and 5.2), and 1 clade had a particularly high ratio (25).

Here we did not consider clades whose 0.1% quantile is above 0.5, since it is our absolute threshold to determine whether a TE-TE hit come from HT. Looking into more details at the clade with a ratio of 25, we saw that such a big difference between both quantiles was due to 7 BUSCO genes whose dS equal 0. Such a score could indicate that these 7 genes were recently acquired horizontally. Yet, this clade of Chordata diverged only 47Myrs ago, so it is also possible that these 7 genes were inherited vertically and are very conserved. Thus, we could not detect with confidence any BUSCO genes resulting from HT with this method, which suggests that the dS distributions we calculated provide a good approximation of the divergence after speciation. Yet, if some BUSCO gene(s) of our dS distribution do come from HT, it will increase our number of false negative, not the opposite.

13.4 . TE annotation

We *de novo* annotated the 247 genomes with RepeatModeler v2 with the option LTRStruct. RepeatModeler produced output for 245 genomes but could not work on two genomes, both quite big: 1725Mb and 5161Mb. RepeatModeler did not work on this two genomes, even when giving just a subset of these genomes as input (we tried until 18% of the genome size). These two genomes are two pillbugs: *Armadillidium vulgare* and *Trachelipus rathkii*. Since *A. vulgare* has been annotated in a previous study, we downloaded the associated TE consensus (Chebbi *et al.* 2019). We concatenated these TE consensus to those of the 245 outputs of RepeatModeler generated in the present study. We also took advantage of previous annotation by adding to our database of TE consensus the TE consensus generated in two previous large-scale studies on HTT detection (one among 195 insects (Peccoud *et al.* 2017) and one among 307 vertebrates (Zhang *et al.* 2020)), plus the TE consensus of Repbase, without the SINE elements, the satellites, nor tandem (downloaded in February 2022). Doing so, we obtained one single database, which we used to annotate TE in the 247 genomes of our dataset. Using a unique database had two advantages. (i) Genomes of low quality will benefit of the annotations of their related species, although we could expect low quality to not have a major impact on RepeatModeler. This is because TE are in numerous copies, so we could expect them to be present even if the genome assembly is incomplete, and anyway RepeatModeler works with a subpart of the genome (except for the LTR detection). (ii) If a HT took place very recently, the TE burst might not have taken place yet. Since RepeatModeler detect TE based on their repetitive nature, it might miss such TE. However, these TE should be present in numerous copies in the donor genome, so the TE present in the recipient species will be annotated by the consensus reconstructed in the donor. Using this unique, but large, database considerably increases the run time of RepeatMasker. This is why we decreased the size of the database in three ways: keeping only classified TE consensus, keeping only consensus above 300bp, and clustering to remove redundancy. Clustering was achieved with the workflow easy-cluster of MMseqs2 (Steinegger and Söding 2017) with –cluster-reassign, and we tested several parameters: -c of 0.8 or 0.9, and –min-seq-id of 0.8, 0.85, 0.9, or 0.95, *i.e.* a total of 8 variations. -c is the minimum coverage and –min-seq-id is the minimum sequence identity to cluster two sequences together. To test these different parameters, we ran RepeatMasker with the 8 resulting databases on 6 genomes (from 340Mb to 2700Mb). Although the run time was reduced (greatly for the biggest genome), the percentage of genome annotated as TE was about the same regardless of the parameters used to cluster the database (see a test in which RepeatModeler was run on 217 genomes only in figure 13.5). This is why we chose the most stringent parameters to cluster

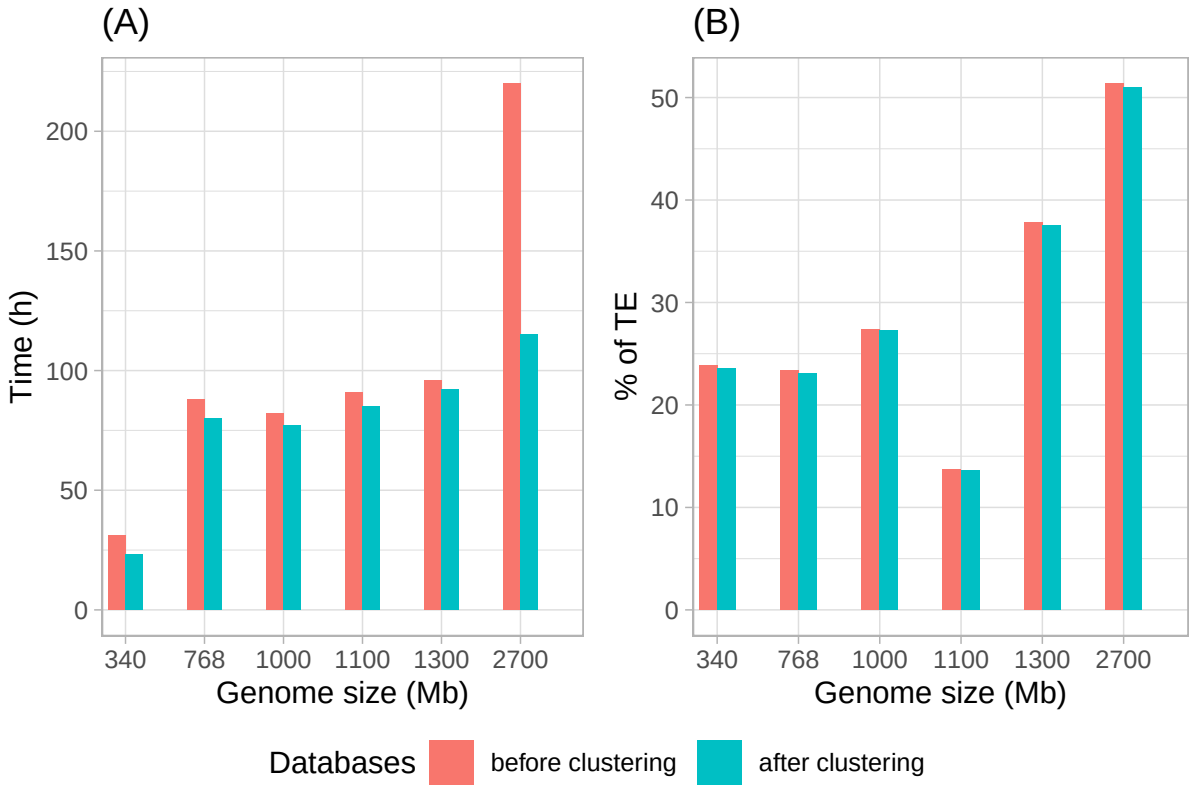


Figure 13.5: **Testing TE annotation on 6 genomes with two variations of databases**. The initial database (before clustering) was generated by concatenating the outputs of RepeatModeler ran on 217 genomes of the current study with the TE consensus of RepBase and of two previous studies, and keeping only TE consensus longer than 300bp: 247,896 TE consensus. The database after clustering was generated with the workflow easy-cluster of MMseqs2 with the options -c 0.8 and –min-seq-id 0.8, and it is composed of 185,980 TE. **A.** Run time (in hours with -pa 20) to run RepeatMasker. Both versions of a same genome were run at the same time. **B.** Percentage of the genome annotated as TE.

the database (-c 0.8 –min-seq-id 0.8), which led to a database of 217,391 TE consensus when using all the outputs of RepeatModeler.

We also identified dubious TE among this single database of TE consensus, which are TE consensus that might not be TE elements. For this, we did a similarity search using blastx of these TE elements on two databases: nr (a non-redundant database of proteins) and repeatPeps (a transposable element protein database provided in the RepeatModeler pipeline). TE consensus that do not show a homology of at least 35% over amino acids on repeatPeps and that show an homology on an nr protein over at least 90 amino acids is considered a dubious TE. We identified 27,102 such dubious TE consensus, yet we kept them in the database. We will investigate them later on, if some of them are involved in a HT.

We used this unique database of 217,391 TE consensus to run RepeatMasker on the 247 genomes of the dataset with the following parameters: -nolow -no_is -norna -engine ncbi. Copies smaller than 300bp and copies annotated with an asterisk (*) by RepeatMasker were discarded, *i.e.* copies included in higher-scoring match. Doing so, we extracted a total of 111,102,018 TE copies. In all phyla (Annelida, Arthropoda, Chordata, and Mollusca) the percentage of the genome annotated as transposable elements varies quite a lot depending on species (figure 13.6A). Yet, one can notice a

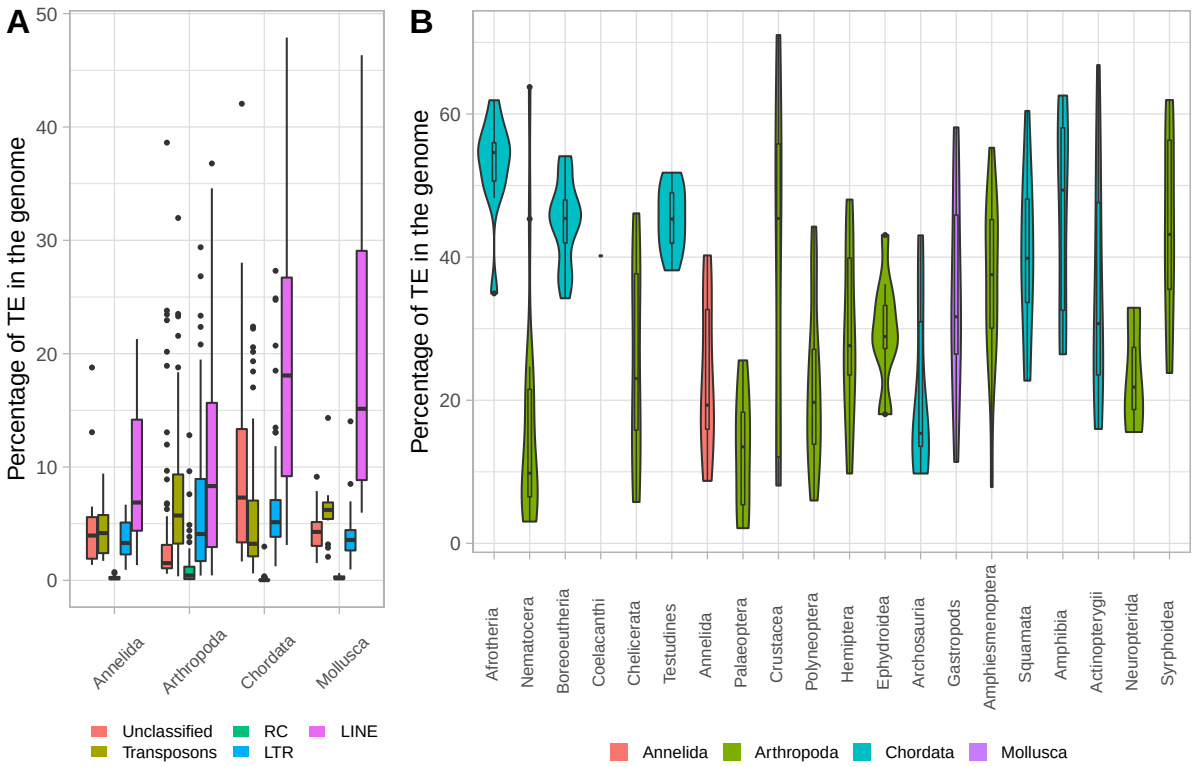


Figure 13.6: **TE composition**. **A.** Boxplot showing the percentage of annotated transposable elements in the genomes (by phylum and by subclass of TE). "Unclassified" TE correspond to TE consensus whose format name did not allow RepeatMasker to classify the copies automatically, but they do correspond to classified TE. **B.** Violon plot showing the percentage of annotated transposable elements in the genomes (by taxa).

higher percentage of LINE elements compared to the other subclasses of TE, and this is true for all four phyla overall. The percentages of LTR elements and DNA elements are quite similar within all phyla, at the exception of Mollusca which have more DNA elements than LTR elements. Although rolling-circles elements are the scarcer in all phyla, one can notice that they are more numerous in Arthropoda. Since this dataset was built around 20 "taxa", and since the number of HT will be compared within each of these taxa, we plotted the percentage of TE by taxa (figure 13.6B). One can notice a particularly low TE content in Archosauria compared to other Chordata, which correlates with the known low genome size of birds (Kapusta *et al.* 2017). This high variability of TE content in term of total amount, but also in term of relative amount for each subclass of TE, is in line with the known variability across animal species. For example, it was estimated that the Arthropoda *Drosophila melanosgaster* has about 15-22% of TE, mostly composed of LTR elements, the Actinopterygii *Danio rerio* has about 50%, mostly DNA elements, while the Chordata *Mus musculus* has about 40% of TE, mostly LINE elements (Huang *et al.* 2012).

13.5 . Similarity searches between TE copies

After extracting the TE copies longer than 300bp for each genome with bedtools getfasta (111,102,018 TE copies in total), we did reciprocal similarity searches on all possible pair of species, except those that diverged less than 40 Myrs ago. It represents 30,313 pairs of species, *i.e.* 60,626 reciprocal searches. We did not investigate pair of species that diverged less than 40 Myrs ago because

they are so related that it woud not be trivial to make the difference between TE inherited horizontally and vertically. Several tools exist for similarity searches, we chose to test three of them on our two biggest files of TE copies: blastn, MMseq2 easy-search, and MMseq2 search (options by default). Although MMseq2 easy-search was slower than blastn, MMseqs2 search was way faster. The number of hits was higher when using MMseq2. We thus decided to use MMseq2 search, and we tried several parameters in order to decrease the run time even more, without impacting the number of hits. We decided to go with the default sensitivity (5.7) but to set the option –max-seqs to 50 (set at 300 by default). This parameter decreased the run time by two, but the number of hits was similar. After the 60,626 MMseqs2 searches, we used the module filterdb of MMseqs2 to keep only the best hit with the option –extract-lines 1. We retained alignments of at least 300 bp in length, with sequence identity $\geq 75\%$, quality score ≥ 200, and with the best hit of the reciprocal search. Keeping only hits with more than 75% of identity means that we will only be able to detect recent HTT events, which is better in the context of this study since we do not want the HTT events to be older than the transitions of habitat (or than the transition of mode of fertilization). In Peccoud *et al.* (2017), where they focused on insects, they estimated that they recovered HTT events that took place in the last 10 Myrs. Here, using the same filters we can expect to recover events as old in insects, but a little older in vertebrates since their mutation rate is slower than in insects. The most recent transition of habitat in our dataset is estimated to be about 18Myrs old, and it corresponds to the transition to aquatic habitat by aquatic snakes (Galbraith *et al.* 2020).

We obtained a total of 247,248,663 hits (corresponding to 47,844,407 TE copies). To figure out which hits result from a HT event, instead of vertical transfer (VT), we calculated the synonymous distance (dS) between copies involved in each hit. We expect the dS of sequences transferred horizontally to be under the dS distribution of the rest of the genome inherited vertically (distribution calculated with BUSCO genes, see section 13.3 and figure 5.1A). dS being calculable on coding regions only, we first had to identify protein-coding regions among TE copies involved in hits. For this, we achieved five successive similarity searches with Diamond blastx of TE copies against the RepeatPeps Database. This step also allowed us to classify TE copies: copies that hit all the time against the same super family were given that super family name, the other copies were discarded. We retained only TE-TE hits involving TE protein regions ≥ 300 bp and involving TE copies of the same super family, *i.e.* 97,187,587 hits. To reduce the workload, we performed a single-linkage clustering, connecting two hits if they have a copy in common, and we kept a maximum of 2000 hits per cluster and per pair of species, choosing the ones with the highest alignment length on a coding region. This step greatly reduced the number of hits (down to 17,982,140 hits), but few clusters were concerned (0.17%). In fact, most clusters were composed of less than 2000 hits per pair of species, whereas a minority was composed of a very high number. Homologous TE regions of each retained TE–TE hit were extracted from TE copy sequences with seqtk and realigned using the Biostrings R package. Every aligned base in each TE copy was attributed a position within a codon based on the Diamond blastx alignment coordinates of TE copies on proteins. Nucleotides of undetermined or mismatched within-codon positions between copies were deleted, so were indels and resulting truncated codons. On the remaining codons, dS were computed with Li's method implemented in the seqinr R package.

13.6 . Identification of TE-TE hits resulting from HT

To distinguish which of the 17,982,140 TE-TE hits result from HT, we compared their dS to the dS distribution of single copy BUSCO genes of their respective clades.

In Zhang *et al.* (2020), the authors considered that a TE-TE hit resulted of a HT if the dS calculated between the copies involved in this hit is under the 0.5% quantile of the dS distribution calculated on BUSCO genes extracted from the clade involved in the TE-TE hit of interest (figure 5.1A). The dS calculated between the copies involved in the TE-TE hit also has to be under the absolute value of 0.5. Here, we thought it might be interesting to use a higher quantile in order to recover more HTT events, which would increase our statistical power to test the influence of the factors of interest on HTT. Using a higher quantile would of course increase the number of true positives, but it would also increase the false positives, which might add noise. Yet, the outcome of the statistical tests should not be influenced by the choice of the quantile, except if for some reason (possibly a random correlation) species of one habitat have more TE in their genomes for example. We tested four quantiles: 0.1%, 0.5%, 1%, and 5%. The main argument in favor of a lower quantile is the fact that for clustering steps (see section 13.8) it is very important to have only hits resulting of HT. Hits from VT might have the consequence to connect hits from independent HT events, decreasing our estimation of HT events. A second argument is that it would be difficult to computationally handle the number of hits recovered with higher quantiles (5.8 millions hits with the 0.1% quantile *versus* 7.5 millions with the 5% quantile). Altogether, we chose to not increase the quantile and to keep the 0.5% quantile.

Thus, we retained TE–TE hits for which the dS value was lower than the 0.5% quantile of the dS distribution of core genes of the appropriate clades. We also removed all hits whose dS value was ≥ 0.5, or computed on less than 100 codons. At this point, we obtained 6,277,064 hits, all resulting from HT. However, the number of HT events is much lower and clustering steps are necessary to obtain an estimation (see section 13.8).

At this step, we can already notice a lack of TE-TE hits resulting from HT in 11 species of our dataset. These 11 species belong to various taxa (4 Nematocera, 3 Chelicerata, 2 Crustacea, 1 Palaeoptera, and 1 Annelida) and to both habitats. Actually, this lack of HTT is more probably due to their low TE content (from 2.3% of their genome to 10.85%), associated to small assembly sizes (from 32Mb to 474Mb, with a median of 92Mb). They all have a BUSCO score above 90%, at the exception of the smallest assembly which has 67.8% of complete single BUSCO genes. These high BUSCO scores indicate that the assembly sizes are close to the real genome sizes.

13.7 . Refining hits

MMseq2 has the advantage to be fast, but it gives an estimated percentage of identity, whereas we are going to need precise values for the clustering steps. This is why we then run blastn (with the task dc-megablast), which was computationally possible at this step because we focused only on the copies involved in HT (2,271,889 copies out of 111,102,018), on pairs of species for which we detected HTT (10,147 pairs out of 30,313), and we performed the searches separately by TE super family. When running MMseq2, we also had to focus on the best hit, also for computation reason, but for this similarity search, we were able to increase the option -max_target_seqs to 100. Doing

so, each target sequence is allowed up to 100 hits, which should help forming more bridges during the clustering steps. We used the same method as with MMseq2, doing reciprocal searches, keeping only hits with more than 75% identity on more than 300pb with a score above 200, and keeping the best hit for a pair of query-subject. Most of the resulting hits at this step should result of HT, but among the additional hits some can result from vertical transfer, so we had to go through all the steps following MMseq2 again (see red arrows in figure 12.1): keeping hits covering a coding region of 300bp or more (we did not have to run again the steps with blastx though), performing single-linkage clustering to keep a maximum of 2000 hits per cluster and pair of species, calculating dS for each retained hits, and selecting hits that result from HT. This refining step allowed us to increase the number of hits involved in HT to 17,983,960, which should increase the performance of clustering. In fact, if we filter out too many hits, we would remove hits allowing to form bridges between other hits, under-performing the clustering step.

13.8 . Clustering

Clustering is very important at two levels: (i) since TE are in numerous copies, a single HT event will lead to many TE-TE hits, and (ii) several TE-TE hits could correspond to two non-overlapping parts of the same TE (resulting from the same transfer). We tried clustering with two approaches: clustering by pair of species or by clade. Clustering by pair of species consist in looking at hits by pair of species, and looking whether some hits of that pair could correspond to the same HTT event. In this approach we repeat the same method as many times as we have pair of species. When clustering by clade we do not look at the hits of just a pair of species, instead we look at the hits of all the species of each clade (or each node of the tree). In this case, we investigate whether the copies of the left side of a node could correspond to the same HT event than the copies of the right side of the node. Zhang *et al.* (2020) clustered by clade, which has two advantages: copies of all species of a same clade are taken together, which improves the efficiency of clustering (since more copies allow to form more bridges), and it is a first way to count the independent number of HTT events (see section 13.10). On the other hand, clustering by pair of species has the advantage to recover HTT events independently for each pair, which removes the bias induced by sampling. In a clustering by pair of species, the count of each HTT event in a pair of species would depend on these two species only, and will not be influenced by the other species of the dataset, which is what we need in this study (see section 13.11).

Before clustering we had to reduce again the number of hits for computational reasons. We used the same method as previously, connecting all hits that have a copy in common with single-linkage clustering. We did not keep more than 200 hits (instead of 2000 in the previous steps) per cluster and per pair of species. Such a clustering decreases the number of hits from 17,983,960 to 4,618,455 hits. Here again, only a minority of clusters are concerned. To note, the goal of refining hits was to increase the number of hits for the cases which had only a very small number of hits, so those which are not reduced at by this step.

Regardless of whether we clustered by pair of species or by clade, we used the same two steps as in Zhang *et al.* (2020): the first one takes into account that TE are in numerous copies, and the second one takes into account that several TE-TE hits might correspond to two non-overlapping parts of the same TE. Yet, the approach for both steps of clustering were quite different depending on our

clustering approach.

Clustering by pair of species was simpler and computationally faster. The first step consisted in comparing the percentage of identity of copies inter- and intra-species. When resulting of the same HT event, one can expect the intra-species percentage of identity to be higher than the inter-species percentage of identity among TE copies. We already know the inter-species percentage of identities, they are the ones of the TE-TE hits. Regarding the intra-species percentage of identities, we had to calculate them. For this, we extracted all copies involved in TE-TE hits per species and per TE family, and we did a similarity search of each file against itself (options: -task dc-megablast -max_target_seqs 100000 -max_hsps 1). We built graphs by connecting every two hits whose intra-species copies had a higher percentage of identity (in at least one species of the pair) than at least one of the two inter-species percentage of identities (one percentage per hit), which we refer as criterion 1. Then we used the algorithm cluster_fast_greedy of the R package igraph which is able to find community structure in a graph, which allowed us to obtain 267,745 communities of hits. For the second step of the clustering, we were very parsimonious, and we considered that any communities whose copies do not have any nucleotidic homology but have homology with a same TE family, without overlap (or an overlap <100aa), might correspond to the same HTT event. For the homologies with TE families, we used the output of blastx of the TE copies against the proteins of RepBase (generated previously, see figure 12.1). To know whether two proteins of RepBase overlap, we did similarity searches between each pair of proteins involved, using blastp. Clustering allowed us to obtain 141,389 hit groups. However, 40% of these hit groups are supported by one hit only. One can expect all hit groups to be supported by several hits because TE are in numerous copies so a single HT event should lead to several hits between the two involved species. This is why we were puzzled by the high number of hit groups that contained only one hit. We could think of three explanations regarding hit groups supported by just one hit: (i) a genome has so little TE copies that we recovered only one TE-TE hit for this HTT event, (ii) both genomes of the pair have many TE copies but we do not recover TE-TE hits for all of them because of the stringent filters we used at each step of the pipeline, and (iii) the TE-TE hit we recovered as a HT would actually correspond to the left tail of the distribution of dS calculated between TE copies inherited vertically. This last possibility would correspond to false positives. To have an idea of the real explanation, we randomly investigated four cases of hit groups supported by one TE-TE hit each. In the two first examples, we found that one genome of the pair has only two and three TE copies for the associated TE consensus, supporting our first hypothesis. In the second example, we found in both genomes more than 900 TE copies for this TE consensus, and more than 900 TE-TE hits with MMseq. However when keeping only TE-TE hits whose TE copies have a coding region of at least 300bp, only one copy remained for each genome, supporting our second hypothesis. These three first examples are very reassuring and suggest that hit groups supported by one TE-TE hit are *bona fide* HTT. However in the fourth example, both genomes of the pair have more than 1000 TE copies for the associated TE consensus, half of them are longer than 300bp, yet only 18 TE-TE hits are recovered with MMseq2. This means that most percentage of identities were under 75% between copies of this consensus. 12 hits remained when keeping only those covering a coding region >300bp, and 2 hits left when keeping only those whose dS is under the 0.5% quantile of the dS distribution obtained with BUSCO genes. This last example might thus correspond to a false positive. To avoid such cases, we did some evaluations (see section 13.9).

For clustering by clade, we used the same ideas as by pair of species, but to evaluate the first step we had to generate one file of TE copies for all nodes in the tree (instead of one per species) and per superfamiliy of TE. Then we did as many similarity searches as the number of files, using the same file for query and subject. Contrary to clustering by pair of species, we evaluated the first step in two rounds. In the first round we focused on young nodes only (clades that diverged in the last 40Myrs, for which we did not even look for HT), and we connected every two hits passing criterion 1, similarly to clustering by pair of species, which allowed us to form 185,677 communities. In the second round, we investigated whether every two pairs of communities could correspond to a single HTT event. Here however, criterion 1 had to be passed by $\geq 5\%$ of all possible pairs of hits taken from the two communities, to which we added a second criterion, which relies on the inferred age of the transfer. To reflect the same HTT, the transfers (represented by the two communities of hits) should not be more recent than both clades. For this, we compared the dS of TE-TE hits within communities to the dS of BUSCO genes of the involved clades. If both criteria were passed, and if they could correspond to two non overlapping parts of a same TE (determined the same way as in clustering by per of species), the two communities were connected. Finally, we used complete-linkage clustering to delineate hit groups. Doing so, we obtained 59,870 hit groups, with only 23% of them supported by one hit only (contrary to 40% when clustering by pair of species). This result confirms that clustering by clade is more efficient than per pair of species. Yet, this 23% of hit groups supported by just one hit might also correspond to false positives, at least partially.

13.9 . Testing false positives

As discussed in the previous section, some of the hit groups supported by just one hit might correspond to false positives, regardless of how we achieved the clustering. As an additional argument in this sense, we randomly sampled 5000 hit groups supported by just one hit, and 5000 supported by more than one hit, and we plotted these transfers on the phylogenetic tree of our dataset (not shown). We could clearly see an over representation of hit groups supported by only one hit among mammals and reptiles, the two taxa for which we were already concerned about false positives. This is because these taxa contain closely related species in our dataset, and because their rate of mutation is lower than in other taxa (figure 13.4). For these two reasons, some pairs of species in these taxa have very low BUSCO gene dS and our power to discriminate between horizontal and vertical transmission will be very limited. As a first security, we chose to keep only hit groups supported by at least five copies (filter A in figure 13.7). This filter removed 71,279 of the 141,389 hit groups clustered by pair of species (54.4%). When clustering by clade, this filter removed 20,554 of the 59,870 hit groups (34.3%). When clustering per clade, copies of all species of the clade are adding up, this is why filter A removed way less hit groups with this method than when clustering per pair of species. Yet, it does not mean there were more false positives when clustering per pair of species.

Also concerned by false positives in closely related clade when working on vertebrates, Zhang *et al.* (2020) chose to not recover HTT between species that diverged in the last 120Myrs, instead of 40Myrs in their previous study on insects (Peccoud *et al.* 2017). They chose this threshold of 120Myrs after plotting the average ratio dN/dS of TE-TE hits of class II elements as a function of the time of divergence. They suggested that class I elements that diverged through transposition within genomes undergo purifying selection because of a cis-preference (functional TE express retrotransposon proteins

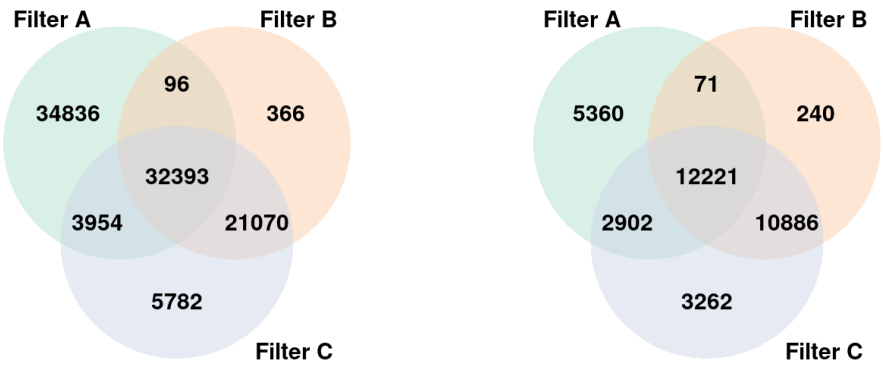


Figure 13.7: **Venn diagram of the number of hit groups removed with each filter**. Numbers are the ones following clustering per pair of species (left) or per clade (right). **Filter A**: remove hit group with less than 5 copies. **Filter B**: remove all hit groups of a clade for which the median dN/dS of TE-TE hits of class II elements is above 0.9. **Filter C**: remove hit groups whose dS distribution might correspond to the left tail of a normal distribution, expect if its maximum dS is far under the 0.5% quantile of the dS distribution calculated on BUSCO genes.

that process their own mRNA), whereas class II elements do not have any kind of cis-preference, and thus no sign of purifying selection is detected on their sequences over vertical transmission. However for HT, one can expect only functional copies to transfer (both for class I and class II elements), which should lead to a signal of purifying selection. Looking at the ratio dN/dS of class II elements can thus be used as a proxy to assess whether a TE was inherited horizontally or vertically. In Zhang *et al.* (2020), they observed a ratio under 1 for class II elements (indicator of HT) only for clades older than 120Myrs. For hit groups involving younger clades, they obtained median dN/dS ratios very close to 1, suggesting that at least some of the HTT they inferred between these younger clades are false positives and rather correspond to vertical inheritance. In this study, we chose to firstly recover HT between any pair of species older than 40Myrs, and we then used a method similar to the one used in Zhang *et al.* (2020) to choose the final threshold. More specifically, we plotted the median dN/dS ratio of TE-TE hits of class II elements as a function of the median of the distribution of BUSCO gene dS (figure 13.8), instead of the divergence time. We did this because the divergence time can be imprecise in some cases, and because a same divergence time in different phyla has very different evolutionary implications. In figure 13.8 one can observe purifying selection in divergent clades (median dN/dS < 1) which supports that TE-TE hits of class II elements recovered in these clades are *bona fide* HT. Yet, one can observe some medians near 1, and even above, suggesting that at least some hit groups of these clades are false positives. We decided to remove all hit groups of a clade for which the median dN/dS of TE-TE hits of class II elements is above 0.9. This filter removed 53,925 hit groups when clustering per pair of species (38.1%). When clustering per clade, it removed 23,418 hit groups (39.1%) (filter B in figure 13.7). The fact that this filter remove about the same percentage of hit groups when clustering per pair of species or per clade is a good indicator that both approaches led to similar results.

Another way to check whether a hit group is a false positive is to look at the dS distribution of the TE-TE hits composing this hit group. If it is a true positive, one can expect a normal distribution. If it is a false positive though (and thus a result of vertical transmission, rather than horizontal transmission), one can expect that we recovered TE-TE hits having dS corresponding to the left tail

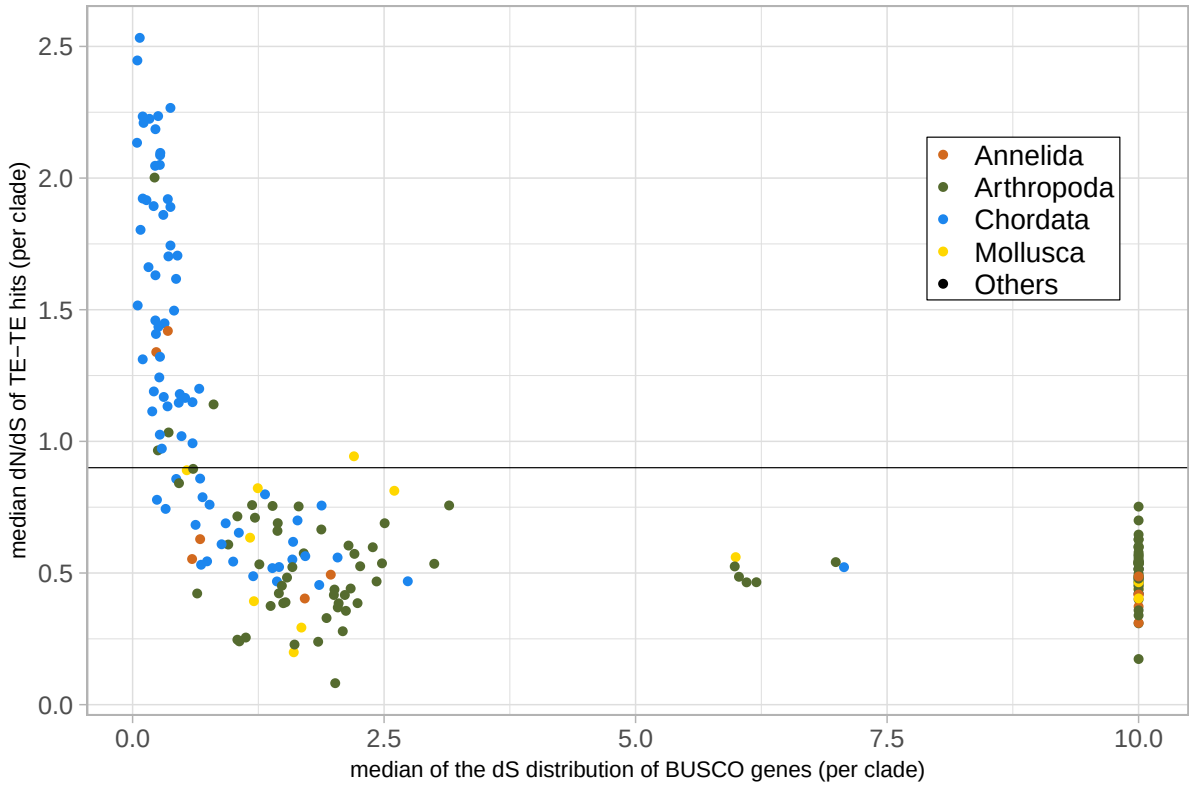


Figure 13.8: **Median ratio dN/dS of TE-TE hits of class II elements as a function of the median of the dS distribution of BUSCO genes**. Each dot represents one clade, and its color indicates its phylum. "Others" means that the clade involves different phyla. The black line indicates the threshold of 0.9, above which we removed all hit groups of a clade.

of the normal distribution of dS generated by vertical inheritance. We used the script implemented in Peccoud *et al.* (2017) to recognize such distributions in each hit group. We removed hit groups whose dS distribution looks like it is truncated and might correspond to the left tail of a normal distribution, except if its maximum dS is far under the 0.5% quantile of the dS distribution calculated on BUSCO genes (under this value minus 0.2). If the maximum dS is that low, one can be confident that this hit group is a true positive, regardless of its distribution. This method removed 63,199 hit groups when clustering per pair of species (44.7%). When clustering per clade, it removed 29,271 hit groups (48.9%) (filter C in figure 13.7). Here again, the fact that this filter remove about the same percentage of hit groups when clustering per pair of species or per clade is a good indicator that both approaches led to similar results.

When using the three methods together (filters A, B, and C), we removed a total of 98,491 hit groups out of 141,389 when clustering per pair of species (69.7%), including 32,393 that would have been removed with any of these methods. When clustering per clade, the three methods together removed 34,942 hit groups out of 59,870 (58.4%), including 12,221 that would have been removed with any of these methods (figure 13.7). Filter B is very redundant with filter C, which is very reassuring: using two independent methods removes roughly the same hit groups. Among the removed hit groups, 5762 or 3262 (per pair of species or per clade) are specific to filter C which suggests that this method might be more efficient to detect false positives than filter B. 34,836 or 5360 (per pair of species or per clade) removed hit groups are specific to filter A. This is not unexpected

since we show in section 13.8 that at least some hit groups have few copies simply because there are few copies in the genome, or because the copies to not pass all the stringent filters of our pipeline. So we probably removed many true positives with filter A, *i.e.* hit groups not inherited vertically. We though safer to keep filter A, mostly because this filter can also prevent false positives due to contamination. In the case of contamination one can expect the assembled genome to harbor only a small number of the contaminant copies (contaminants usually represents a small fraction of the overall DNA).

To summarize, we will continue the analyses with 42,898 hit groups when clustering per pair of species and with 24,928 hit groups when clustering per clade.

A last thing to check are the dubious TE. Among all TE consensus, we identified 27,102 dubious TE, which are sequences for which we are not sure they are really TE (see section 13.4). In our dataset clustered per pair of species or per clade, we have 28 and 29 of these TE, respectively. They are involved in 101 and 89 hit groups, respectively. We will have to investigate more in details these 29 TE consensus to figure out whether they are really transposable elements.

13.10 . Counting independent HTT events

We aimed at estimating a minimum number of independent HTT events that took place among the 247 genomes, taking into account the fact that several species of our dataset share some events. This can happen either when a HTT took place in the common ancestor of several of our species, or when the donor species is quite related to other species of our dataset which might also have the TE involved in HTT. In the statistical analysis, we plan to pick species 2 by 2, so we do not need this step (see section 13.11). Yet, we still would like to have an idea of the number of independent HTT events in our dataset, so we can compare this number with the 2248 events inferred in Peccoud *et al.* (2017) and the 975 HTT inferred in Zhang *et al.* (2020).

For this, we used the table of hits that was clustered by clade, and we investigated whether some hit groups could be explained by other hit groups. If this is the case, one can expect the copies of a hit group resulting from a transfer to be similar to the hits of the other hit group of the same transfer group. In addition, the species involved in these two hit groups should be related. We obtained this information thanks to the outputs of similarity searches of TE intra-clades generated at the step of clustering by clade.

Doing so, we estimated that 11,573 independent HT events took place in our dataset (figure 13.2). These events did not necessary involve a direct transfer from one of our species to another, instead the transfer might have taken place in an ancestor, from a species not present in our dataset but related to a species of our dataset, or it might be an indirect transfer that involved an intermediate species. More specifically, we found 6660 transfers within Arthropoda, 2292 within Chordata, 148 within Mollusca, 101 within Annelida, and 2372 between species of different phyla. We then normalized these numbers of HTT events per the number of pair of species between which we looked for HTT (figure 13.9). Doing so, Mollusca have the highest number of HTT events per pair of species (1.63), followed by Annelida and Arthropoda which have a similar number (0.85 and 0.89), then Chordata (0.65), while pairs of different phyla have the lowest number of HTT (0.13). Such a lower rate of HTT between species of different phyla is in line with the positive correlation between the number of HTT events and the phylogenetic proximity found in Peccoud *et al.* (2017) at the scale of insects.

In addition, it is expected to find more HTT in Arthropoda than in Chordata since previous studies found 2248 HTT events among 197 insects and 975 HTT events among 307 vertebrates (Peccoud *et al.* 2017; Zhang *et al.* 2020). Yet, the numbers recovered here could look very high in comparison of these two study. Two main reasons can explain why we recovered more HTT in the present study: (i) we improved the pipeline at each step in order to recover more HTT, and (ii) in the case of Arthropoda, the previous study was composed of many related species, preventing the authors from looking for HTT between many pairs. In the case of Chordata, 66.2% of the HTT involved an Actinopterygii and 39.6% of the HTT involved an Amphiba. 10.8% of the transfers involved neither an Actinopterygii nor an Amphiba. Such a high proportion of HTT in Actinopterygii is in line with the study of Zhang *et al.* (2020) in which 93.7% of the HTT they recovered involved Teleost fishes (a group of Actinopterygii). Although this proportion is lower in the present study, it can simply be explained by a lower proportion of Actinopterygii in our dataset (14.5% of the Chordata vs 20.8% in Zhang *et al.* (2020)) and a higher proportion of Amphibia in our dataset (10.6% of the Chordata vs 1.6% in Zhang *et al.* (2020)). Regarding Annelida and Mollusca, no large-scale studies ever tried to recover HTT in these phyla, although some examples were reported in Mollusca at least (Metzger *et al.* 2018). It is thus very interesting to see that Annelida are involved in about as many HTT events as Arthropoda, and that Mollusca are involved in twice as many (figure 13.9). In addition, most HTT events we recovered involved DNA elements (83.8%), and this for all phyla (figure 13.9). This result is reassuring for two reasons. The first reason is that it was already shown that DNA elements have more HT than retrotransposon (80.6% in Peccoud *et al.* (2017)), and the second reason is that the most abundant annotated TE in the assemblies are LINE elements (figure 13.6A). If our pipeline would recover contamination as HT, it would be biased to find more HT among the most abundant subclass, *i.e.* LINE elements. Thus recovering more HT of DNA elements, rather than RNA elements, is a good indicator that the HTT events we recovered in this study are *bona fide* HT, instead of contamination.

Regarding the two factors we are testing here (the habitat and the mode of fertilization), preliminary results tend to indicate an absence of correlation between the habitat and the number of HTT events, but the presence of a correlation between the mode of fertilization and the number of HTT. More precisely, we found as many HTT events between aquatic species than between terrestrial species, yet we found less HTT events between species of different habitats (figure 13.10A). It would be interesting to assess whether the HTT between species of different habitats are more often from an aquatic species to a terrestrial one, or whether it is the opposite. When looking at taxa individually (figure 13.10B), we could not observe a global trend, some taxa having more HT between aquatic species (e.g. Afrotheria, Amphibia, Archosauria, Testunides), others having more HT between terrestrial species (e.g. Chelicerata, Crustacea, Gastropoda, Hemiptera, Polyneoptera). It would be interesting to look at the habitat more precisely: some "aquatic" species are actually just partially aquatic, and among the aquatic ones some live in freshwater whereas others live in a marine environment. Regarding the mode of fertilization, only four taxa have aquatic species with internal fertilization: Actinoptergyii, Amphibia, Coelacanthi, and Amniota (Boreoeutheria + Afrotheria + Squamata + Archosauria + Testudines). We looked at the number of HTT events between aquatic species of these four taxa (looking at both modes of fertilization), and we found more HTT events in pairs of species both using external fertilization than in pairs of species using different modes of fertilization, and even less in pairs of species both using internal fertilization (figure 13.10C). This

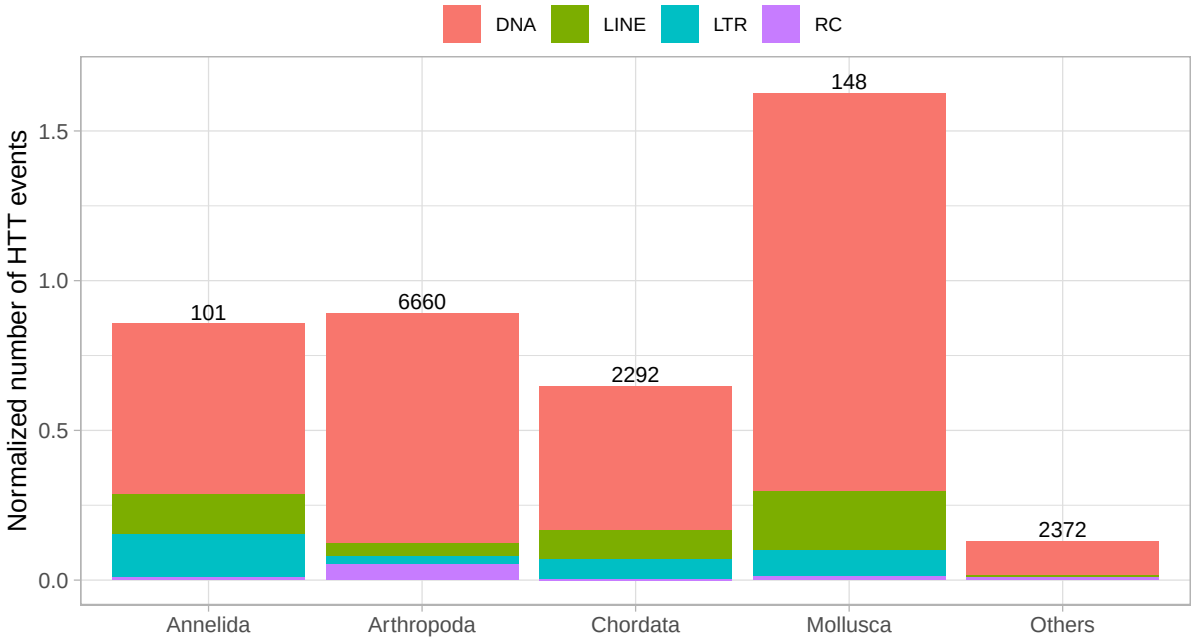


Figure 13.9: **Normalized number of HTT events per phylum**. The number of HTT events was normalized by dividing the absolute numbers by the number of pairs of species among which we looked for HTT for each phylum. Absolute numbers of HTT events are shown on top of each bar. "Others" are HTT that took place between species of different phyla. Colors represent the subclass of TE.

trend is true for all the taxa we investigated. Among Actinopterygii, there are 15 times more HTT events when both species use external fertilization than when both species use internal fertilization. For Amphibia, there is only one aquatic species that use internal fertilization in our dataset, so we do not have any pair of species both using internal fertilization. Yet, we observed twice more HTT events when both species of the pair use external fertilization than when one species of the pair use external fertilization and the other one use internal fertilization. For the other pairs of species of interest (Actinopterygii vs Amphibia or non-Amphibia Sarcopterygii (Coelacanthi+Amniota) vs Actinopterygii or vs Amphibia or vs another non-Amphibia Sarcopterygii), we observed 11 times more HTT events when both species of the pair use external fertilization than when both species of the pair use internal fertilization. Although these preliminary results suggest a positive effect of the external fertilization on the number of HTT events, we cannot conclude on the implication of this factor yet, nor on the factor of the habitat. Instead, we will have to test these two factors with a proper statistical analysis (see section 13.11).

13.11 . Statistical analysis

Although the preliminary results hereinbefore give us an idea of the implication of the habitat and of the mode of fertilization on the number of HTT, a proper statistical analysis is necessary. This analysis will allow us to assess the impact of these two factors independently of the phylogeny, and taking into account possible biases due to sampling. We did not perform this statistical analysis at the time of the writing of this manuscript. Yet, the design of this statistical analysis is the first thing we thought about, which allowed us to appropriately design a strategy to sample genomes, compatible

with the statistical analysis. I will describe hereinafter the approach we are planning to use.

To not be biased by the sampling, we will use the table of hits that was clustered by pair of species. Doing so, each pair of species is totally independent, and the count of HTT does not depend on other species. We will generate a 247*247 half matrix to stock the number of HTT events for each pair of species. Then, to not be biased by a different number of species in different taxa, and to avoid to count twice the same HTT event (that would be recovered in several related species of our dataset), we will compare species two by two (one aquatic vs one terrestrial). For this, we will randomly sample one aquatic and one terrestrial species per taxa, and we will subset HTT that involve these species in our matrix. Doing so, we will work with only a subset of HTT. If this number is too low, we might not have enough statistical power to conclude on the implication of our two factors. This is why we tried to improve the pipeline at each step, trying to increase the number of HTT we can recover. On the subset of HTT involving the species we randomly picked, we will look how many pairs have more HTT events in their aquatic species compared to the terrestrial one. These species are not necessarily representative of their group, but doing this sampling many times will work like replicates, and will allow us to obtain a distribution. Then, we will compare this distribution to the expected one to conclude on the implication of our two factors on the number of HTT events. To obtain this expected distribution, we will permute the type of habitat among our dataset, randomling assigning the aquatic or the terrestrial habitat for each species, and we will then process the data the same way as for the observed distribution: sampling one aquatic and one terrestrial species per taxa, and looking how many pairs have more HTT events in their aquatic species compared to the terrestrial one. Repeating this about 100 times (the permutation and the measure) will allow us to obtain the expected distribution under our null hypothesis: no impact of the habitat on the number of HTT. If our distribution is significantly different, it will suggest that habitat does have an effect on the number of HTT (the terrestrial or the aquatic one depending on the direction of the shift).

In the case of "aquatic" species which are not fully aquatic, it would be interesting to achieve the test with and without them. If we find an effect of the aquatic habitat on the number of HTT events, these species could help us understand at what stage of the development the aquatic habitat favors the number of HTT.

Regarding the second factor, the mode of fertilization, we will use the same approach but at a smaller scale since we can test this factor in just a subset of taxa. The taxa we will be interested in are the taxa in figure 13.10C. Taxa for which all aquatic species use external fertilization and all terrestrial species (which necessarily all use internal fertilization) are not so informative because of the correlation between both tested factors. Actinopterygii is thus the most interesting taxa with 9 species using external fertilization and 5 species using internal fertilization. The second most interesting taxa are Amphibia, with 6 species that use external fertilization and 4 species that use internal fertilization, although only one is aquatic. We also have one species of Coelacanthi which uses internal fertilization and all the species of aquatic Amniota which all use internal fertilization.

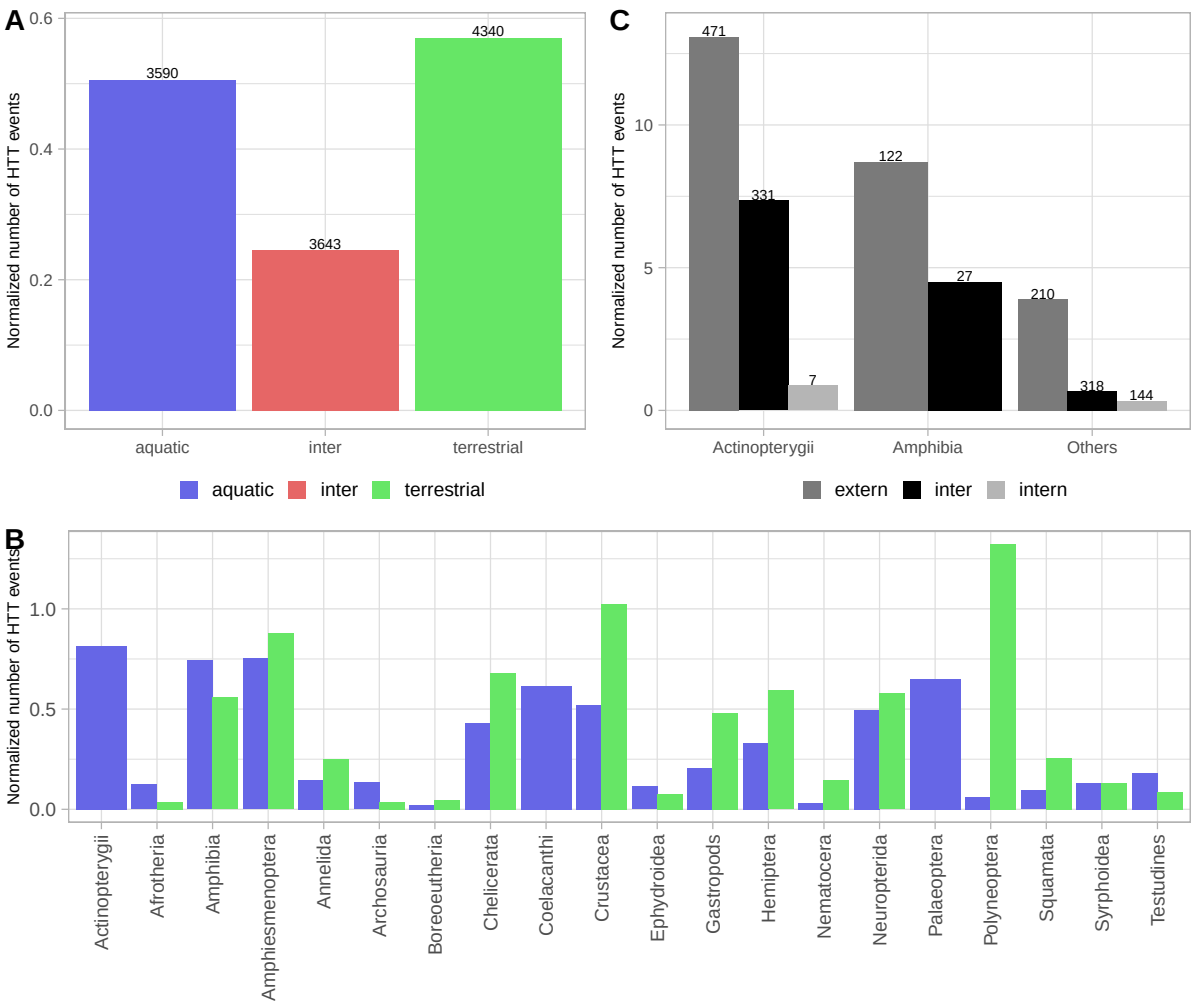


Figure 13.10: **The habitat and the mode of fertilization as possible factors involved in HTT**. **A.** Normalized number of HTT events between pairs of species living both in aquatic habitats (in blue), both in terrestrial habitat (in green), or in different habitats (in red, legend "inter"). **B.** Normalized number of HTT events in aquatic or terrestrial species of each of the 20 taxa (without taking into account the taxa nor the habitat of the paired species). **C.** Normalized number of HTT events in a subset of the dataset: Actinoptergyii, aquatic Amphibia, Coelacanthi, and aquatic Amniota. "extern" corresponds to HTT between a pair of species both using external fertilization, as opposed to "intern". "inter" corresponds to HTT between a species using internal and another using external fertilization. "Actinopterygii" and "Amphibia" show the number of independent HTT events when both species of the pair belong the the taxa of interest. "Others" shows the number of independent HTT events between all other possible pair of combination: an Actinoptrygii vs an Amphibia, a non-Amphibia Sarcopterygii (Coelacanthi+Amniota) vs an Actinoptrygii or vs Amphibia or vs another non-Amphibia Sarcopterygii.

14 - The phylogenetic and geographical proximity as factors promoting HTT

Although Peccoud *et al.* (2017) found that phylogenetic relatedness and geographical proximity may facilitate HTT in insects, they were able to test these two factors only at a large scale. To test the impact of these two factors at a smaller scale, we designed a large-scale study in which we produced our own dataset. This study was also in progress at the time of the writing of this manuscript.

This project was already started before the beginning of my PhD: insects were already collected at two sites, and the experiments were already in progress for these two sites (Cameroon and French Guiana). I collected myself the insects of two other sites (in France and in Spain), and I achieved all the experiments (from DNA extractions to genome sequencing) and the genome assemblies of these two sites plus three others (USA, Costa-Rica, and Sweden). Once the genomes of the seven sites assembled, I took care of the following analyses.

14.1 . Designing the dataset

The initial plan was to sample insects in the wild at 10 sites around the world (in order to test the geographical proximity), and at each site, we planned to sample 3 Diptera, 3 Lepidoptera and 3 Hymenoptera (in order to test the phylogenetic proximity: for example, are there more HTT between the Diptera than between a Diptera and a Lepidoptera?). However because of various impediments (the COVID pandemic and delays with collaborators for sampling), we were able to collect samples from seven sites only (figure 14.1): 63.26N, 19.02E (Sweden), 48.70N,2.14E (France), 36.61N,-6.24W (Spain), 46.35N,-121.52W (USA, WA), 8.70009N,-83.20175W (Costa Rica), 5.34000N,11.33000E (Cameroon), 4.08800N,-52.68010W (French Guyana). Each sample was named after its site in the main language of the site (S for Sweden [Sverige], F for France [France], E for Spain [España], W for Washington, CR for Costa Rica [Costa Rica], C for Cameroon [Cameroun] and G for french Guyana [Guyane] and its taxonomic order (L for Lepidoptera, H for Hymenoptera, and D for Diptera), followed by a number.

We captured most insects with an insect-exhauster or a net, without using any traps. Although it is possible to use traps that attract insects with different stimuli (light or food for example), or that catch them in their fly, the use of a single trap tends to limit the diversity of insects we would trap. Since we wanted a diversity as high as possible, among our three orders of interest, we simply captured any insects of these orders we would see at the site. At the two sites I sampled myself, I did use three kind of stimuli, in addition of captures without stimuli, in order to capture insects I could not have captured otherwise: various flowers naturally at the site, a light at night (which attracts moths for example) and myself at sundown (which attracts mosquitoes). At each site, the perimeter of capture was quite reduced, allowing us to estimate that all insects captured at this site are sympatric. As an example, among the two sites I sampled myself, the two most distant samples were captures at 2.8 km in France and at 3.7 km in Spain. The insects were individually stocked in alcohol until the time of DNA extraction. To note, the insects were not precisely morphologically identified, so their identification is based at the genetic level only (see section 14.5). The application

Figure 14.1: **Map of the sites sampled for this study.** In blue are the sites where we collected the insects ourselves in the wild and did a *de novo* assembly of their genomes. In yellow is the site where we sampled genomes directly from the website Darwin Tree of Life. This figure was designed with the website https://app.datawrapper.de/select/map.

iNaturalist (available from https://www.inaturalist.org) was very helpful during sampling thanks to its ability to recognize organisms based on a picture. Although we did not directly use this identification in the study, it has been very helpful on the field to give me an idea of what I should capture next.

Before DNA extraction, we took a picture of each insect. In the case of the two sites I sampled, I took the pictures the day of the capture, before stocking the insects in alcohol. Then, we crushed the whole body of each insect, or a part of the body depending on its size (see column "Extracted" of table 14.1), and we performed DNA extraction with the kit Nucleobond AXG20. In the case of the five sites I took care of, I did the extractions on a single individual in order to reduce polymorphism, which will help when assembling genomes. I had to do the extractions on several individuals for three samples though because they were too small: I pooled two to four individuals (see the number in column "Extracted" of table 14.1). We looked at the integrity of DNA on an agarose gel, checking that the fragments were high enough for long read sequencing. Then, I sequenced each COI gene with Sanger. This amplification did not work on seven samples, but it allowed us to identify the taxonomic family for all the other samples. Thanks to this information, we were able to pick the nine samples per site: we avoided insects of the same family as much as possible. Otherwise, we checked that the identity of their COI was under 90%. Having a problem with four samples among the seven sites, our dataset is composed of 59 insects (7*9-4=59).

Sample	Extracted [#]	Depth (X)	Assembly size (Mb)	Closest CO1 [pID, cover]	Identification
WL5	whole body [1]	7 | 98	526	*Glaucopsyche lygdamus* [98.70]	family
WL3	whole body [1]	11 | 101	397	*Erynnis tristis* [93.55]	family
WL2	head & thorax [1]	4 | 96	435	*Argynnis coronis* [99.79]	genus
WH4	whole body [1]	8 | 47	206	*Ammophila sabulosa* [91.04]	family
WH2	whole body [1]	8 | 38	229	*Pimpla luctuosa* [83.68]	family
WH1	whole body [1]	16 | 37	312	*Agenioideus nigerrimus* [81.37, 92]	family
WD3	whole body [1]	5 | 35	940	*Drosophila jambulina* [87.65][a]	family
WD2	whole body [1]	5 | 52	456	*Syrphus vitripennis* [99.74]	genus
WD1	whole body [1]	18 | 101	243	*Bombylius major* [85.75]	family
SL15	whole body [1]	9 | 125	447	*Perizoma didymatum* [98.29]	family
SL10	whole body [1]	7 | 141	429	*Pennisetia hylaeiformis* [99.94]	family
SH4	whole body [1]	15 | 64	198	*Allorhynchium sp.* [85.71]	family
SH1	head & thorax [1]	14 | 44	268	*Tenthredo mesomela* [92.59]	family
SD8	whole body [1]	9 | 61	570	*Tipula fascipennis* [89.64]	family
SD5	head & thorax [1]	10 | 83	239	*Machimus atricapillus* [94.48]	family
SD14	whole body [1]	5 | 51	693	*Hydrotaea armipes* [89.84]	family
GL3	whole body [1]	3 | 59	1053	*Callimorpha dominula* [91.03][b]	family
GL2	thorax & abdomen [1]	9 | 59	544	*Anartia amathea* [99.10]	family
GL1	whole body [5]	5 | 26	481	*Eudonia lacustrata* [89.59]	family
GH3	thorax & abdomen [1]	23 | 115	248	*Paraponera clavata* [96.41]	family
GH2	whole body [1]	5 | 58	370	*Melipona fasciculata* [91.03]	family
GH1	whole body [2]	11 | 126	264	*Angiopolybia obidensis* [96.9, 86]	family
GD3	whole body [1]	5 | 78	442	*Ptecticus aurifer* [87.43]	family
GD2	whole body [1]	4 | 45	1442	*Mesembrinella sp.* [89.77][b]	superfamily
GD1	whole body [15]	5 | 59	270	*Drosophila melanogaster* [90.18][a]	family
FL8	whole body [1]	3 | 83	446	*Polyommatus icarus* [100]	genus
FL6	whole body [1]	11 | 87	701	*Agriphila geniculea* [100]	genus
FL2	thorax [1]	10 | 58	849	*Xestia xanthographa* [99.81]	family
FH16	whole body [6]	NA | 45	508	*Diplolepis rosae* [99.72]	species
FH13	whole body [1]	13 | 35	368	*Lasioglossum lativentre* [88.77]	genus
FH11	whole body [1]	18 | 48	188	NA	species
FD5	whole body [1]	8 | 43	682	*Beris chalybata* [87.88]	family
FD27	head & thorax [1]	6 | 30	1180	*Pollenia labialis* [89.98][b]	superfamily
FD26	whole body [2]	8 | 35	843	*Tipula paludosa* [99.93]	genus
EL8	whole body [1]	12 | 118	433	*Caradrina kadenii* [94.63]	family
EL3	head & thorax [1]	17 | 114	509	*Pelopidas mathias* [91.67]	family
EL2	whole body [1]	5 | 58	745	*Phragmatobia fuliginosa* [89.48][b]	family
EH2	whole body [2]	12 | 29	304	*Messor barbarus* [99.74]	family
EH19	whole body [4]	41 | 58	254	*Mimumesa dahlbomi* [88.61, 98]	family
EH1	head & thorax [1]	30 | 65	292	*Antodynerus aff. limbatus* [86.85, 96]	family
ED8	whole body [1]	6 | 70	621	*Stomorhina lunata* [100]	genus
ED4	whole body [1]	9 | 56	821	*Sarcophaga villeneuveana* [98.04]	family
ED20	whole body [1]	10 | 66	502	*Episyrphus balteatus* [99.96][c]	family
CRL6	whole body [1]	7 | 78	558	*Pholisora catullus* [92.23, 98]	genus
CRL4	whole body [1]	9 | 147	268	*Eurema blanda* [88.46]	family
CRL3	whole body [1]	2 | 67	336	*Anartia jatrophae* [99.74]	genus
CRH7	whole body [1]	5 | 31	285	*Melipona bicolor* [92.50]	family
CRH5	whole body [1]	8 | 64	153	*Antodynerus aff. limbatus* [84.99, 98]	family
CRD4	whole body [1]	13 | 42	582	*Oxysarcodexia varia* [97.40]	family
CRD3	whole body [1]	8 | 72	228	*Chrysops niger* [92.37]	family
CRD1	whole body [1]	2 | 44	523	*Volucella latifasciata* [90.45]	family
CL3	whole body [1]	6 | 61	718	*Pycnarmon pantherata* [91.95]	family
CL2	head & thorax [1]	2 | 39	614	*Amerila alberti* [95.31]	family
CL1	head & thorax [1]	7 | 47	801	*Thyas honesta* [93.23]	family
CH2	whole body [3]	14 | 133	233	*Exoneura angophorae* [92.37]	family
CH1	whole body [1]	17 | 80	312	*Anochetus minans* [85.56, 97]	family
CD2	whole body [1]	1 | 43	680	*Atylotus miser* [88.07]	family
CD1	whole body [1]	11 | 45	473	*Carpomya vesuviana* [86.48][b]	superfamily

Table 14.1: **Metadata of the 59 *de novo* assemblies**. Extracted: part of the body used for extraction [number of individual(s) in the extraction]. Depth: MinION sequencing | Illumina sequencing based on the expected genome size. Closest CO1: species name of the best hit of the CO1 annotated by mitoZ [percentage of identity, and percentage covered when <99%]. Identification: taxonomical level at which we identified our sample for NCBI submission.

a. sample identified as another family despite this hit (see text). b. the next hit has a similar score but on a different family. c. hit on the mtDNA because the CO1 was not annotated.

Sample	Species	Accession	Genome size	N50	BUSCO
BL1	*Furcula furcula*	GCA_911728495.1	736Mb	27Mb	99.4%
BL2	*Tinea trinotella*	GCA_905220615.1	372Mb	14Mb	98.8%
BL3	*Apotomis betuletana*	GCA_932273695.1	684Mb	25Mb	99.0%
BH1	*Andrena dorsata*	GCA_929108735.1	273Mb	89Mb	99.2%
BH2	*Ectemnius lituratus*	GCA_910593735.2	235Mb	17Mb	98.9%
BH3	*Athalia rosae*	GCA_917208135.1	172Mb	25Mb	99.6%
BD1	*Sicus ferrugineus*	GCA_922984085.1	312Mb	45Mb	97.5%
BD2	*Pollenia amentaria*	GCA_943735925.1	1270Mb	235Mb	99.4%
BD3	*Tachina fera*	GCA_905220375.1	752Mb	142Mb	99.4%

Table 14.2: **Dataset from Darwin Tree of Life.** BUSCO shows the percentage of complete BUSCO (single and duplicated) using the insecta_odb10 database.

We were able to add a 8th site in Britain thanks to the project Darwin Tree of Life, the aim of which is to sequence the genomes of 70,000 eukaryotes species in Britain and Ireland. The sampling map on their website allowed us to pick nine insect genomes (named B after Britain) that were all sampled at the exact same locality (51.7719N,-1.3378W) (figure 14.1 and table 14.2). This 8th site brought the total number of samples to 68.

14.2 . Genome sequencing and assembly

Working on HT of TE, we were concerned by the assembly of these elements. TE being in numerous copies, the assembly of these elements can be challenging. It was shown that long read sequencing is more efficient than short reads sequencing to assemble these elements (Shahid and Slotkin 2020). This is why we chose to generate long reads with MinION. However, MinION sequencing has a high error rate (∼15%) and it demands high amounts of DNA (ideally 3μg at the beginning of the protocol). Yet, the amount of DNA we can extract is limited in such small insects. Because of these two limitations, we chose to not use only MinION sequencing, instead we used a hybrid strategy to assemble our 59 genomes, generating also short reads with Illumina sequencing. This technology is less prone to sequencing errors (<1%) and it is easier and cheaper to obtain a deeper sequencing depth, even with a small amount of DNA (few hundreds of ng).

For Illumina sequencing, we sub-contracted Novogene to build a paired end library (2x150 pb; insert size = 350 bp). For MinION sequencing, we performed long-read sequencing in our laboratory. For this, we selected fragments longer than 10kb with the protocol SRE XS of Circulomics, except when samples were too fragmented. Then, we prepared libraries with the SQK-LSK109 kit and we use one flowcell for two to three samples.

We assembled the 59 genomes with the MaSuRCA hybrid assembler v4.0.1 to v4.0.5 depending on samples. We chose this software, firstly because it is able to handle hybrid assemblies, but also because its optimal is with a high depth of short reads (∼100X) and with a moderate depth of long reads (∼10X), which is cost-efficient and in line with what we are able to generate with our constraints (Zimin *et al.* 2013). Although we aimed for these sequencing depths, we obtained a median sequencing depth of 60X with Illumina (70X on average), and of 7X with MinION (8X on average) (more detail in table 14.1). We set all parameters to default, except for USE_LINKING_MATES that was set to 1 (as suggested when one has less than 15X coverage of long reads) and for Jellyfish hash size

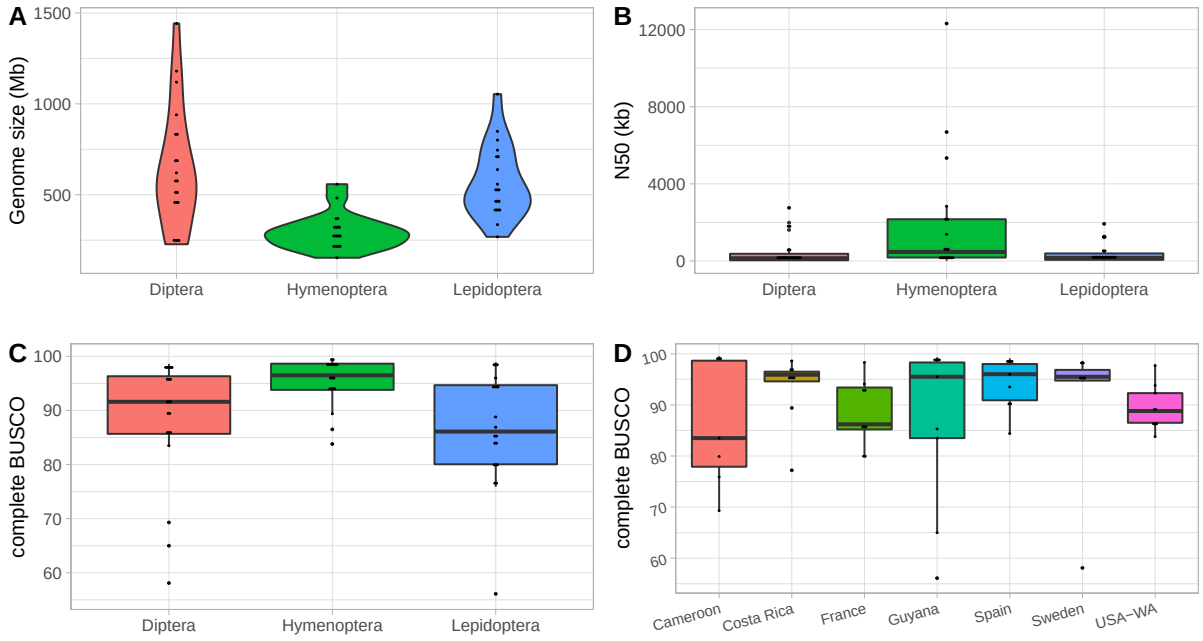


Figure 14.2: **Assembly statistics**.

(JF_SIZE = 10000000000 for Hymenoptera and 20000000000 for Lepidoptera and Diptera).

FH16 was processed by a collaborator, Antoine Branca, so it followed a different procedure. DNA from 6 whole bodies of adults that came from the same gale were extracted with the kit DNeasy of QIAGEN, and it was sequenced with Illumina sequencing only. The reads were assembled with Minia, followed by an additional step of scaffolding with Ragtag thanks to the reference genome of the same species (also assembled by Antoine Branca, with a hybrid strategy).

On average, the assemblies have a N50 of 895kb, and a median of 216kb. The minimum value is that of GD2 (4kb) and the maximum value is that of SH4 (12Mb). The completeness of the assemblies were assessed with BUSCO v5.1.2 with the database insecta_odb10. The samples have an average BUSCO score of 90.0%, and a median of 93.8%. The assembly sizes range from 153Mb (for CRH5) to 1442Mb (for GD2) (see details in table 14.1). Hymenoptera have the smallest assemblies, and the less variation (from 153 to 558Mb). On the opposite, Diptera have the biggest assemblies, but some also have a smaller assembly size; there is a high variation (from 278 to 1442Mb). These smaller assembly sizes for Hymenoptera reflect their higher quality: a median N50 of 470kb *versus* 146kb and 161kb in Diptera and Lepidoptera, and a median complete BUSCO score of 96.5% *versus* 91.1 and 83.1% in Diptera and Lepidoptera, respectively (figure 14.2).

When comparing genome quality per site, Cameroon, France and USA-Washington have the lowest median BUSCO scores, whereas Costa-Rica and Sweden have the highest ones (figure 14.2). Genomes with the lowest scores (<80%) are from Guyana (GD3: 65.0%; GL3: 56.1%), Sweden (SD14: 58.1%), Cameroon (CL3: 79.9%; CL2: 75.9%; CD2: 69.3%), Costa-Rica (CRL6: 77.2%), and France (FL6: 79.8%). Because of this difference of quality among sites, it will be important to check for the absence of correlation between genome quality and the number of HTT we will recover. If Cameroon has the lowest number of HTT, it might be due to its general lower quality of genomes, instead of the absence of other closely located site.

In addition, we assembled the mitochondrial DNA with the module all of mitoZ v3.4 (Meng

et al. 2019), with the parameters –clade Arthropoda –requiring_taxa Arthropoda –genetic_code 2 –fastq_read_length 150 –data_size_for_mt_assembly 0 (to use all the Ilumina reads of the file), –assembler megahit –kmer_megahit 59 79 99 119 141 (as advised in the wiki of mitoZ). 19 of the 59 samples have a complete and circular mitochondrial assembly, and 29 have an almost complete mitochondrial assembly (*i.e.*, assembly in one contig, which is >10kb, but not circular). The 11 remaining samples have a fragmented assembly (from 3 to 9 contigs).

14.3 . Cleaning assemblies

Working on horizontal transfers, we were concerned by contamination, so we made a database that includes common sources of contamination in laboratories: the human genome hg38, *E. coli* (we concatenated the 347 genomes of chromosomal level available in November 2022), and UniVec (a non-redundant database of a large number of vectors). We looked for similarities between our 59 *de novo* assemblies and this database with the module search of MMseq2 (options -s 5.7 –search-type 4 and –max-seqs 50), and we focused on hits with a percentage of identity >90%. While we did not obtain a single hit between our samples and UniVec, we obtained some hits on the human genome and on *E. coli* (figure 14.3). Most hits with the human genome were very short (<250bp), which might reflect some very conserved regions between insects and vertebrates. Yet, some hits were longer (up to 1256bp) with 100% of identity, which might reflect some human contamination, although in low amount. Regarding hits on *E. coli*, their percentage of identity were lower, which might reflect contamination by another bacteria, possibly a parasite. Overall, the result of this similarity search was very reassuring, and it shows that if some samples were contaminated by humans or *E. coli*, it is in very low amount. Based on the assumption that contamination at a low rate would assemble badly, one could expect such contamination to be assembled in small scaffolds. This is why we chose to remove from our assemblies all scaffolds covered at least at 25% of their lengths by a hit with more than 90% identity on a sequence of our database of contaminants (triangles in figure 14.3). This led us to remove 111 scaffolds across 26 samples of our 59 genomes (from 1 to 23 scaffolds per genome). The resulting cleaned assemblies of the 59 genomes will be made available on NCBI. Regarding the genomes from Britain downloaded on Darwin Tree of Life, they were of very high quality (table 14.2), so we did not go through this step of decontamination.

In the context of this study, in which we will investigate HT between each of the 68 samples, we would like to be confident about the absence of contamination (that could have been introduced during experiments, or that could correspond to parasites), which would lead to false positive. Since we sequenced the whole bodies of the insects most of the time, we expect most of our *de novo* samples to be contaminated by various symbionts. To remove them from our analysis, we ran Blobtools and we kept only the scaffolds covered at least at 8X by short reads and identified as Arthropods or viruses. We chose a minimum sequencing depth because scaffolds covered by just few reads are more prone to sequencing errors, which would impede our ability to detect direct contamination. We chose to keep scaffolds identified as viruses because of the numerous endogenized viruses found in most genomes and because of the fact that some LTR transposable elements are sometimes annotated as viruses. Many of the scaffolds we removed from the assemblies for this study are probably truly part of the genomes of interest, however we chose to be conservative by focusing only on the retained scaffolds. Although the assembly size decreased quite a lot in some of our samples, the number of complete

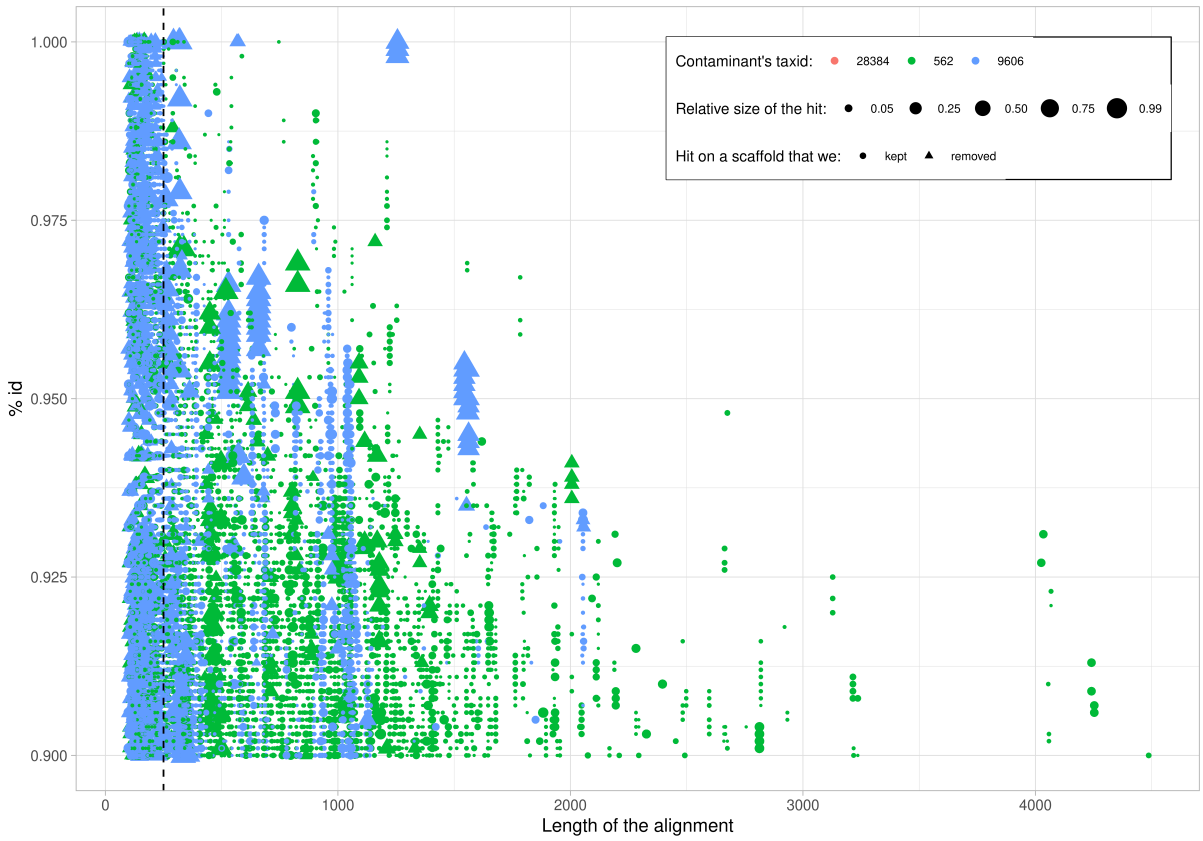


Figure 14.3: **Hits between the 59 *de novo* assemblies and possible contaminants**. Colors show the taxid of the contaminant (26384 for UniVec, 562 for *E. coli*, and 9606 for *H. sapiens*). The relative size of each hit (length of the alignment divided by the length of the scaffold) is proportional to the size of the dot. Circles show hits for which the relative size is $\leq$0.25%, while triangles show hits for which the relative size is >0.25%. ggplot draw circles first, this is why it looks like there are more triangles, yet they represent only 0.47% of the total hits. The dash line is at 250nt, illustrating the length of the majority of the hits.

BUSCO genes stayed similar, at the exception of three samples for which we lost more than 20% of complete BUSCO genes (figure 14.4). This loss was due to the fragmented assemblies (Blobtools could not assign the short scaffolds to any taxonomy), rather than a high contamination rate. For all the rest of the study, we worked on these subsets of genomes. Regarding samples from Britain, we also ran Blobtools, which showed that most of the scaffolds were assigned to Arthropoda.

14.4 . Check for contamination

Here we make the distinction between three kinds of contamination: (i) direct contamination, (ii) indirect cross-contamination, and (iii) non shared contamination.

Direct contamination is when the DNA of species A directly contaminates sample B. Such contamination can take place while handling the insects, during DNA extractions or during DNA sequencing. Since all extractions of the *de novo* assemblies were achieved at the same laboratory, such contamination could have taken place. Direct contamination is actually the easiest to identify: the contaminant should be 100% identical between samples A and B. Sequencing errors could decrease this percentage though, this is why we kept only scaffolds covered at least at 8X after running Blobtools. To de-

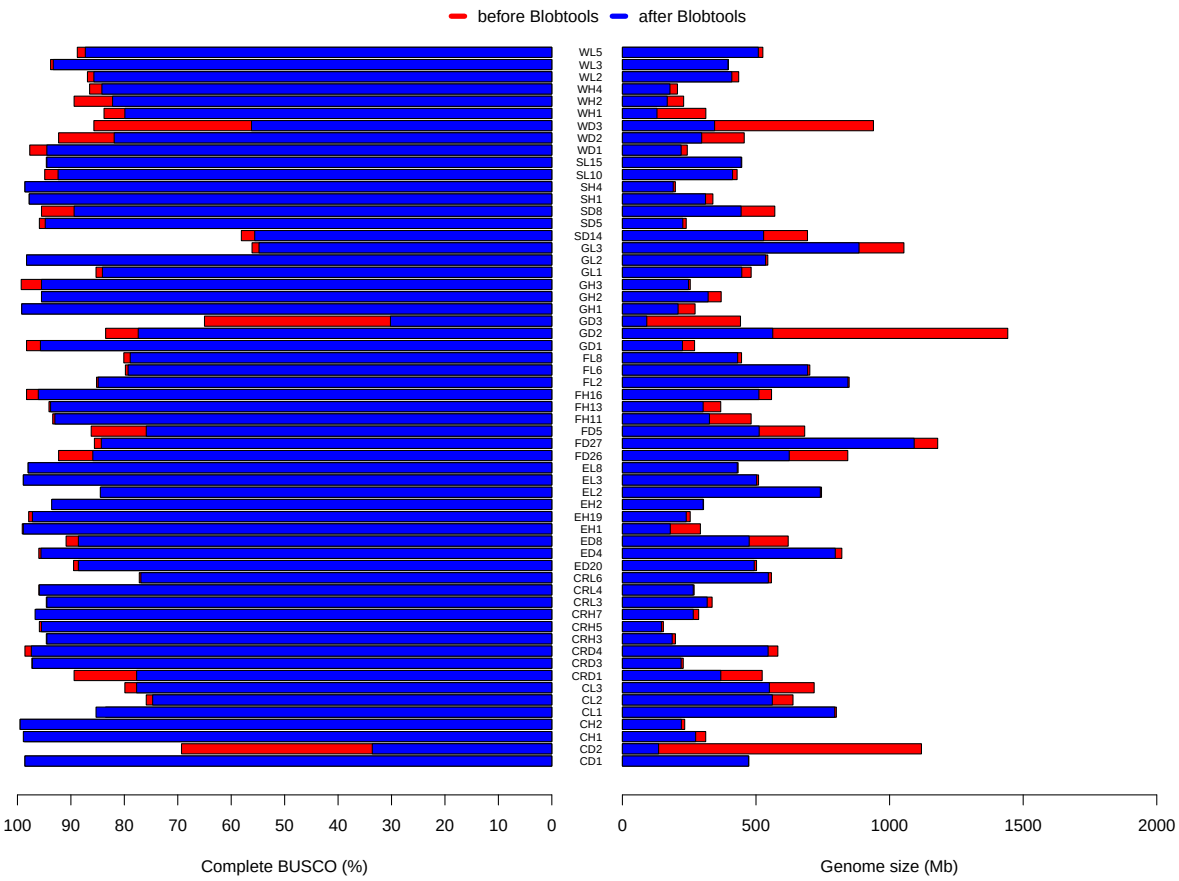


Figure 14.4: **Statistics of the 59 *de novo* assemblies**. (left) Complete BUSCO genes. (right) Genomes size in megabases. On both sides, red indicates statistics of the version we are going to submit on NCBI, whereas blue indicates statistics after running Blobtools, which are those of the versions of the assemblies we are using in this study.

tect direct contamination, we could have compared the sequence identity of BUSCO genes between samples. Yet, these genes are in single copies, whereas TE are in numerous copies. So if the contamination is low, one can expect to assemble contaminant TE, but not contaminant BUSCO genes. Because of this, we chose to detect contamination thanks to mitochondrial DNA, which is also in numerous copies. For this, we did a similarity search between all the *de novo* assemblies and their mitogenomes with megablast (we kept only hits longer than 400bp with 95% of identity or more). The output revealed four hits, but with only 95-96% of identity, and between pair of samples which are taxonomically related. This method allowed us to conclude on the absence of direct contamination between our 59 assemblies.

Indirect cross-contamination is when two samples (or more) are contaminated by the same contaminant. This contaminant can originate from the laboratory, but it can also be a biological shared parasite. In any case, the presence of such a contaminant would lead to false positive HT. Since we kept only scaffolds identified as Arthropoda (or viruses) by Blobtools, such contaminants should be quite limited in the subset of genomes we are using in this study.

Non shared contamination is when a contaminant (from the laboratory or a parasite) contaminated only one of our samples. Although such a contaminant is the most difficult to identify, it should not impede our study, except if a HT took place between the contaminant species and one species of our dataset.

Once HT will be identified, it will also be possible to check some of them *a posteriori*. For this, we could look at the flanking regions of the TE and make sure it is different in the pair of samples involved in the HT (figure 5.1D). In the case of the 59 genomes we assembled ourselves, we also have the possibility to check some insertions by PCR (figure 5.1D). We could also investigate HT whose TE-TE hits is 100% identical. Such a high percentage would reflect either a very recent HT, or a contamination. Yet, TE from contamination will not necessarily harbor a percentage of identity of 100%, which can be explain in two scenarios. The first one is when two related species each contaminated a sample of our dataset. So the percentage of identify of the false HT will be the percentage of identity between the two contaminant species. The second scenario is due to the fact that TE are in numerous copies, which are not totally identical. So if we do not sequence the contaminant copy, the percentage of identity will be <100%. Because detecting contaminant is not that obvious, we will be very careful with false HT. A noteworthy precision is that we do not expect some false HT due to contamination in some species of the dataset to bring an overall signal on the two factors we are testing here.

14.5 . Taxonomy and phylogeny

In the case of the samples from Britain, their species names were specified on Darwin Tree of Life (Table 14.2). For the samples we *do novo* assembled though, only two samples were identified at the species level by experts working on these species on a regular basis at the laboratory: *Colletes hederae* (FH11) by Fabrice Requier and *Diplolepis rosae* (FH16) by Antoine Branca. All the other samples were assigned a taxonomy based on the similarity between their CO1 gene and the nr/nt database. We already had the sequences of these genes for most samples, which we obtained though PCR amplification, yet we used the CO1 sequences annotated by mitoZ which were more complete (table 14.1). Except in the case of a hit with 100% identity (using the CO1 annotated by mitoZ or the CO1 amplified by PCR), in which case we assigned the sample at the genus level, we assigned the taxonomies at the family level to avoid miss classification. The taxonomy of WD3 was unclear: the CO1 amplified with PCR was related to *Cyrtopogon montanus* [98.11%] (Asilidae), whereas the CO1 annotated by mitoZ was related to *Drosophila jambulina* [87.65%] (Drosophilidae), hits on Asilidae species having a lower percentage of identity [85.76% on the Asilidae *Dasypogon diadema*]. Looking at the morphology, we could confidently tell WD3 is not a Drosophilidae, despite the low quality of the picture that was taken once the individual was already in alcohol (figure 14.5). Similarly, the best hit of the CO1 of GD1 annotated by mitoZ is on *Drosophila melanogaster* [90.18%]. However, looking at the morphology, we could confidently tell it is not a Drosophilidae (figure 14.5). The CO1 of GD1 was poorly anotated (708bp) and its mtDNA poorly assembled (3 contigs of 5091, 3840, and 2021bp). When using each contig to perform a similarity search on the nr/nt datatbase, instead of just the CO1 gene, the best hits of the two first contigs are on *Rhamphomyia insignis* [85.75%] and *Empis stercorea* [86.04%], reciprocaly, two Empididae species. The best hit of the last contig, which contains the CO1 gene, is on *Drosophila formosana* [90.84%] but it covers only 89% of the query. Taking all this information together, we decided to assign the family Empididae to GD1. The taxonomical family was unclear for seven other species, but we solved four of them thanks to the phylogenetic analysis (see bellow). For the remaining three samples, we could only identify them at the superfamily level.

Figure 14.5: **Pictures of some samples of interest**. From left to right: FL8, EL3, ED20, EH1, WD3, GD1, CRH3, and EH19. The first row is here simply to show the cutest (or scariest on the right) insects I captured for this project.

In order to build a phylogenetic tree, we extracted 300 BUSCO genes in single copies, choosing the ones found in the majority of the 68 assemblies. We aligned these genes with MAFFT v7.490 (–auto), and we trimmed the resulting alignments with trimAL v1.4 (-strictplus). We then built a tree with iqtree2 (-m MFP -B 1000) without any constrains (figure 14.6). We found two nodes with weak bootstrap supports (50 and 56) which are not in accordance with what is known in the literature: Cynipidae should group with Ichneumonidae (Peters *et al.* 2017) and Muscidae should be outside the Oestroidea (composed here of the families Polleniidae, Tachinidae, Rhiniidae, and Sarcophagidae) (Narayanan Kutty *et al.* 2019). Within Oestroidea, our phylogeny is poorly resolved, as it is the case in literature (Narayanan Kutty *et al.* 2019). We ran iqtree2 a second time, but this time with two constrains to fix the topology (option -g), using 23 dates found on TimeTree to date the tree (options –dates and –date-tip 0). The resulting tree is shown in figure 14.7. To make sure that the topology of this new tree is not significantly different than the unconstrained tree, we ran the AU test, as suggested in the manual of iqtree2. The AU test did not rejected our dated phylogenetic tree.

The unconstrained tree allowed us to solve the family of four samples: GL3 and EL2 are Erebidae (the taxonomical analysis suggested either Erebidae or Noctuidae), CRH5 is a Vespidae, and CRH3 is a Formicidae despite its unusual morphology for an ant (see its large thorax in figure 14.5). GL3 and EL2 both group with CL1 and CL2, which are Erebidae (figure 14.6). The PCR amplification on the CO1 gene of CRH5 did not work, its morphology was ambiguous, yet both our taxonomical and our phylogenetic approaches detected this sample as a Vespidae (figure 14.6). The mtDNA of CRH3 is badly assembled and its CO1 gene is incomplete (∼900nt). While this partial CO1 gene blasted both on Formicidae and Crabronidae species, the CO1 we amplified by PCR was related to *Eciton hamatum* (99.3% of identity), a Formicidae of the subfamily Dorylinae. Looking into more details at the morphology of Dorylinae, we found that some species of this subfamily also have quite a large thorax (Brady *et al.* 2014). Finally, CRH3 grouped with the other Formicidae in the phylogenetic tree (figure 14.6), so we decided to assign the family of Formicidae to CRH3.

Another dubious case is that of EH19. The best hit of EH19's CO1 gene was on a *Mimumesa* (a Crabronidae), with a percentage of identity of 88.6%. However EH19 does not group in the phylogenetic tree with BH2 which is also supposed to be a Crabronidae (figure 14.6). Yet, the morphology of EH19 looked quite similar to the one of *Mimumesa* (see its picture in figure 14.5), so we decided to keep EH19 as a Crabronidae.

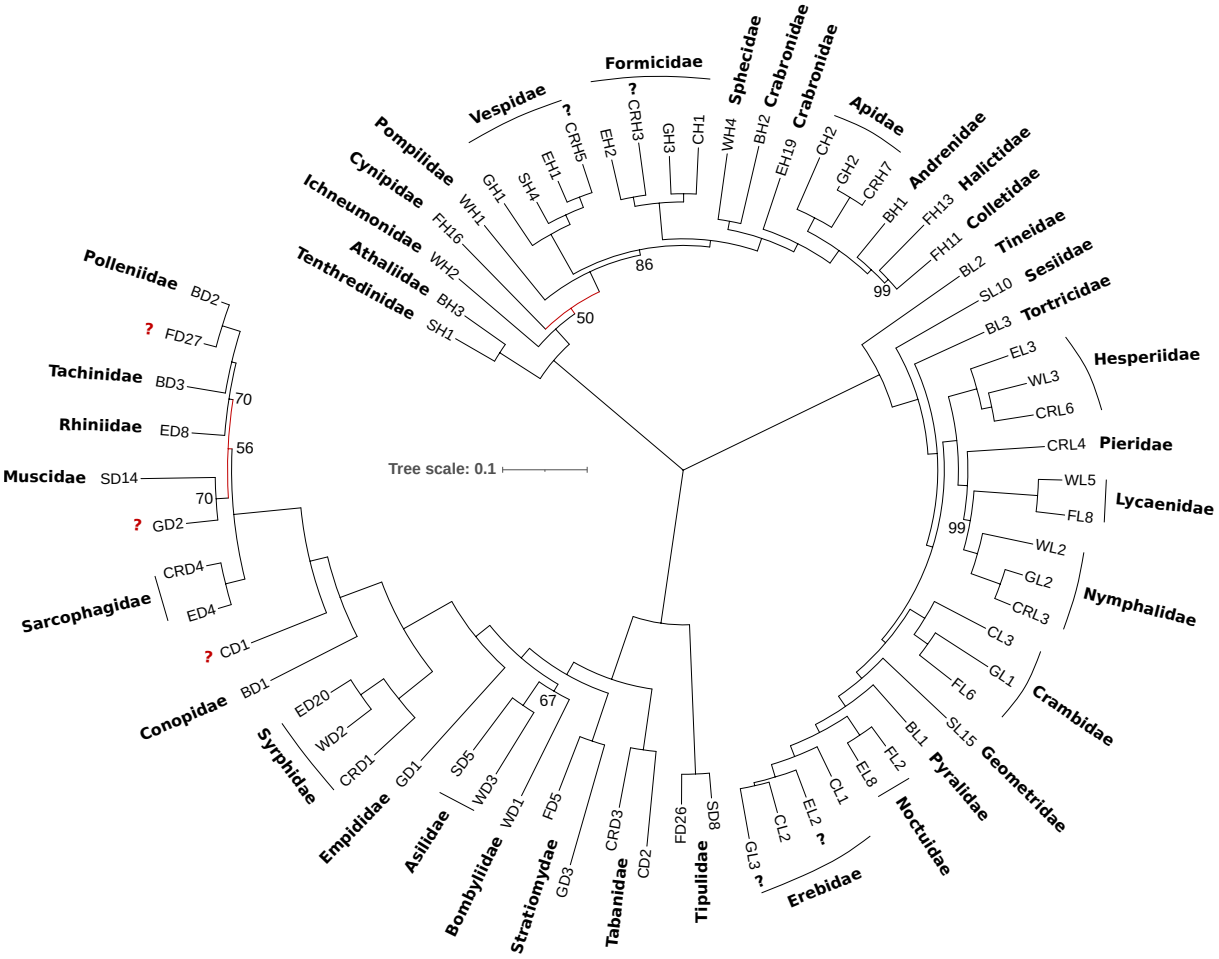


Figure 14.6: **Phylogenetic tree of the 68 samples**. Bootstraps under 100 are indicated at the corresponding nodes. Nodes with a Bootstrap under 60 are colored in red. The family of each sample identified with the taxonomical approach is indicated in bold (interrogation points when we could not identify the family). The family of black interrogation points were resolved thanks to the phylogenetic tree, but not the red ones. Tree visualized on itol.embl.de, and annotated by hand.

Doubts remained on the family of CD1, FD27, and GD2. CD1 being the only Acalyptratae of our dataset, the phylogenetic tree could not help to solve its family. FD27 is either a Polleniidae or a Calliphoridae according to our taxonomical approach, but we only have one species of Polleniidae (BD2) in our dataset and no Calliphoridae. GD2 is either a Muscidae or a Calliphoridae according to our taxonomical approach, but we only have one species of Muscidae (SD14) in our dataset and still no Calliphoridae. To note, Calliphoridae is not a monophyletic family (Narayanan Kutty *et al.* 2019), so we cannot exclude that both FD27 and GD2 belong to this family even though they do not group together in the tree. Thus, we were only able to identify these three samples at the superfamily level: Acalyptratae for CD1, and Calyptratae for FD27 and GD2.

Looking at the dating of the phylogenetic tree, nine pairs of species diverged less than 40 Myrs ago, so we will not be able to look for HT between these pairs of species (figure 14.7).

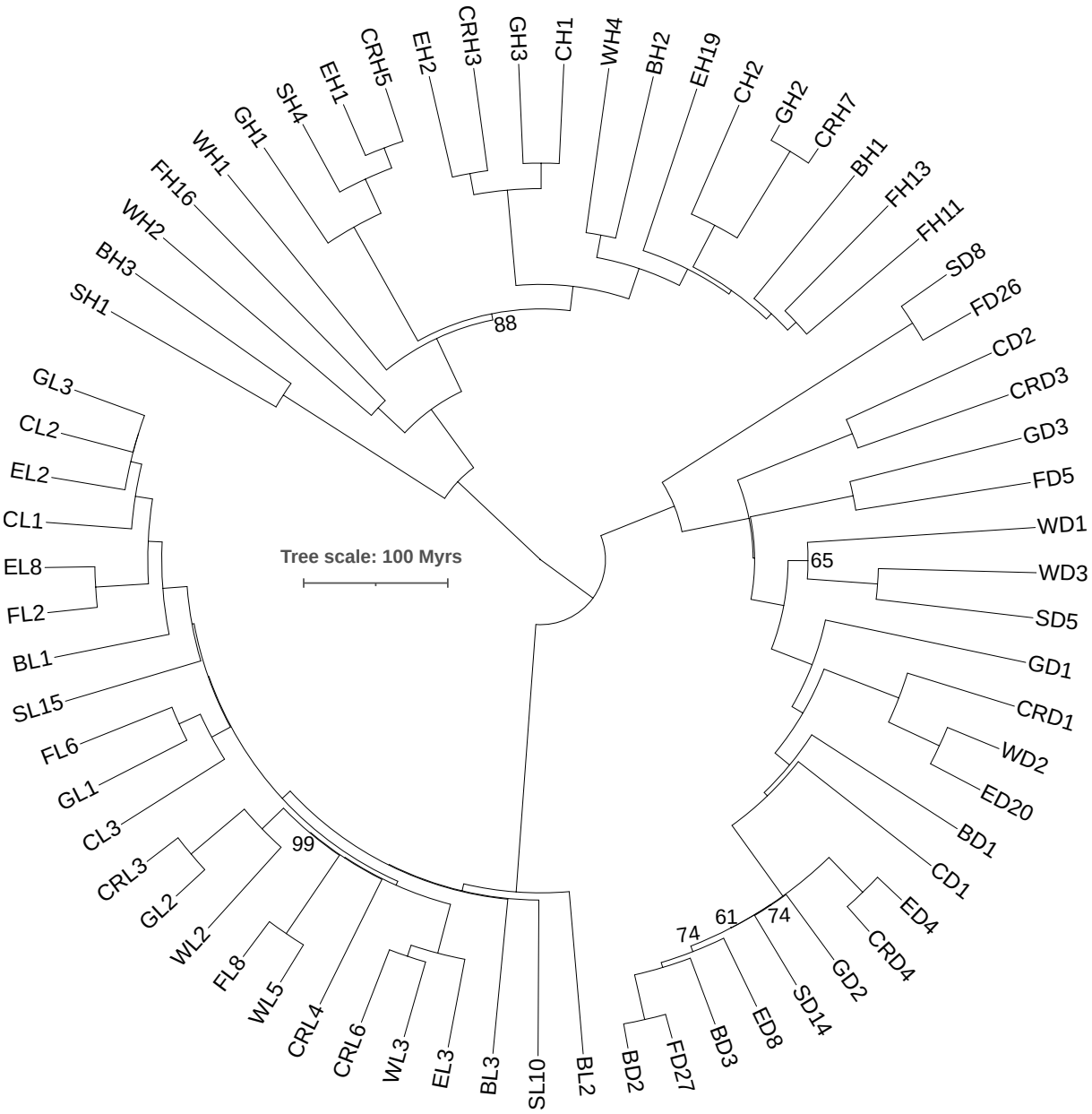


Figure 14.7: **Dated phylogenetic tree of the 68 samples**. Bootstraps under 100 are indicated at the corresponding nodes. Tree visualized on itol.embl.de

14.6 . Rest of the pipeline

For the rest of the pipeline, from TE annotation to the identification of TE resulting of HT, we will use the same approach as for the previous large-scale study (figure 12.1). At the time of the writing of this manuscript, we already ran RepeatModeler on 64 of the 68 assemblies. Since the rest of the pipeline is now ready to use, and since the genomes of this study are quite small (compared to many genomes of the previous study), each step of the pipeline will be much faster to run. This is why we expect to obtain the remaining results in just few months, instead of a year with the previous study.

14.7 . Perspectives

Since we designed the dataset in order to test the phylogenetic and the geographical distances, we could use this dataset to test these two factors on other events than HTT. For example, we could test these factors on HT of other sources, such as genes, including bacterial and viral genes. And we could also quantify HT from these different sources and compare them.

Regarding the 59 *de novo* assemblies, the potential of this dataset is much higher. Since we performed DNA extractions on the whole body for most samples, we have access to their respective parasites (bacteria, viruses, but also eukaryotes) that lived in or on those organisms. These parasites can be extracellular or intracellular, and it would be interesting to assess whether they are involved in HT. They could be involved at two scales: HT between the host and the parasite, and they could also act as vector to transfer this newly acquired sequence to another host. To investigate the first inquiry, we could look for HT between the parasites we recovered in our samples and the genomes of these samples. For the second inquiry, we could look whether two samples that share a common parasite also shared more HT events than pairs of samples that do not share a common parasite.

General discussion

15 - Free viruses as vectors of horizontal transfers

In the first part, I shortly investigated whether free viruses could act as vectors of HT between animals, *i.e.* transport DNA between two hosts. This DNA could either stay free in the virus, or it could be integrated in its genome (first step), before being inserted in the genome of another host during a secondary infection (second step).

In the published study, we showed that a viral infection (here AcMNPV) can impact the TE expression of the host, although moderately. An increase in TE activity could favor the chance of a TE to transpose in the virus, fulfilling the first step of a HT between animals mediated by a free virus. In addition, we found a host-derived TE integrated in the viral genomes, which was highly expressed in the host cells. This means that once inserted in the virus, TE might still be active, and could thus transpose in the secondary host, fulfilling the second step of a HT between animals mediated by a free virus.

During my PhD, I was supposed to investigate the role of free viruses as vectors of HT into more details, yet we chose to shorten this part, in order to add the large-scale study on the 247 animal genomes of part III. The initial plan was to try to recover a full HT between two insects mediated by AcMNPV, using the AcMPNV from the study of Gilbert *et al.* (2016), in which they infected *Spodoptera exigua*. Following this infection, they estimated that about 5% of the viral genomes harbored at least one TE that transferred from *Spodoptera exigua*. I was thus supposed to use these AcMNPV to do a second round of infection on larvae of *Sesamia nonagrioides*, cross the resulting adults, and see whether I could find any of the TE recovered in the first study in the offspring. *S. nonagrioides* is very permissive to AcMNPV, and was used in previous studies for infection by this virus (Loiseau *et al.* 2021).

It would also be very interesting to do the same experiment with other viruses, and also to repeat it using multiple viruses to infect the same host in order to assess whether TE can transpose from one virus to the other. AcMNPV is actually a very good candidate as a vector because of its broad range of hosts and of its double stranded DNA; it was shown that most HVT are from this type of viruses (Irwin *et al.* 2022). Any other viruses with the same features would be good candidates and are worth further investigation.

16 - Domesticated viruses as vectors of horizontal transfers from parasitoid wasps

In the second part, I focused on a particular kind of viruses: the polydnaviruses (PDV), which are domesticated viruses of some parasitoid wasps. After characterizing the integration of CtBV (the polydnavirus of *C. typhae*) in its natural host *S. nonagrioides* (Article n°3), I investigated the chances that such integrations are transmitted over generations (Article n°4). In this sense, I found numerous integrations in three lepidopteran species not naturally targeted by *C. typhae* (Article n°4), which means that these species might have more chances to survive parasitism, and thus to transmit their integrations. Yet, current work in the laboratory suggests that this is true only for some non target species, whereas others die following parasitism. In the case of our three non target species, parasitism significantly increased the mortality of *N. typhae* and *C. phragmitella*, but not the one of *G. sparganii*. Regarding transmission of integration to the next generation, I could not detect any transmission among the 500 offspring of surviving *S. nonagrioides*, despite the persistence of these integrations in their parents at the adult stage (Article n°4). This result suggests that if transmission is possible, it happens at a frequency too low to be detectable by my experiment, at least in the system CtBV/*S. nonagrioides*. At a broader scale, we found that such transmissions did happen several times during lepidopteran evolution (Article n°5). We found traces of HIM-mediated integration of polydnaviruses in 124 lepidopteean species out of the 775 we investigated, with 23 having both IV and BV integrations.

All these results support PDV as a major vector of HT from parasitoid wasps to lepidopterans. A study is currently in process at the laboratory to assess whether DNA was also transferred over the course of evolution from wasps to other non-lepidopteran hosts. For this, the same method as the one used among the 775 lepidoptera genomes is being used (Article n°5), but extended to all arthropod genomes available on NCBI.

Although we showed that HT from PDV took place several times during lepidopteran evolution (Article n°5), we could not calculate the frequency at which such transmissions take place, because it appears to be lower than what was detectable by our pipeline (Article n°4). In order to bring more light on this, Inès Matrougui, doing her M2 internship during my last year of PhD, is estimating the number of integration events of CtBV in the gonads and eggs of surviving *S. nonagrioides*. Her study will help understand to what extent gonads are targeted by CtBV, and should help to understand why no CtBV integrations were transmitted to any of the 500 offspring I investigated (Article n°4). An improvement of the pipeline could also help to increase our power of detection: taking into account the information of the paired-end reads (as in figure 5.1D), in addition of chimeric reads, might allow us to detect integration despite the absence of chimeric reads.

All this work focused on PDV as a source of HT, yet it would also be interesting to investigate whether PDV could also act as vector of HT of other sources, by transporting non-PDV DNA in its viral particles. In this context, it is interesting to notice that helitron TE were proposed to be able to hitchike on PDV particles to hop on among multiple insect species (Heringer and Kuhn 2022). In addition, we found a HT of a helitron element between a parasitoid wasp and *S. nonagrioides*, although we could not tell whether this transfer was facilitated by PDV (Article n°2).

In addition to enlightening a mechanism of HT between insects, this work also brings new inputs on PDV in general, and rises new questions such as (i) "What is the evolutionary scenario that led to similar features between IV and BV?", (ii) "How did proviral segments evolve among related wasp genomes?", (iii) "To what extent the patterns of PDV integrations are similar between related wasp species?", (iv) "By what mechanism(s) the DNA circles are integrated?", (v) "What is the role of PDV integration during parasitism?", and (vi) "What makes a host permissive to the wasp development?". Some of these questions are currently being investigated at the laboratory. In order to answer to the questions (ii) and (iii), we need to annotate PDV in more wasp species, and to characterize their integrations in different hosts. In this sense, we are currently working at the laboratory on *Cotesia icipe*, a species which is quite distant to the other *Cotesia* investigated until now. This wasp lays only one egg per host, so it will be interesting to see to what extent its PDV involve similar mechanisms. For question (iv), the protein recognizing the HIM in order to integrate the DNA circle in the host genome was not clearly identified yet, nor the mechanism by which non HIM-mediated integrations take place. Two integrases encoded by genes of nudiviral origin might be involved (Chevignon *et al.* 2018), yet it was also proposed that host factors might be involved (Wang *et al.* 2021a). This latter scenario is supported by our study in which we found different patterns of integration depending on the host species (Article n°4). It seems very counter intuitive that a host factor would be involved, because we would expect the host to quickly get rid off such an element since parasitized hosts usually die. The only possibility for such a host factor to be maintained is if this factor has another function which would be vital for the host, or if this element is selfish. I shortly investigated that second possibility, and it appears that the host integrase that Wang *et al.* (2021a) identified is a LTR/Gypsy transposable element. How this host TE is recruited and whether this mechanism of integration is specific to some species only still remains to be investigated. For question (v), it was initially thought that the goal of integration was to persist long enough in the host, whereas the free DNA circles would be degraded. However, we showed that free DNA circles persist at high levels during the entire duration of the parasitism, making the role of integration unclear (Article n°3 and n°4). What is clear though, is that there is a strong selective pressure to maintain such integrations, as shown by the conservation of the HIM motif between proviral segments but also between quite distant wasp species. This question is also quite related to question (vi), in which we can wonder whether integrations are as numerous in non permissive hosts. We are currently investigating this question at the laboratory, by comparing the number of integration events of two strains of CsBV (the BV of *Cotesia sesamiae*) in *Busseola fusca*: the inland strain develops well on *B. fusca*, while the coast strain is generally not able to develop on *B. fusca*.

17 - Impact of horizontal transfers on evolution

Genomes are shaped by natural selection and genetic drift, changing frequencies of alleles in populations, sometimes to fixation. Novelty can arise from spontaneous mutation, *de novo* gene genesis, gene duplication, gene loss, recombination, introgression, or horizontal transfer.

In prokaryotes, it is clear that HT has a major impact on their evolution, with 81% of their genes that would have been acquired horizontally at some point during their evolution (Dagan *et al.* 2008). The vast majority of expansions of protein families in prokaryotes would be due to HT (Treangen and Rocha 2011).

In eukaryotes, it was believed that HT were not possible, and although we now have some examples of such events, some researchers still doubt its veracity. Martin (2017) even goes as far as saying: "Could it be that eukaryote HGT does not really exist to any significant extent in nature, but is an artefact produced by genome analysis pipelines?". He claims that if HGT in eukaryotes was really an important phenomenon (he discusses only HGT since he accepts HTT), we should be able to find evidence of genetic mechanisms for HGT and we should observe a cumulative effect of HGT over time, which means that the number of genes acquired by HGT should increase as a function of time (as it is the case with pangenomes in prokaryotes or with sequence divergence). He does raise some interesting points like the fact that many HGT recovered were actually due to contamination (Lander *et al.* 2001; Bemm *et al.* 2016), and that one should not automatically conclude to a HGT in the presence of unexpected branches in a phylogeny, without looking at other possible explanations such as gene loss and gene duplication, and of course to random phylogenetic error. He also points out the fact that most HGT analyses are based only on genes in one genome, without evolutionary information about all genes in all genomes.

Nevertheless, I do not think that the impact of HT on evolution is only a matter of frequency. Even if the endogeneization of the retrovirus involved in the formation of the placenta was the only event of HGT in this taxa, could we really say that it had no impact on the evolution of this taxa? This is why I think that the answer to the question "What is the impact of HT in animal evolution?" is bipartite: we should look at the quantitative aspect, but also at the qualitative aspect, since just one transfer can have a major impact on the evolution of a clade. In the chapter 1, I give some examples of HGT from prokaryotes or viruses to multicellular organisms with striking impacts on the receiving species (such as adaptation to extreme environments, protection against bacteria, or vision in vertebrates), and in the section 2.3, I give some examples of HGT *between* multicellular organisms, also with striking impacts (such as the exchange of genes involved in resistance against toxins, or in pigmentation). To come back in the context of my PhD in which I focused on HT *between* metazoa only, it is true that there are not many examples of HGT with a described effect. To my knowledge there is only one example not related to polydnaviruses: the gene coding for an antifreeze protein that was transferred between two species of fishes (Graham and Davies 2021). In addition, they are three examples of genes with poydnavirus origins : (i) *SI gasmin* that transferred from the polydnavirus of a wasp to a lepidoptera, which now plays a role in anti-bacterial immune response in the latter (Gasmi *et al.* 2015; Lelio *et al.* 2019), (ii) *Se-BLL2* that confers a resistance of caterpillars to baculovirus infection (Gasmi *et al.* 2015), and (iii) *Se-BLL3* that reduces the mortality of *Spodoptera frugiperda*

larvae caused by the densovirus JcDV (Gasmi *et al.* 2018).

In addition to these three genes, many other sequences from polydnaviruses were recovered. Over the course of my PhD we found many events of HT from polydnaviruses to their lepidopteran hosts, enlightening the quantitative aspect of these HT (Article n°5). Such a high frequency suggests that HT from polydnavirus has a major impact on the evolution of Lepidoptera. Future qualitative studies, mostly looking at the co-opted sequences, should bring more light on the real impact of such HT on their evolution. In addition, my results on polydnaviruses also describe into details a mechanism for HT, which was an indispensable (lacking) part for the acceptance of HGT for Martin (2017). The fact that we have a precise mechanism here proves that the integrations we recovered are real, since we could not have found the expected pattern of a HIM-mediated integration just by chance. Nonetheless, this mechanism applies to HT from polydnaviruses only, so a lot of work remains to be done in order to understand how HT takes place in other systems.

Regarding HTT, many examples were recovered, in addition to which I also found thousands of new events in the large-scale study I achieved in metazoa (chapter 13). I also expect to find many additional events in the second large-scale study in insects, which was still in progress when writing this manuscript (chapter 14). It is known that TE have various impacts on the evolution of eukaryotic genomes (Schaack *et al.* 2010; Bourque *et al.* 2018): their amplification increases genome size, they are a source of mutations and genetic polymorphisms, they can be involved in genome rearrangements, they can regulate genes, and they can be domesticated as new genes. One can expect most of these impacts to have negative effects on the host (such as cancers), yet they contribute to genome diversity. For example, (i) the insertion of a TE in an intron of the peppered moth led to wing pattern melanisation bringing a selective advantage in polluted areas, (ii) RAG proteins, which are important for V(D)J recombination in jawed vertebrates, is a co-opted DNA transposon, (iii) the tail loss in great apes is due to the insertion of a SINE in an intron, and (iv) about one fourth of the human promotor regions would contain sequences derived from TE (Hayward and Gilbert 2022). HT being often accountable for TE spreading, one can expect HTT to have an important impact on genomes. In this sense, Gilbert and Feschotte (2018) analyzed 28 TE with phenotypic consequences, and found strong evidence that at least 6 of them originated from HT. This result supports the idea that HTT can have major evolutionary consequences in plants, their dataset being biased toward that clade due to a more thorough knowledge on phenotypic consequences of TE for agronomical selection. Their examples cover different ways HTT can impact genomes: (i) the LTR retrotransposon Tcs2 was transferred (directly or not) from asparagus to orange, leading to the cold-dependent overexpression of the gene Ruby, conferring a deeper pigmentation, and (ii) the LTR retrotransposon Rider was transferred (directly or not) from spinach to tomato, leading to the duplication of the gene SUN in Roman tomato (and thus to the overexpression of its gene, conferring an elongated form) and to the disruption of the gene PSY1 in yellow tomatoes due to insertion in this gene. It would be interesting to recover such striking examples in metazoans. In order to have a better idea of the strength of HTT compared to other evolutionary forces, future works could look at whether high rates of speciation are linked to high rates of HTT.

As a counterbalance to Martin's position, some researchers suggested that HT were so preponderant in early evolution that it would explain (i) the universality of the genetic code, and (ii) the difficulty to solve the deepest clades of the tree of life, and particularly the emergence of eukaryotes (Vetsigian *et al.* 2006; Syvanen 2012). (i) Many theories were proposed to explain the universality of

the genetic code across all living organisms, the main one being the LUCA (Last Universal Common Ancestor) theory, in which all living organisms would have arisen from a shared common ancestor, inheriting its genetic code. A second theory is the stereochemical theory, in which inherent chemical interactions between codons and their amino acids would be decisive. Although this second theory was initially not considered as the most likely, its support grew together with increasing evidence of rampant HT: Syvanen (2012) suggested that the genetic code evolved in multiple lineages connected by HT. (ii) The most supported theory is that eukaryotes is the sister group of Archaebacteria, yet more than 20 other theories exist. Pisani *et al.* (2007) analyzed a set of 2300 eukaryotic genes, and found that half of them are more closely related to bacteria (from different lineages) than to archaea, but 9.6% are more related to archaea, and 36.6% have no prokaryotic homolog, illustrating the complexity of the problem, and the multiple horizontal exchanges that led to extant eukaryotes.

Appendix

A - Résumé détaillé en français

A.1 . Introduction

Le matériel génétique est transmis des parents aux descendants par transmission verticale (reproduction), suivant la génétique mendélienne. Cependant, en 1928, Griffith a effectué une expérience lors de laquelle des souris sont mortes suite à une infection à la fois par une bactérie vivante mais non virulente et une bactérie virulente mais morte. En 1944, Avery et McCart ont compris que l'ADN des bactéries virulentes mortes avait été transmis aux bactéries non virulentes vivantes, les rendant virulentes. Cette expérience est la première découverte d'un transfert horizontal d'ADN (TH). Les transferts horizontaux peuvent ainsi être définis comme la transmission de matériel génétique par d'autres moyens que la reproduction (qu'elle soit sexuée ou asexuée), éventuellement entre espèces génétiquement éloignées, par opposition à la transmission verticale.

De nombreuses études ont montré que les TH sont fréquents chez les procaryotes (organismes sans noyau), chez lesquels environ 81% des gènes auraient été acquis horizontalement à un moment donné de leur évolution. Les TH sont notamment impliqués dans l'échange de gènes de résistance aux antibiotiques chez les bactéries. Les TH sont donc une force évolutive majeure chez les procaryotes, et les mécanismes impliqués sont bien connus.

Chez les eucaryotes cependant (organismes dont l'ADN est dans un noyau), les barrières biologiques additionnelles ont poussé les chercheurs à penser que les TH sont une caractéristique propre aux procaryotes. Une première barrière est la présence d'un noyau qu'il faut traverser pour atteindre l'ADN, une seconde est la présence de cellules spécialisées dans la reproduction chez certains organismes multicellulaires comme les animaux. En effet, seuls les TH ayant lieu dans ces cellules (cellules germinales dans le cas des organismes sexués) seront transmis à la descendance. Malgré ces barrières, de plus en plus d'exemples de TH chez les eucaryotes ont été rapportés, démontrant que ces barrières ne sont pas insurmontables. Chez les protistes par exemple (organismes eucaryotes unicellulaires), il a été estimé qu'environ 1% de leurs gènes proviennent de TH. Quant aux organismes multicellulaires, les chercheurs continuent de découvrir de plus en plus de TH, bien que la plupart proviennent de bactéries ou de virus. Par exemple, des algues ont acquis un gène bactérien leurs permettant de survivre dans des conditions extrêmes, des insectes ont acquis un gène bactérien leur permettant de se nourrir de plantes produisant des toxines, l'ancêtre des vertébrés a acquis un gène bactérien important dans l'évolution de la vision, plusieurs mammifères ont acquis indépendamment un gène viral important pour la formation du placenta, et des guêpes ont acquis des gènes viraux leur permettant de produire des particules virales.

Enfin, le TH *entre* organismes multicellulaires, c'est à dire d'un multicellulaire à un autre, semble être le type de transfert se produisant le moins souvent. Bien que les exemples de TH de ce type impliquant des gènes (HGT) sont encore assez anecdotiques, de nombreux exemples de TH d'éléments transposables (HTT) ont été rapportés entre eucaryotes multicellulaires. Trois études à large échelle ont respectivement montré que 65% des plantes analysées ont subi au moins un évènement d'HTT, qu'il y a eu au moins 2248 évènements d'HTT parmi 195 insectes analysés, et 975 évènements d'HTT parmi les 307 vertébrés analysés. De plus, il a été estimé que 24 des 26 lignées de *mariner* (une famille

d'ET) testés sont impliqué dans des HTT entre 20 génomes de *Drosophia*. Une autre étude a pu estimer que les 3 espèces de *Drosophila* analysées échangent des ET à une fréquence de 0.04 HT par famille d'ET et par million d'années.

Un élément transposable (ET) est une séquence d'ADN dite égoïste, qui est mobile et qui peut se multiplier au sein d'un génome. De tels éléments sont présents chez tous les êtres vivants, mais en proportion très diverse selon les espèces. Les ET représentent près de la moitié du génome humain (alors que les gènes codant des protéines ne représentent que 2%), ∼20% du génome de Drosophile, et ∼80% du génome de maïs par exemple. Cependant, seule une partie de ces ET sont toujours actifs dans les génomes, c'est à dire que seule une partie a toujours la capacité de s'amplifier (moins de 0.05% des ET humains, mais environ 30% des ET de Drosophiles). Cette capacité à s'exciser de leurs génomes, à traverser l'enveloppe nucléaire, à s'insérer à une autre position génomique, puis à se multiplier une fois insérés, ainsi que leur grand nombre, expliquerait pourquoi il y a tant de TH d'ET, comparé aux TH de gènes.

Un autre type d'élément mobile d'intérêt ici, sont les polydnavirus (PDV), éléments présents dans le génome d'un grand nombre de guêpes parasitoïdes. Les PDV sont présents dans deux familles de guêpes : les Braconidae et les Ichneumonidae, c'est pourquoi les PDV de la première famille sont appelés bracovirus, alors que ceux de la seconde famille sont appelés ichnovirus. Dans les deux cas, les PDV sont composés de deux éléments, tous deux présents dans le génome de la guêpe : (i) plusieurs segments proviraux et (ii) des gènes d'origine virale (figure 2.4). Les femelles ont la capacité d'amplifier ces segments proviraux, de les exciser de leurs génomes, de les circulariser, puis de les empaqueter dans des particules virales. Ces dernières sont produites grâce aux gènes d'origine virale. Ces gènes ont été acquis par transferts horizontaux plusieurs fois dans l'histoire évolutive des guêpes parasitoïdes : les guêpes Braconidae les ont acquis suite à un TH d'un nudivirus ayant eu lieu il y a environ 100 millions d'années, alors que les guêpes Ichneumonidae les ont acquis plusieurs fois (au moins deux) de façon indépendante, et d'un virus inconnu à ce jour (figure 2.2). Une fois ces particules assemblées, elles sont injectées dans l'hôte en même temps que les oeufs, où les cercles ADN peuvent exprimer leurs gènes de virulence, assurant le bon développement larvaire des guêpes (figure 2.3). De façon intéressante, plusieurs études ont montré qu'au moins certains cercles ADN étaient capables de s'intégrer dans le génome de l'hôte, grâce à un motif ADN bien particulier : le HIM (pour Host Integration Motif). En raison de leur mobilité, de leur capacité à entrer dans les cellules et de leur propre mécanisme d'intégration, les PDV sont une source intéressante d'ADN pour les TH. Malgré le fait que l'intégration soit souvent une impasse évolutive, puisque la plupart des hôtes parasités meurent, plusieurs exemples de TH de PDV ont été rapportés. La première étude rapportant des TH de PDV a trouvé 105 régions dans deux génomes de lépidoptères dérivant de PDV. Depuis, un HT de PDV conférant un avantage évolutif au lépidoptère receveur a même été identifié: le gène *SI gasmin* joue un rôle important lors de la réponse immunitaire anti-bactérienne.

A.1.1 . Mécanismes lors d'un TH

Bien que les PDV possèdent leurs propres mécanismes pour subir un TH, ils ne sont présents que chez certaines guêpes parasitoïdes (ce qui représente tout de même plusieurs dizaines de milliers d'espèces), les mécanismes permettant un TH dans les autres espèces restent donc à découvrir. Nous avons vu que les ET sont eux présents dans toutes les espèces et qu'ils possèdent certaines caractéristiques leurs permettant de subir un TH (l'insertion notamment), cependant ils ne remplissent

pas tous les mécanismes nécessaires, comme sortir de sa cellule, atteindre un autre organisme, et entrer dans les cellules de ce dernier. Quant aux TH de gènes, leurs mécanismes sont une boîte noire. Ci-après, je décris les mécanismes putatifs pouvant expliquer chaque étape requise afin qu'un TH se produise, mais ils ne sont bien sûr pas exclusifs les uns des autres, et pourraient même tous co-exister.

La première barrière à passer est la sortie de l'organisme donneur et l'entrée dans l'organisme receveur. Les TH entre eucaryotes pourraient théoriquement avoir lieu par l'acquisition de matériel génétique nu, par l'alimentation, ou via un vecteur. Du matériel génétique nu circule dans les fluides animaux (sang, salive, etc.), mais on ne sait pas au bout de combien de temps il est dégradé. Concernant l'alimentation, de tels cas de TH n'ont jamais été démontrés. Les vecteurs sont donc les meilleurs candidats à ce jour pour compléter un TH. Un vecteur peut être n'importe quelle entité pouvant transporter du matériel génétique du donneur à la cellule receveuse. Plusieurs vecteurs ont été proposés, mais je me suis intéressée uniquement aux virus lors de ma thèse.

Pour qu'un virus soit vecteur d'un TH, nous devons découvrir (i) si les virus peuvent recevoir et transporter du matériel génétique étranger, (ii) quelles sont les chances que le matériel génétique nouvellement acquis persiste assez longtemps pour être transmis à une autre espèce, et (iii) si du matériel génétique peut s'intégrer dans un eucaryote à partir d'un virus.

Pour le premier point, plusieurs études ont montré que les rétrovirus pouvaient encapsider l'ARN de l'hôte, possiblement jusqu'à 50% de l'ARN total encapsidé. La capacité des virus à encapsider de l'ADN étranger reste toujours à évaluer. Alternativement, le matériel génétique pourrait s'insérer dans le génome du virus, au lieu d'être librement transporté dans ses particules virales. En ce sens, il a été montré que de nombreux gènes codés par des grands virus à ADN proviennent de leurs hôtes eucaryotes, et il a été estimé qu'environ 5% des génomes d'AcMNPV (un grand virus à ADN) abritaient une insertion *de novo* d'ET provenant de papillons infectés. Cette étude démontre que les TH d'un hôte vers un virus sont assez fréquents.

Pour le second point, que le matériel génétique soit inséré dans le virus ou transporté librement, il doit persister assez longtemps afin d'être transmis à une autre espèce. Ainsi, la deuxième étape d'un TH médié par un virus (transmettre la séquence nouvellement acquise à une autre espèce) doit avoir lieu avant que le matériel génétique libre ne soit dégradé, ou avant que le matériel génétique intégré ne soit purgé de la population virale. Les virus ayant une importante densité en gènes, nous pouvons nous attendre à ce que la plupart des insertions dans leurs génomes soient délétères et ne persistent qu'à très faible fréquence et seulement sur quelques cycles de réplication virale.

Enfin, il faut que le matériel génétique porté par un virus (dans son génome ou libre) puisse être transmis à une seconde espèce. Les TH d'un virus à un organisme multicellulaire sont en fait très courant en laboratoire, puisque c'est simplement ce qui se passe lorsque l'on fait de la transgenèse. De plus, j'ai précédemment donné quelques exemples de TH provenant de virus. Bien que la fréquence à laquelle cette endogénéisation se produit reste à évaluer, le fait que la majorité des éléments viraux endogènes (EVE) des Arthropodes soient spécifiques à certaines espèces, et parfois non fixés dans une espèce, suggère que l'endogénéisation est assez fréquente dans ces espèces, et est toujours en cours. En plus des virus libres, les polydnavirus sont aussi de très bon candidats de HT entre les guêpes parasitoïdes et leurs hôtes.

Une fois le matériel génétique dans la cellule receveuse, il doit encore atteindre le noyau et s'insérer dans le génome. Il a été montré que l'ADN libre est rapidement dégradé dans la cellule, mais d'un autre côté il peut aussi être associé à des complexes protéiques qui le protègent et peuvent le transporter

jusqu'au noyau. Pour ce qui est du matériel génétique arrivé dans la cellule grâce à un vecteur, il pourrait aussi atteindre le noyau via ce vecteur si ce vecteur a la capacité d'y entrer. Enfin, le matériel génétique pourrait s'intégrer dans le génome via divers mécanismes : (i) s'il est intégré dans le génome d'un virus dont l'intégration fait partie de son cycle de vie (les rétrovirus), ou tout autre virus qui s'intègrerait par accident, il sera intégré par la même occasion. (ii) dans le cas des polydnavirus et de certains ET, nous avons vu qu'ils ont leurs propres mécanismes leurs permettant de s'intégrer. Pour tous les autres cas (gènes, ou ET non autonomes par exemple), l'ADN pourrait s'intégrer via les mécanismes de réparation ADN de l'espèce receveuse. Quant à l'ARN, il pourrait s'intégrer par transcription inverse.

Même si le matériel génétique passe avec succès toutes ces barrières, il n'est pas nécessairement transmis au fil des générations. Pour cela, l'intégration doit avoir lieu dans une cellule qui sera transmise à la descendance. Dans le cas des organismes sexués, ces cellules se limitent aux cellules germinales, mais seul un sous-ensemble est transmis à la progéniture. C'est pourquoi les vecteurs qui ciblent les cellules germinales sont de meilleurs candidats. Alternativement, le matériel génétique pourrait également s'insérer aux premiers stades de l'embryogenèse, lorsque les cellules souches pluripotentes permettent un accès direct à la lignée germinale de la prochaine génération. Une fois transmises à la progéniture, toutes les cellules possèdent le matériel génétique nouvellement acquis, qui sera donc automatiquement transmis aux générations suivantes, sauf en cas d'absence de reproduction bien sûr. Ensuite, le sort du matériel génétique nouvellement acquis dans la population est aux mains des forces évolutives classiques, comme le reste du génome, c'est-à-dire l'équilibre entre dérive génétique et pression sélective. Étant donné que les insertions dans les régions codant des protéines et dans les régions régulatrices ont un impact négatif, on peut s'attendre à ce qu'un certain nombre d'insertions aient une faible probabilité d'être maintenues dans une population. En revanche, si l'insertion est cooptée et procure un avantage sélectif, cela augmenterait considérablement ses chances de fixation dans la population. Dans le cas des HTT, il a été montré que malgré l'effet délétère de la transposition sur le génome de l'hôte, la transposition augmente leur chance de persistance, sauf bien sûr lorsque le taux de transposition est si élevé qu'il conduit à la stérilité ou la mort de leur hôte.

A.1.2 . Facteurs influençant le nombre de TH

De nombreux facteurs écologiques pouvant influencer le nombre de TH ont été proposés : (i) les relations proies-prédateurs, (ii) les relations hôtes-parasites, (iii) la proximité géographique, et (iv) l'habitat. Dans le cas des relations hôtes-parasites, les insertions de polydnavirus dans les hôtes de guêpe est un exemple. Concernant la proximité géographique, l'étude ayant trouvé 2248 évènements d'HTT parmi 195 insectes a pu détecter une corrélation entre le nombre de HTT et la proximité géographique (ainsi que la proximité phylogénétique). Cependant, les auteurs de l'étude ne connaissaient pas la localisation exacte de leurs insectes, donc il n'ont pu montrer l'impact de ce facteur avec précision. Bien qu'une telle corrélation soit attendue, son impact peut être moins fort si des vecteurs ayant une forte capacité à se diffuser entre régions du monde sont impliqués. Pour ce qui est de la proximité phylogénétique, ce résultat suggère qu'un HTT a plus de chance de se produire entre deux espèces génétiquement proches. Enfin, il a été proposé que le fait de vivre dans un habitat aquatique pourrait augmenter le nombre de TH, l'eau jouant le rôle de connecteur écologique. En ce sens, l'étude rapportant 975 HTT parmi 307 vertébrés a estimé que 93.7% de ces transferts concernent

des poissons. Cependant, les auteurs n'ont pas pu déchiffrer si ce grand nombre de HTT chez les poissons est dû à leur habitat aquatique ou est spécifique aux poissons pour d'autres raisons. Ce grand nombre de HTT pourrait aussi être spécifique aux espèces aquatiques à fécondation externe, l'ADN libre se trouvant dans l'eau pouvant s'attacher au sperme lors de la fécondation, ou même accéder directement à l'oeuf. En plus des facteurs écologiques, des caractéristiques au niveau du génome et de la structure de la population (du donneur ou du receveur) pourraient aussi influencer la probabilité d'un TH.

A.1.3 . Organisation de la thèse

Bien que ce soit désormais un consensus que les TH sont possibles entre organismes multicellulaires, leur fréquence, les mécanismes, et les facteurs impliqués sont toujours au stade de spéculations. Lors de cette thèse, je m'intéresse à ces aspects, me concentrant sur les TH entre animaux, en insistant tout particulièrement sur les insectes. Mon travail est organisé autour de trois parties. La partie I s'intitule "Les virus libres comme vecteurs de transferts horizontaux". Ici, j'ai brièvement cherché à savoir si des virus libres pouvaient agir comme vecteur de HTT, c'est-à-dire transporter l'ADN d'un organisme à un autre. La partie II s'intitule "Les virus domestiqués des guêpes parasitoïdes comme vecteurs de transferts horizontaux", lors duquel j'ai aussi cherché à savoir si des virus peuvent agir comme vecteurs, mais cette fois-ci en me focalisant sur un type de virus très particulier : les polydnavirus, qui sont des virus domestiqués présents dans le génome de nombreuses guêpes parasitoïdes. Cette partie permet aussi d'étudier une interaction parasitoïde/hôte, qui pourrait promouvoir les transferts horizontaux. Pour terminer, la partie III s'intitule "Facteurs influençant les transferts horizontaux", partie dans laquelle j'effectue deux études à large échelle, chacune évaluant l'impact de deux facteurs sur les transferts horizontaux: l'habitat (terrestre ou aquatique) et le mode de fécondation (interne ou externe) pour la première, la proximité géographique et phylogénétique pour la seconde. Ces deux études sont toujours en cours au moment de la rédaction de cette thèse.

A.2 . Résultats et discussion

A.2.1 . Partie I: Virus libres comme vecteurs de transferts horizontaux

Cette partie a fait l'objet d'une publication, dans laquelle nous avons montré qu'une infection virale (par le baculovirus AcMNPV) peut impacter l'expression des ET de l'hôte, bien que modérément (Article n°1). Une augmentation de l'activité des ET pourrait augmenter les chances qu'un ET transpose dans le virus lors de l'infection (première étape d'un TH entre animaux médié par un virus libre). De plus, nous avons trouvé qu'un ET dérivé de l'hôte mais intégré dans les génomes viraux était fortement exprimé (TFP3 dans la figure 3 et la table 2 de l'article n°1). Cela signifie qu'une fois insérés dans le virus, les ET peuvent encore être actifs, et pourraient donc transposer dans un autre hôte lors d'une infection secondaire (seconde étape d'un TH entre animaux médié par un virus libre).

Il serait intéressant dans de futures études d'observer un TH complet médié par un virus libre. Pour cela, nous pourrions utiliser un virus possédant des intégrations d'ET d'une espèce, pour infecter une autre espèce. AcMNPV est un très bon candidat comme vecteur en raison de sa large gamme d'hôtes et de son grand génome à ADN double brin, mais il serait intéressant de tester aussi d'autres virus, et de réaliser des infections multiples.

A.2.2 . Partie II: Virus domestiqués des guêpes parasitoïdes comme vecteurs de transferts horizontaux

Cette partie, qui a fait l'objet de quatre publications, consiste à évaluer si les polydnavirus peuvent promouvoir des TH entre guêpes parasitoïdes et leurs hôtes lépidoptères. Nous nous sommes particulièrement intéressés à la guêpes *Cotesia typhae* et son unique hôte naturel connu, *Sesamia nonagrioides*. *S. nonagrioides*, plus connu sous le nom de Sésamie du maïs, est un ravageur majeur des cultures de maïs en Europe méditerranéenne et en Afrique (photos dans la figure 2.3). *C. typhae*, ne vit qu'en Afrique de l'est où elle parasite les populations locales de *S. nonagrioides*. Il s'agit d'une guêpe parasitoïde actuellement étudiée au laboratoire afin de déterminer si elle pourrait être utilisée comme agent de lutte biologique en France. Pour cela, les éventuels effets secondaires doivent être méticuleusement étudiés, comme la possibilité de transferts horizontaux entre les populations importées de *C. typhae* et les lépidoptères français, comme *S. nonagrioides*, mais aussi éventuellement d'autres espèces non cibles.

Pour cela, il a d'abord fallu assembler le génomes de *C. typhae* et celui de *S. nonagrioides*. Le génome de *S. nonagrioides* a fait l'objet d'une publication à part entière (Article n°2), dans laquelle nous présentons un assemblage des génomes nucléaire et mitochondrial, réalisé à partir d'un séquençage de reads courts et longs, extrait d'un pool de larves consanguines mâles et femelles. L'assemblage du génome nucléaire est de 1021 Mbp, est composé de 2553 scaffolds, et a une N50 de 1105 kbp. Ce génome est plus de deux fois plus grand que toutes les espèces de Noctuidae séquencées à ce jour, principalement en raison d'un taux de répétition élevé (table 2 et figure 1 de l'article n°2). En effet, 62.94% de son génome est annoté comme éléments répétés, dont 33.88% sont assignés à des éléments transposables. Le génome mitochondrial complet est de 15,330 pb et contient tous les gènes codant les protéines et ARN attendus. Nous avons pu trouver un cluster "mitochondrial nuclear DNA" (NUMTs), qui résulte de l'intégration nucléaire récente de deux copies du génome mitochondrial, dont l'une est réarrangée en trois parties (figure 2 de l'article n°2). En plus des 17230 gènes nucléaires codant des protéines que nous avons prédits automatiquement, nous avons annoté à la main des gènes impliqués dans la détermination du sexe (dsx, IMP, PSI) et des gènes de l'alpha-amylase. Une bonne connaissance de la détermination du sexe chez une espèce de ravageur pourrait être utile dans le cadre de la technique de l'insecte stérile utilisée en lutte biologique. Quant aux gènes de l'alpha-amylase, ils pourraient être impliqués dans l'interaction avec les guêpes parasitoïdes. Dans cette première étude, nous n'avons trouvé aucune trace de transfert horizontal récent de gènes de bracovirus de guêpes parasitoïdes.

Le génome de *Cotesia typhae* est lui inclus dans une publication plus conséquente, dans laquelle nous étudions les intégrations de ses bracovirus (CtBV) dans les génomes somatiques de *S. nonagrioides* suite au parasitisme (Article n°3). Pour assembler le génome de *C. typhae*, nous avons extrait l'ADN d'un pool d'individus provenant d'une lignée consanguines de notre laboratoire, que nous avons ensuite séquencé en reads courts et longs, afin de réaliser un assemblage hybride de son génome. Indépendamment, nous avons aussi parasité plusieurs chenilles *S. nonagrioides* par *C. typhae*, puis nous avons extrait l'ADN de ses chenilles parasitées sept jours plus tard. Tout d'abord, le génome nucléaire de *C. typhae* que nous avons assemblé fait 187Mb et est constitué de seulement 72 scaffolds et d'une N50 de 6.81Mb. 98.9% des gènes BUSCO d'insectes sont présents. Ces statistiques démontrent un assemblage de très bonne qualité. Dans ce génome, nous avons pu trouver 27 segments proviraux (figure 1 de l'article n°3), dont 26 ayant probablement la capacité de former des cercles ADN, qui sont

ensuite injectés dans la chenille lors du parasitisme, en même temps que les oeufs de guêpes. Cependant, sur ces 27 segments, seuls 16 possèdent des HIM (Host Integration Motif), motif permettant de s'intégrer dans le génome de la chenille (figure 3 de l'article n°3). Nous avons trouvé que 96.5% des intégrations ont lieu via ces HIMs, ce qui montre que c'est le mécanisme d'intégration majeur des bracovirus de *C. typhae*. De plus, nous avons observé une sur-représentation de micro-homologies de 1 et 2-pb entre les bracovirus et la chenille au niveau des jonctions d'intégration, ce qui démontre que les bracovirus peuvent s'intégrer de deux façons, avec ou sans micro-homologies (figure 4 de l'article n°3). Nous ignorons pour l'instant si les intégrations médiées par micro-homologies sont générées par le même mécanisme que les intégrations dépourvues de micro-homologie, ou si elles se produisent via les mécanismes de réparation de l'ADN de la chenille par exemple. Quels que soient les mécanismes d'intégration, nous avons observé des intégrations dans tous les tissus/structures de la chenille que nous avons étudiés : l'hémolymphe, le corps gras, les chaînes ganglionnaires et la tête. C'est la première preuve d'intégration de bracovirus ailleurs que dans l'hémolymphe, bien que ce tissu soit celui qui contient le plus d'intégrations (figure 6b de l'article n°3). En effet, nous avons estimé qu'en moyenne dans l'hémolymphe, chaque génome haploïde est sujet à 85 évènements d'intégration, contre 12 dans le corps gras par exemple. Nous avons aussi trouvé de nombreux évènements d'intégration dans une chenille pour laquelle le parasitisme a échoué (échantillon "Whole body" de la figure 6b de l'article n°3). Ce résultat est très intéressant car le parasitisme ayant échoué, cette chenille aurait pu survivre jusqu'à l'âge adulte et se reproduire. Cela nous amène donc à nous demander si les chenilles survivant au parasitisme peuvent transmettre cet ADN de guêpe à leur descendants, ce que nous étudions dans l'article suivant. Avant cela, un autre résultat surprenant est celui où nous comparons la quantité de bracovirus sous sa forme circulaire *versus* intégrée. Il était traditionnellement supposé que l'intégration des cercles de polydnavirus est bénéfique pour la guêpe car elle permettrait la persistance (alors que les cercles non intégrés seraient dégradés) et l'expression de ces gènes pendant toute la durée du développement embryonnaire et larvaire, qui peut durer entre 7 et 14 jours en laboratoire, selon les espèces considérées. Ici, nous démontrons que 7 jours après le parasitisme, les bracovirus circulaires sont tout aussi nombreux que les bracovirus intégrés, à l'exception des cercles de bracovirus 1 et 7 qui s'intègrent en très grande quantité, et sont de loin les plus abondants (figure 8 de l'article n°3). Il s'ensuit que l'intégration n'est pas une exigence pour la persistance pendant au moins la moitié de la durée du développement embryonnaire de *C. typhae*. D'autres études seront nécessaires pour éclairer le rôle de l'intégration jusqu'à l'étape de nymphose des larves de guêpes.

Suite aux résultats de l'article précédent, nous nous sommes demandé si les chenilles survivant au parasitisme pouvaient transmettre cet ADN de guêpe intégré à leurs descendants (Article n°4). Pour cela, nous avons parasité 62 chenilles *S. nonagrioides* par *C. typhae*, et nous avons formé des couples avec les chenilles survivantes. Sur les 15 couples formés, seuls cinq ont produit des descendants. Nous avons séquencé en Illumina 500 de ces descendants, ainsi que deux femelles adultes ayant survécu au parasitisme: une ayant eu de nombreux descendants, l'autre aucun. Nous avons trouvé de nombreuses intégrations de bracovirus dans ces deux femelles, bien qu'en plus grande quantité dans la seconde (figure 3b de l'article n°4). Cette persistance d'intégration chez les adultes était encourageante pour notre recherche sur les descendants, cependant, nous n'avons pas détecté la moindre intégrations chez ces descendants. Il est surprenant de n'avoir détecté aucune transmission alors que nous avons estimé que la femelle ayant eu des descendants possédait en moyenne deux évènements d'intégration par génome haploïde. Cette absence de transmission pourrait être expliquée de deux façons: (i) un plus

faible nombre d'évènements d'intégration dans les gamètes que dans les autres types cellulaires (ce qui pourrait être possible si les gamètes sont moins accessibles par les bracovirus), ou (ii) les gamètes sont autant sujets aux intégrations de bracovirus que les autres types cellulaires, mais les gamètes portant les intégrations pourraient être négativement impactés, empêchant leur transmission. Afin d'essayer de répondre à cette question, Inès Matrougui en stage de M2 lors de ma dernière année de thèse, est en train d'étudier cette question en estimant le nombre d'évènements d'intégration de CtBV dans les gonades de *S. nonagrioides* survivant au parasitisme, ainsi que dans leurs oeufs.

Dans ce même article, nous nous sommes aussi demandé si les bracovirus pouvaient s'intégrer dans les génomes somatiques d'espèces autres que *S. nonagrioides*. Cette question est particulièrement pertinente dans le cadre de la lutte biologique, lors de laquelle *C. typhae* pourrait être introduite en France afin de contrôler les populations de *S. nonagrioides* ravageuses de cultures de maïs. Dans un tel contexte, il n'est pas impossible que *C. typhae* parasite d'autres chenilles non-cibles, auquel cas nous devons estimer les risques sur ces populations de chenilles, comme le risque de subir des intégrations de bracovirus dans leurs génomes. Nous avons donc choisi trois espèces non-cibles partageant la même niche écologique que *S.nonagrioides* en France : *Nonagria typhae*, *Globia sparganii*, et *Chilo phragmitella*. Nous avons parasité un individu de chaque espèce. Nous les avons ensuite séquencés en Illumina à leur mort qui a eu lieu au stade larvaire pour les trois individus. Nous avons pu identifier de nombreuses intégrations dans les trois espèces, dans des quantités similaires à ce que nous avons observé pour les chenilles de *S.nonagrioides* parasitées (figure 3b de l'article n°4). Ce résultat démontre que les bracovirus de *C. typhae* sont capables de rentrer dans les cellules et de s'intégrer dans les génomes d'une grande variété d'espèces de lépidoptères, même dans un Crambidae (*C. phragmitella*) qui a divergé il y a plus de 100 millions d'années de *S. nonagrioides*. Il semble donc que ces deux processus ne soient pas liés à la permissivité des hôtes, puisque des études en cours au laboratoire suggèrent que ces trois espèces non-cibles permettent très rarement aux oeufs de *C. typhae* de se développer. Enfin, nous avons aussi noté que les bracovirus s'intègrent, en partie, différemment dans le Crambidae, l'espèce la plus divergente. En effet, chez *C. phragmitella* seulement 74,3% des intégrations ont eu lieu via HIM, contre 96,5% chez *S. nonagrioides* (figure 5 de l'article n°4). Cette différence montre que des facteurs hôtes sont impliqués lors de l'intégration, bien qu'ils restent à déterminer.

Pour finir, dans le dernier article de cette partie, nous nous sommes intéressés, non plus au bracovirus de *C. typhae*, mais à un polydnavirus d'une autre famille : un ichnovirus (Article n°5). Grâce à la même méthode que celle que nous avons développée dans les deux articles précédents, Camille Heisserer, étudiante en stage de M2 que j'ai co-encadrée, a pu trouver de nombreuses intégrations d'ichnovirus de la guêpe *Hyposoter didymator* dans son hôte *Spodoptera frugiperda* (figure 3 de l'article n°5). Le processus d'intégration chez cet ichnovirus est très similaire à ce que nous avons pu observer chez *C. typhae*, aussi bien en termes de quantité d'intégration par cellule, que de mécanismes d'intégration via HIM (figure 4 de l'article n°5), avec là aussi une sur-représentation de micro-homologies aux jonctions ichnovirus/chenille (figure 5 de l'article n°5). Cette similarité est surprenante car ces deux familles de polydnavirus ont une origine évolutive indépendante et les séquences des HIMs ne sont d'ailleurs pas orthologues. Cette similarité pourrait être due à de la convergence évolutive qui serait apparue suite à une pression de sélection similaire dans les deux systèmes.

Dans un deuxième temps dans cet article, nous avons réalisé une recherche de polydnavirus, bracovirus et ichnovirus, dans les 775 génomes de lepidoptères disponibles sur NCBI. Nous avons pu

retrouver des transferts horizontaux médiés par HIM dans 124 de ces espèces, ce qui suggère que les intégrations via HIM sont probablement une voie majeure de transferts horizontaux de matériel génétique de guêpes à lépidoptères (figure 7 de l'article n°5).

A.2.3 . Partie III: Facteurs influençant les transferts horizontaux

Cette partie consiste à évaluer l'influence de quatre facteurs sur les transferts horizontaux. Deux facteurs, l'habitat (aquatique ou terrestre) et le mode de fécondation (interne ou externe), sont étudiés dans une première étude à large échelle incluant diverses espèces d'animaux, alors que les deux autres facteurs, la proximité géographique et la proximité phylogénétique, sont étudiés dans une seconde étude à large échelle se focalisant sur trois ordres d'insectes. Dans les deux études, nous nous sommes concentrés uniquement sur les transferts horizontaux d'éléments transposables (HTT) en raison de leur fort potentiel. Au moment de la rédaction de cette thèse, les deux études sont encore en cours.

Dans la première étude, nous avons inclus 247 génomes d'NCBI, en échantillonnant un maximum de transitions d'habitat et de transitions de mode de fécondation. Afin de maximiser ce nombre de transitions, nous avons travaillé à l'échelle des animaux et non des insectes. Nous avons annoté les éléments transposables dans ces 247 génomes afin de rechercher les évènements de HTT entre chaque paire d'espèces. Nous avons pu identifier 11,573 évènements d'HTT indépendant (figures 13.2 et 13.9). Bien que nos résultats préliminaires suggèrent que le milieu aquatique ne facilite pas plus le HTT que le milieu terrestre (figure 13.10A), ils suggèrent que les espèces à fécondation externe subissent plus de HTT que les espèces à fécondation interne (13.10C). Cependant, il reste encore à effectuer un teste statistique plus poussé avant de vraiment pouvoir conclure sur l'influence de ces deux facteurs sur le nombre d'évènement de HTT.

Dans la seconde étude, nous avons échantillonné 8 localités à travers le monde (afin de tester la proximité géographique, voir la carte d'échantillonage figure 14.1) et pour chaque localité, nous avons échantillonné 3 diptères, 3 lépidoptères et 3 hyménoptères (afin de tester la proximité phylogénétique : par exemple, y a-t-il plus de HTT entre les diptères qu'entre un diptère et un lépidoptère ?). Parmi ces 8 localités, les génomes anglais ont été directement téléchargés sur le site du projet Darwin Tree of Life, alors que nous avons récolté nous-mêmes dans la nature les insectes des 7 autres localités (quelques photos figure 14.5), avant d'extraire leurs ADN et d'assembler leurs génomes. Ayant un souci avec 4 échantillons, notre jeu de données final est de 68 génomes d'insectes (8*9-4=68). Après avoir réalisé un arbre phylogénétique de ce jeu de données (figure 14.6), nous avons commencer l'annotation des éléments transposables dans les génomes, mais nous n'avons pas encore identifié les HTT.

A.2.4 . Impact évolutif des transferts horizontaux

Les génomes sont façonnés par la sélection naturelle et la dérive génétique, changeant les fréquences des allèles (différentes versions d'un gène) dans les populations, parfois jusqu'à fixation (tous les individus d'une population ont le même allèle). La nouveauté génomique peut résulter de mutations spontanées, de la genèse de gènes, de duplications, de perte de gènes, de recombinaisons, d'introgressions (reproduction avec une autre espèce) ou de transferts horizontaux.

Chez les procaryotes, il est clair que les TH ont un impact majeur sur leur évolution, avec 81% de leurs gènes qui auraient été acquis horizontalement à un moment donné de leur évolution. On a d'abord pensé que les TH étaient propres aux procaryotes, mais nous avons désormais de nombreux

exemples de TH chez les eucaryotes, bien qu'en bien plus faible quantité que chez les procaryotes. On peut donc se demander quel est le réel impact des TH sur l'évolution des génomes eucaryotes ?

A travers cette thèse, nous avons pu identifier des centaines d'évènements de TH de polydnavirus chez les lépidoptères (partie II) et des milliers d'évènements de TH d'ET lors de notre première étude à large échelle (partie III). Bien que plusieurs exemples de domestication de ces insertions de polydnavirus par les lépidoptères ont été décrits, une analyse fonctionnelle à plus grande échelle reste à réaliser pour avoir une meilleure idée de l'impact de ces insertions sur l'évolution des lépidoptères. Quant aux ET, on sait qu'ils peuvent avoir divers impacts sur l'évolution des génomes eucaryotes : leur amplification augmente la taille du génome, ils sont source de mutations et de polymorphismes génétiques, ils peuvent être impliqués dans des réarrangements génomiques, ils peuvent réguler des gènes et ils peuvent être domestiqués en tant que nouveaux gènes. On peut s'attendre à ce que la plupart de ces impacts aient des effets négatifs sur l'hôte (induisant par exemple des cancers), mais ils contribuent aussi à la diversité du génome. Les TH étant souvent responsables de la diffusion d'ET dans les génomes, on peut s'attendre à ce que les TH d'ET aient un impact important. En ce sens, une étude a analysé 28 ET ayant des conséquences phénotypiques chez les plantes, et a trouvé des preuves solides que 6 d'entre eux proviennent de TH. Dans certains cas au moins, les TH d'ET peuvent donc induire des changements phénotypiques importants, tel que la forme allongée des tomates romaines ou la couleur des tomates jaunes.

Ces nombreux évènements de TH et les impacts possibles des ET listés ci-dessus, suggèrent que les TH d'ET ont probablement un impact important sur les génomes analysés, bien que la force de ce processus évolutif comparée aux autres forces évolutives reste à déterminer. Quant au nombre élevé TH de polydnavirus, ils suggèrent eux aussi que ces TH ont un impact important sur l'évolution des lépidoptères.

Néanmoins, je ne pense pas que l'impact des TH sur l'évolution soit uniquement une question de fréquence. Même si l'endogénéisation du rétrovirus impliqué dans la formation du placenta était le seul événement de TH dans ce groupe, pourrait-on vraiment dire qu'il n'a eut qu'un impact mineur sur l'évolution de ce groupe ? C'est pourquoi je pense que la réponse à la question "Quel est l'impact des TH sur l'évolution des animaux ?" est bipartite : en plus de l'aspect quantitatif, il faut aussi regarder l'aspect qualitatif, puisqu'un seul transfert peut avoir des conséquences majeures sur l'évolution d'un groupe. En ce qui concerne les TH de gènes, ils ont beau se produire en faible fréquence, il est clair que ces transferts peuvent apporter de remarquables adaptations. Des exemples de tel transferts provenant de bactéries et de virus (adaptation aux environnements extrêmes, protection aux infections bactériennes, amélioration de la vision chez les vertébrés), mais aussi entre animaux (résistance au froid ou encore protection aux infections bactériennes) ont été décrits, ainsi que des exemples provenant d'autres organismes multicellulaires (résistance à des toxines, pigmentation, parade nuptiale).

Bibliography

Ahmed, M. Z., S.-J. Li, X. Xue, X.-J. Yin, S.-X. Ren, *et al.*, 2015 The Intracellular Bacterium Wolbachia Uses Parasitoid Wasps as Phoretic Vectors for Efficient Horizontal Transmission. PLOS Pathogens **11**: e1004672.

Allio, R., S. Donega, N. Galtier, and B. Nabholz, 2017 Large Variation in the Ratio of Mitochondrial to Nuclear Mutation Rate across Animals: Implications for Genetic Diversity and the Use of Mitochondrial DNA as a Molecular Marker. Molecular Biology and Evolution **34**: 2762–2772.

Altincicek, B., J. L. Kovacs, and N. M. Gerardo, 2012 Horizontally transferred fungal carotenoid genes in the two-spotted spider mite Tetranychus urticae. Biology Letters **8**: 253–257.

Archibald, J. M., 2015 Endosymbiosis and Eukaryotic Cell Evolution. Current Biology **25**: R911–R921.

Bai, H., G. M. S. Lester, L. C. Petishnok, and D. A. Dean, 2017 Cytoplasmic transport and nuclear import of plasmid DNA. Bioscience Reports **37**: BSR20160616.

Baidouri, M. E., M.-C. Carpentier, R. Cooke, D. Gao, E. Lasserre, *et al.*, 2014 Widespread and frequent horizontal transfers of transposable elements in plants. Genome Research **24**: 831–838.

Balaj, L., R. Lessard, L. Dai, Y.-J. Cho, S. L. Pomeroy, *et al.*, 2011 Tumour microvesicles contain retrotransposon elements and amplified oncogene sequences. Nature Communications **2**: 180.

Bartolomé, C., X. Bello, and X. Maside, 2009 Widespread evidence for horizontal transfer of transposable elements across Drosophilagenomes. Genome Biology **10**: R22.

Beck, M. H., S. Zhang, K. Bitra, G. R. Burke, and M. R. Strand, 2011 The Encapsidated Genome of Microplitis demolitor Bracovirus Integrates into the Host Pseudoplusia includens. Journal of Virology **85**: 11685–11696.

Beckage, N. and J.-M. Drezen, editors, 2011 *Parasitoid Viruses: Symbionts and Pathogens*. Academic Press.

Beckage, N. E. and D. B. Gelman, 2004 Wasp parasitoid disruption of host development: Implications for new biologically based strategies for insect control. Annual Review of Entomology **49**: 299–330.

Beiko, R. G., T. J. Harlow, and M. A. Ragan, 2005 Highways of gene sharing in prokaryotes. Proceedings of the National Academy of Sciences **102**: 14332–14337.

Béliveau, C., A. Cohen, D. Stewart, G. Periquet, A. Djoumad, *et al.*, 2015 Genomic and Proteomic Analyses Indicate that Banchine and Campoplegine Polydnaviruses Have Similar, if Not Identical, Viral Ancestors. Journal of Virology **89**: 8909–8921.

Bemm, F., C. L. Weiß, J. Schultz, and F. Förster, 2016 Genome of a tardigrade: Horizontal gene transfer or bacterial contamination? Proceedings of the National Academy of Sciences **113**: E3054–E3056.

Bézier, A., J. Herbinière, B. Lanzrein, and J.-M. Drezen, 2009 Polydnavirus hidden face: The genes producing virus particles of parasitic wasps. Journal of Invertebrate Pathology **101**: 194–203.

Biémont, C. and C. Vieira, 2006 Junk DNA as an evolutionary force. Nature **443**: 521–524.

Blackburn, D., 2005 Evolutionary origins of viviparity in fishes. In *Viviparity in Fishes*, pp. 303–317, New Life Publications.

Bongiovanni, L., A. Andriessen, M. H. M. Wauben, E. N. M. N.-t. Hoen, and A. de Bruin, 2021 Extracellular Vesicles: Novel Opportunities to Understand and Detect Neoplastic Diseases. Veterinary Pathology **58**: 453–471.

Bourque, G., K. H. Burns, M. Gehring, V. Gorbunova, A. Seluanov, *et al.*, 2018 Ten things you should know about transposable elements. Genome Biology **19**: 199.

Brady, S. G., B. L. Fisher, T. R. Schultz, and P. S. Ward, 2014 The rise of army ants and their relatives: Diversification of specialized predatory doryline ants. BMC Evolutionary Biology **14**: 93.

Buffalo, V., 2021 Quantifying the relationship between genetic diversity and population size suggests natural selection cannot explain Lewontin's Paradox. eLife **10**: e67509.

Burke, G. R., T. J. Simmonds, S. A. Thomas, and M. R. Strand, 2015 Microplitis demolitor Bracovirus Proviral Loci and Clustered Replication Genes Exhibit Distinct DNA Amplification Patterns during Replication. Journal of Virology **89**: 9511–9523.

Cai, J., G. Wu, P. A. Jose, and C. Zeng, 2016 Functional transferred DNA within extracellular vesicles. Experimental Cell Research **349**: 179–183.

Camargo, A. M., D. A. Andow, P. Castañera, and G. P. Farinós, 2018 First detection of a Sesamia nonagrioides resistance allele to Bt maize in Europe. Scientific Reports **8**: 3977.

Catoni, M., E. Noris, A. M. Vaira, T. Jonesman, S. Matić, *et al.*, 2018 Virus-mediated export of chromosomal DNA in plants. Nature Communications **9**: 5308.

Chebbi, M. A., T. Becking, B. Moumen, I. Giraud, C. Gilbert, *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**: 727–741.

Chevignon, G., G. Periquet, G. Gyapay, N. Vega-Czarny, K. Musset, *et al.*, 2018 Cotesia congregata Bracovirus Circles Encoding PTP and Ankyrin Genes Integrate into the DNA of Parasitized Manduca sexta Hemocytes. Journal of Virology **92**.

Chou, S., M. D. Daugherty, S. B. Peterson, J. Biboy, Y. Yang, *et al.*, 2015 Transferred interbacterial antagonism genes augment eukaryotic innate immune function. Nature **518**: 98–101.

Chuong, E. B., N. C. Elde, and C. Feschotte, 2017 Regulatory activities of transposable elements: From conflicts to benefits. Nature Reviews Genetics **18**: 71–86.

Coakley, G., R. M. Maizels, and A. H. Buck, 2015 Exosomes and Other Extracellular Vesicles: The New Communicators in Parasite Infections. Trends in Parasitology **31**: 477–489.

Cohen, S., S. Au, and N. Panté, 2011 How viruses access the nucleus. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research **1813**: 1634–1645.

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**: 186.

Crisp, A., C. Boschetti, M. Perry, A. Tunnacliffe, and G. Micklem, 2015 Expression of multiple horizontally acquired genes is a hallmark of both vertebrate and invertebrate genomes. Genome Biology **16**: 50.

Dagan, T., Y. Artzy-Randrup, and W. Martin, 2008 Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. Proceedings of the National Academy of Sciences **105**: 10039–10044.

Daniels, S. B., K. R. Peterson, L. D. Strausbaugh, M. G. Kidwell, and A. Chovnick, 1990 Evidence for horizontal transmission of the P transposable element between Drosophila species. Genetics **124**: 339–355.

Desjardins, C. A., D. E. Gundersen-Rindal, J. B. Hostetler, L. J. Tallon, D. W. Fadrosh, *et al.*, 2008 Comparative genomics of mutualistic viruses of Glyptapanteles parasitic wasps. Genome Biology **9**: R183.

Dotto, B. R., E. L. Carvalho, A. F. da Silva, F. Z. Dezordi, P. M. Pinto, *et al.*, 2018 HTT-DB: New features and updates. Database: The Journal of Biological Databases and Curation **2018**.

Dunning Hotopp, J. C., 2011 Horizontal gene transfer between bacteria and animals. Trends in genetics : TIG **27**: 157–163.

Eckwahl, M. J., H. Arnion, S. Kharytonchyk, T. Zang, P. D. Bieniasz, *et al.*, 2016 Analysis of the human immunodeficiency virus-1 RNA packageome. RNA **22**: 1228–1238.

Farinós, G. P., P. Hernández-Crespo, F. Ortego, and P. Castañera, 2018 Monitoring of Sesamia nonagrioides resistance to MON 810 maize in the European Union: Lessons from a long-term harmonized plan. Pest Management Science **74**: 557–568.

Feschotte, C. and C. Gilbert, 2012 Endogenous viruses: Insights into viral evolution and impact on host biology. Nature Reviews Genetics **13**: 283–296.

Flot, J.-F., B. Hespeels, X. Li, B. Noel, I. Arkhipova, *et al.*, 2013 Genomic evidence for ameiotic evolution in the bdelloid rotifer Adineta vaga. Nature **500**: 453–457.

Galbraith, J. D., A. J. Ludington, A. Suh, K. L. Sanders, and D. L. Adelson, 2020 New Environment, New Invaders—Repeated Horizontal Transfer of LINEs to Sea Snakes. Genome Biology and Evolution **12**: 2370–2383.

Gama Sosa, M. A., R. De Gasperi, and G. A. Elder, 2010 Animal transgenesis: An overview. Brain Structure and Function **214**: 91–109.

Gangadharan, S., L. Mularoni, J. Fain-Thornton, S. J. Wheelan, and N. L. Craig, 2010 DNA transposon Hermes inserts into DNA in nucleosome-free regions in vivo. Proceedings of the National Academy of Sciences **107**: 21966–21972.

Gasmi, L., H. Boulain, J. Gauthier, A. Hua-Van, K. Musset, *et al.*, 2015 Recurrent Domestication by Lepidoptera of Genes from Their Parasites Mediated by Bracoviruses. PLoS Genetics **11**.

Gasmi, L., A. K. Jakubowska, J. Ferré, M. Ogliastro, and S. Herrero, 2018 Characterization of two groups of Spodoptera exigua Hübner (Lepidoptera: Noctuidae) C-type lectins and insights into their role in defense against the densovirus JcDV. Archives of Insect Biochemistry and Physiology **97**: e21432.

Gasmi, L., E. Sieminska, S. Okuno, R. Ohta, C. Coutu, *et al.*, 2021 Horizontally transmitted parasitoid killing factor shapes insect defense to parasitoids. Science **373**: 535–541.

Gelvin, S. B., 2009 Agrobacterium in the Genomics Age. Plant Physiology **150**: 1665–1676.

Gilbert, C. and C. Belliardo, 2022 The diversity of endogenous viral elements in insects. Current Opinion in Insect Science **49**: 48–55.

Gilbert, C., A. Chateigner, L. Ernenwein, V. Barbe, A. Bézier, *et al.*, 2014 Population genomics supports baculoviruses as vectors of horizontal transfer of insect transposons. Nature Communications **5**: 3348.

Gilbert, C. and R. Cordaux, 2017 Viruses as vectors of horizontal transfer of genetic material in eukaryotes. Current Opinion in Virology **25**: 16–22.

Gilbert, C. and C. Feschotte, 2018 Horizontal acquisition of transposable elements and viral sequences: Patterns and consequences. Current Opinion in Genetics & Development **49**: 15–24.

Gilbert, C. and F. Maumus, 2022 Multiple Horizontal Acquisitions of Plant Genes in the Whitefly Bemisia tabaci. Genome Biology and Evolution **14**: evac141.

Gilbert, C., J. Peccoud, A. Chateigner, B. Moumen, R. Cordaux, *et al.*, 2016 Continuous Influx of Genetic Material from Host to Virus Populations. PLoS Genetics **12**.

Gilbert, C., S. Schaack, J. K. Pace, P. J. Brindley, and C. Feschotte, 2010 A role for host-parasite interactions in the horizontal transfer of DNA transposons across animal phyla. Nature **464**: 1347–1350.

Gilly, A., M. Etcheverry, M.-A. Madoui, J. Guy, L. Quadrana, *et al.*, 2014 TE-Tracker: Systematic identification of transposition events through whole-genome resequencing. BMC Bioinformatics **15**: 377.

Gladyshev, E. A., M. Meselson, and I. R. Arkhipova, 2008 Massive Horizontal Gene Transfer in Bdelloid Rotifers. Science **320**: 1210–1213.

Goic, B., K. A. Stapleford, L. Frangeul, A. J. Doucet, V. Gausson, *et al.*, 2016 Virus-derived DNA drives mosquito vector tolerance to arboviral infection. Nature Communications **7**: 12410.

Gozashti, L., S. W. Roy, B. Thornlow, A. Kramer, M. Ares, *et al.*, 2022 Transposable elements drive intron gain in diverse eukaryotes. Proceedings of the National Academy of Sciences **119**: e2209766119.

Graham, L. A. and P. L. Davies, 2021 Horizontal Gene Transfer in Vertebrates: A Fishy Tale. Trends in Genetics **37**: 501–503.

Guinet, B., D. Lepetit, S. Charlat, P. N. Buhl, D. G. Notton, *et al.*, 2023 Endoparasitoid lifestyle promotes endogenization and domestication of dsDNA viruses. eLife **12**: e85993.

Gundersen-Rindal, D. E. and D. E. Lynn, 2003 Polydnavirus integration in lepidopteran host cells in vitro. Journal of Insect Physiology **49**: 453–462.

Gyles, C. and P. Boerlin, 2014 Horizontally transferred genetic elements and their role in pathogenesis of bacterial disease. Veterinary Pathology **51**: 328–340.

Hayward, A. and C. Gilbert, 2022 Transposable elements. Current Biology **32**: R904–R909.

Heath, B. D., R. D. J. Butcher, W. G. F. Whitfield, and S. F. Hubbard, 1999 Horizontal transfer of Wolbachia between phylogenetically distant insect species by a naturally occurring mechanism. Current Biology **9**: 313–316.

Heringer, P. and G. C. S. Kuhn, 2022 Multiple horizontal transfers of a Helitron transposon associated with a parasitoid wasp. Mobile DNA **13**: 20.

Herniou, E. A., E. Huguet, J. Thézé, A. Bézier, G. Periquet, *et al.*, 2013 When parasitic wasps hijacked viruses: Genomic and functional evolution of polydnaviruses. Philosophical Transactions of the Royal Society B: Biological Sciences **368**: 20130051.

Hilgenboecker, K., P. Hammerstein, P. Schlattmann, A. Telschow, and J. H. Werren, 2008 How many species are infected with Wolbachia? – a statistical analysis of current data. FEMS Microbiology Letters **281**: 215–220.

Holzerlandt, R., C. Orengo, P. Kellam, and M. M. Albà, 2002 Identification of New Herpesvirus Gene Homologs in the Human Genome. Genome Research **12**: 1739–1748.

Hotopp, J. C. D., M. E. Clark, D. C. S. G. Oliveira, J. M. Foster, P. Fischer, *et al.*, 2007 Widespread Lateral Gene Transfer from Intracellular Bacteria to Multicellular Eukaryotes. Science **317**: 1753–1756.

Huang, C. R. L., K. H. Burns, and J. D. Boeke, 2012 Active Transposition in Genomes. Annual Review of Genetics **46**: 651–675.

Husnik, F. and J. P. McCutcheon, 2018 Functional horizontal gene transfer from bacteria to eukaryotes. Nature Reviews Microbiology **16**: 67–79.

Intrieri, M. C. and M. Buiatti, 2001 The Horizontal Transfer of Agrobacterium rhizogenes Genes and the Evolution of the Genus Nicotiana. Molecular Phylogenetics and Evolution **20**: 100–110.

Irwin, N. A. T., A. A. Pittis, T. A. Richards, and P. J. Keeling, 2022 Systematic evaluation of horizontal gene transfer between eukaryotes and viruses. Nature Microbiology **7**: 327–336.

Kaiser, L., J. Fernandez-Triana, C. Capdevielle-Dulac, C. Chantre, M. Bodet, *et al.*, 2017 Systematics and biology of Cotesia typhae sp. n. (Hymenoptera, Braconidae, Microgastrinae), a potential biological control agent against the noctuid Mediterranean corn borer, Sesamia nonagrioides. ZooKeys pp. 105–136.

Kalluraya, C. A., A. J. Weitzel, B. V. Tsu, and M. D. Daugherty, 2023 Bacterial origin of a key innovation in the evolution of the vertebrate eye. Proceedings of the National Academy of Sciences **120**: e2214815120.

Kambayashi, C., R. Kakehashi, Y. Sato, H. Mizuno, H. Tanabe, *et al.*, 2022 Geography-Dependent Horizontal Gene Transfer from Vertebrate Predators to Their Prey. Molecular Biology and Evolution **39**: msac052.

Kapusta, A., A. Suh, and C. Feschotte, 2017 Dynamics of genome size evolution in birds and mammals. Proceedings of the National Academy of Sciences **114**: E1460–E1469.

Kawamura, Y., A. Sanchez Calle, Y. Yamamoto, T.-A. Sato, and T. Ochiya, 2019 Extracellular vesicles mediate the horizontal transfer of an active LINE-1 retrotransposon. Journal of Extracellular Vesicles **8**: 1643214.

Kay, E., T. M. Vogel, F. Bertolla, R. Nalin, and P. Simonet, 2002 In Situ Transfer of Antibiotic Resistance Genes from Transgenic (Transplastomic) Tobacco Plants to Bacteria. Applied and Environmental Microbiology **68**: 3345–3351.

Kofler, R., T. Hill, V. Nolte, A. J. Betancourt, and C. Schlötterer, 2015 The recent invasion of natural Drosophila simulans populations by the P-element. Proceedings of the National Academy of Sciences of the United States of America **112**: 6659–6663.

Koutsovoulos, G. D., S. G. Noriot, M. Bailly-Bechet, E. G. J. Danchin, and C. Rancurel, 2022 AvP: A software package for automatic phylogenetic detection of candidate horizontal gene transfers. PLOS Computational Biology **18**: e1010686.

Lacroix, B. and V. Citovsky, 2016 Transfer of DNA from Bacteria to Eukaryotes. mBio **7**: e00863–16.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, *et al.*, 2001 Initial sequencing and analysis of the human genome. Nature **409**: 860–921.

Lawrence, J. G. and H. Ochman, 1997 Amelioration of Bacterial Genomes: Rates of Change and Exchange. Journal of Molecular Evolution **44**: 383–397.

Le Rouzic, A., T. S. Boutin, and P. Capy, 2007 Long-term evolution of transposable elements. Proceedings of the National Academy of Sciences **104**: 19375–19380.

Le Rouzic, A. and P. Capy, 2005 The First Steps of Transposable Elements Invasion: Parasitic Strategy vs. Genetic Drift. Genetics **169**: 1033–1043.

Lechardeur, D., K.-J. Sohn, M. Haardt, P. B. Joshi, M. Monck, *et al.*, 1999 Metabolic instability of plasmid DNA in the cytosol: A potential barrier to gene transfer. Gene Therapy **6**: 482–497.

Leclercq, S., J. Thézé, M. A. Chebbi, I. Giraud, B. Moumen, *et al.*, 2016 Birth of a W sex chromosome by horizontal transfer of Wolbachia bacterial symbiont genome. Proceedings of the National Academy of Sciences **113**: 15036–15041.

Legeai, F., B. F. Santos, S. Robin, A. Bretaudeau, R. B. Dikow, *et al.*, 2020 Genomic architecture of endogenous ichnoviruses reveals distinct evolutionary pathways leading to virus domestication in parasitic wasps. BMC Biology **18**: 89.

Legendre, M., A. Lartigue, L. Bertaux, S. Jeudy, J. Bartoli, *et al.*, 2015 In-depth study of Mollivirus sibericum, a new 30,000-y-old giant virus infecting Acanthamoeba. Proceedings of the National Academy of Sciences **112**: E5327–E5335.

Lelio, I. D., A. Illiano, F. Astarita, L. Gianfranceschi, D. Horner, *et al.*, 2019 Evolution of an insect immune barrier through horizontal gene transfer mediated by a parasitic wasp. PLOS Genetics **15**: e1007998.

Li, C.-X., M. Shi, J.-H. Tian, X.-D. Lin, Y.-J. Kang, *et al.*, 2015 Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. eLife **4**: e05378.

Li, Y., Z. Liu, C. Liu, Z. Shi, L. Pang, *et al.*, 2022 HGT is widespread in insects and contributes to male courtship in lepidopterans. Cell **185**: 2975–2987.e10.

Loiseau, V., J. Peccoud, C. Bouzar, S. Guillier, J. Fan, *et al.*, 2021 Monitoring Insect Transposable Elements in Large Double-Stranded DNA Viruses Reveals Host-to-Virus and Virus-to-Virus Transposition. Molecular Biology and Evolution **38**: 3512–3530.

Loreto, E. L. S., C. M. A. Carareto, and P. Capy, 2008 Revisiting horizontal transfer of transposable elements in Drosophila. Heredity **100**: 545–554.

Martin, W. F., 2017 Too Much Eukaryote LGT. BioEssays **39**: 1700115.

Matveeva, T. V., D. I. Bogomaz, O. A. Pavlova, E. W. Nester, and L. A. Lutova, 2012 Horizontal Gene Transfer from Genus Agrobacterium to the Plant Linaria in Nature. Molecular Plant-Microbe Interactions®**25**: 1542–1551.

McDaniel, L. D., E. Young, J. Delaney, F. Ruhnau, K. B. Ritchie, *et al.*, 2010 High Frequency of Horizontal Gene Transfer in the Oceans. Science **330**: 50–50.

McKelvey, T. A., D. E. Lynn, D. Gundersen-Rindal, D. Guzo, D. A. Stoltz, *et al.*, 1996 Transformation of Gypsy Moth (Lymantria dispar) Cell Lines by Infection withGlyptapanteles indiensisPolydnavirus. Biochemical and Biophysical Research Communications **225**: 764–770.

Meng, G., Y. Li, C. Yang, and S. Liu, 2019 MitoZ: A toolkit for animal mitochondrial genome assembly, annotation and visualization. Nucleic Acids Research **47**: e63.

Mérel, V., M. Boulesteix, M. Fablet, and C. Vieira, 2020 Transposable elements in Drosophila. Mobile DNA **11**: 23.

Metegnier, G., T. Becking, M. A. Chebbi, I. Giraud, B. Moumen, *et al.*, 2015 Comparative paleovirological analysis of crustaceans identifies multiple widespread viral groups. Mobile DNA **6**: 16.

Metzger, M. J., A. N. Paynter, M. E. Siddall, and S. P. Goff, 2018 Horizontal transfer of retrotransposons between bivalves and other aquatic species of multiple phyla. Proceedings of the National Academy of Sciences **115**: E4227–E4235.

Miyao, A., M. Nakagome, T. Ohnuma, H. Yamagata, H. Kanamori, *et al.*, 2012 Molecular Spectrum of Somaclonal Variation in Regenerated Rice Revealed by Whole-Genome Sequencing. Plant and Cell Physiology **53**: 256–264.

Moran, N. A. and T. Jarvik, 2010 Lateral Transfer of Genes from Fungi Underlies Carotenoid Production in Aphids. Science **328**: 624–627.

MOYAL, PASCAL., PATRICE. TOKRO, AHMET. BAYRAM, MATILDA. SAVOPOULOU-SOULTANI, ERIC. CONTI, *et al.*, 2011 Origin and taxonomic status of the Palearctic population of the stem borer Sesamia nonagrioides (Lefèbvre) (Lepidoptera: Noctuidae). Biological Journal of the Linnean Society **103**: 904–922.

Mu, J., X. Zhuang, Q. Wang, H. Jiang, Z.-B. Deng, *et al.*, 2014 Interspecies communication between plant and mouse gut host cells through edible plant derived exosome-like nanoparticles. Molecular Nutrition & Food Research **58**: 1561–1573.

Narayanan Kutty, S., K. Meusemann, K. M. Bayless, M. A. T. Marinho, A. C. Pont, *et al.*, 2019 Phylogenomic analysis of Calyptratae: Resolving the phylogenetic relationships within a major radiation of Diptera. Cladistics **35**: 605–622.

Ono, R., Y. Yasuhiko, K.-i. Aisaki, S. Kitajima, J. Kanno, *et al.*, 2019 Exosome-mediated horizontal gene transfer occurs in double-strand break repair during genome editing. Communications Biology **2**: 1–8.

Palazzo, A., P. Lorusso, C. Miskey, O. Walisko, A. Gerbino, *et al.*, 2019 Transcriptionally promiscuous "blurry" promoters in Tc1/mariner transposons allow transcription in distantly related genomes. Mobile DNA **10**: 13.

Peccoud, J., R. Cordaux, and C. Gilbert, 2018 Analyzing Horizontal Transfer of Transposable Elements on a Large Scale: Challenges and Prospects. BioEssays **40**: 1700177.

Peccoud, J., V. Loiseau, R. Cordaux, and C. Gilbert, 2017 Massive horizontal transfer of transposable elements in insects. Proceedings of the National Academy of Sciences **114**: 4721–4726.

Peters, R. S., L. Krogmann, C. Mayer, A. Donath, S. Gunkel, *et al.*, 2017 Evolutionary History of the Hymenoptera. Current Biology **27**: 1013–1018.

Pisani, D., J. A. Cotton, and J. O. McInerney, 2007 Supertrees Disentangle the Chimerical Origin of Eukaryotic Genomes. Molecular Biology and Evolution **24**: 1752–1760.

Quispe-Huamanquispe, D. G., G. Gheysen, and J. F. Kreuze, 2017 Horizontal Gene Transfer Contributes to Plant Evolution: The Case of Agrobacterium T-DNAs. Frontiers in Plant Science **8**.

Ravenhall, M., N. Škunca, F. Lassalle, and C. Dessimoz, 2015 Inferring Horizontal Gene Transfer. PLOS Computational Biology **11**: e1004095.

Rodrigues, M. L., L. Nimrichter, D. L. Oliveira, J. D. Nosanchuk, and A. Casadevall, 2008 Vesicular Trans-Cell Wall Transport in Fungi: A Mechanism for the Delivery of Virulence-Associated Macromolecules? Lipid Insights **2**: 27–40.

Salzberg, S. L., 2017 Horizontal gene transfer is not a hallmark of the human genome. Genome Biology **18**: 85.

Samuel, M., M. Bleackley, M. Anderson, and S. Mathivanan, 2015 Extracellular vesicles including exosomes in cross kingdom regulation: A viewpoint from plant-fungal interactions. Frontiers in Plant Science **6**: 766.

Schaack, S., C. Gilbert, and C. Feschotte, 2010 Promiscuous DNA: Horizontal transfer of transposable elements and why it matters for eukaryotic evolution. Trends in Ecology & Evolution **25**: 537–546.

Schnable, P. S., D. Ware, R. S. Fulton, J. C. Stein, F. Wei, *et al.*, 2009 The B73 maize genome: Complexity, diversity, and dynamics. Science (New York, N.Y.) **326**: 1112–1115.

Schneider, S. E. and J. H. Thomas, 2014 Accidental Genetic Engineers: Horizontal Sequence Transfer from Parasitoid Wasps to Their Lepidopteran Hosts. PLOS ONE **9**: e109446.

Schönknecht, G., W.-H. Chen, C. M. Ternes, G. G. Barbier, R. P. Shrestha, *et al.*, 2013 Gene Transfer from Bacteria and Archaea Facilitated Evolution of an Extremophilic Eukaryote. Science **339**: 1207–1210.

Shahid, S. and R. K. Slotkin, 2020 The current revolution in transposable element biology enabled by long reads. Current Opinion in Plant Biology **54**: 49–56.

Sibley, L. D., 2004 Intracellular Parasite Invasion Strategies. Science **304**: 248–253.

Song, S. U., T. Gerasimova, M. Kurkulos, J. D. Boeke, and V. G. Corces, 1994 An env-like protein encoded by a Drosophila retroelement: Evidence that gypsy is an infectious retrovirus. Genes & Development **8**: 2046–2057.

Steglich, C. and S. W. Schaeffer, 2006 The ornithine decarboxylase gene of Trypanosoma brucei: Evidence for horizontal gene transfer from a vertebrate source. Infection, Genetics and Evolution **6**: 205–219.

Steinegger, M. and J. Söding, 2017 MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nature Biotechnology **35**: 1026–1028.

Syvanen, M., 2012 Evolutionary Implications of Horizontal Gene Transfer. Annual Review of Genetics **46**: 341–358.

Thézé, J., S. Leclercq, B. Moumen, R. Cordaux, and C. Gilbert, 2014 Remarkable Diversity of Endogenous Viruses in a Crustacean Genome. Genome Biology and Evolution **6**: 2129–2140.

Thézé, J., J. Takatsuka, M. Nakai, B. Arif, and E. A. Herniou, 2015 Gene acquisition convergence between entomopoxviruses and baculoviruses. Viruses **7**: 1960–1974.

Thomas, C. M. and K. M. Nielsen, 2005 Mechanisms of, and Barriers to, Horizontal Gene Transfer between Bacteria. Nature Reviews Microbiology **3**: 711–721.

Treangen, T. J. and E. P. C. Rocha, 2011 Horizontal Transfer, Not Duplication, Drives the Expansion of Protein Families in Prokaryotes. PLOS Genetics **7**: e1001284.

Tzfira, T. and V. Citovsky, 2006 Agrobacterium-mediated genetic transformation of plants: Biology and biotechnology. Current Opinion in Biotechnology **17**: 147–154.

van der Kuyl, A. C. and B. Berkhout, 2020 Viruses in the reproductive tract: On their way to the germ line? Virus Research **286**: 198101.

Van Etten, J. and D. Bhattacharya, 2020 Horizontal Gene Transfer in Eukaryotes: Not if, but How Much? Trends in Genetics **36**: 915–925.

Venner, S., V. Miele, C. Terzian, C. Biémont, V. Daubin, *et al.*, 2017 Ecological networks to unravel the routes to horizontal transposon transfers. PLOS Biology **15**: e2001536.

Vetsigian, K., C. Woese, and N. Goldenfeld, 2006 Collective evolution and the genetic code. Proceedings of the National Academy of Sciences **103**: 10696–10701.

Volkoff, A.-N., V. Jouan, S. Urbach, S. Samain, M. Bergoin, *et al.*, 2010 Analysis of Virion Structural Components Reveals Vestiges of the Ancestral Ichnovirus Genome. PLOS Pathogens **6**: e1000923.

Wagner, A., R. J. Whitaker, D. J. Krause, J.-H. Heilers, M. van Wolferen, *et al.*, 2017 Mechanisms of gene flow in archaea. Nature Reviews. Microbiology **15**: 492–501.

Wallau, G. L., P. Capy, E. Loreto, A. Le Rouzic, and A. Hua-Van, 2016 VHICA, a New Method to Discriminate between Vertical and Horizontal Transposon Transfer: Application to the Mariner Family within Drosophila. Molecular Biology and Evolution **33**: 1094–1109.

Wallau, G. L., C. Vieira, and É. L. S. Loreto, 2018 Genetic exchange in eukaryotes through horizontal transfer: Connected by the mobilome. Mobile DNA **9**: 6.

Wang, X. and X. Liu, 2016 Close ecological relationship among species facilitated horizontal transfer of retrotransposons. BMC Evolutionary Biology **16**: 201.

Wang, Z., X. Ye, Y. Zhou, X. Wu, R. Hu, *et al.*, 2021a Bracoviruses recruit host integrases for their integration into caterpillar's genome. PLOS Genetics **17**: e1009751.

Wang, Z.-h., Y.-n. Zhou, J. Yang, X.-q. Ye, M. Shi, *et al.*, 2021b Genome-Wide Profiling of Diadegma semiclausum Ichnovirus Integration in Parasitized Plutella xylostella Hemocytes Identifies Host Integration Motifs and Insertion Sites. Frontiers in Microbiology **11**.

Werren, J. H., L. Baldo, and M. E. Clark, 2008 Wolbachia: Master manipulators of invertebrate biology. Nature Reviews Microbiology **6**: 741–751.

Werren, J. H., W. Zhang, and L. R. Guo, 1997 Evolution and phylogeny of Wolbachia: Reproductive parasites of arthropods. Proceedings of the Royal Society of London. Series B: Biological Sciences **261**: 55–63.

White, F. F., D. J. Garfinkel, G. A. Huffman, M. P. Gordon, and E. W. Nester, 1983 Sequences homologous to Agrobacterium rhizogenes T-DNA in the genomes of uninfected plants. Nature **301**: 348–350.

Whitfield, Z. J., P. T. Dolan, M. Kunitomi, M. Tassetto, M. G. Seetin, *et al.*, 2017 The Diversity, Structure, and Function of Heritable Adaptive Immunity Sequences in the Aedes aegypti Genome. Current biology: CB **27**: 3511–3519.e7.

Whittaker, G. R., M. Kann, and A. Helenius, 2000 Viral entry into the nucleus. Annual Review of Cell and Developmental Biology **16**: 627–651.

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

Wybouw, N., W. Dermauw, L. Tirry, C. Stevens, M. Grbić, *et al.*, 2014 A gene horizontally transferred from bacteria protects arthropods from host plant cyanide poisoning. eLife **3**: e02365.

Xia, J., Z. Guo, Z. Yang, H. Han, S. Wang, *et al.*, 2021 Whitefly hijacks a plant detoxification gene that neutralizes plant toxins. Cell **184**: 1693–1705.e17.

Yoth, M., S. Jensen, and E. Brasset, 2022 The Intricate Evolutionary Balance between Transposable Elements and Their Host: Who Will Kick at Goal and Convert the Next Try? Biology **11**: 710.

Zhang, H.-H., J. Peccoud, M.-R.-X. Xu, X.-G. Zhang, and C. Gilbert, 2020 Horizontal transfer and evolution of transposable elements in vertebrates. Nature Communications **11**: 1362.

Zhou, S., B. Liu, Y. Han, Y. Wang, L. Chen, *et al.*, 2022 ZOVER: The database of zoonotic and vector-borne viruses. Nucleic Acids Research **50**: D943–D949.

Zimin, A. V., G. Marçais, D. Puiu, M. Roberts, S. L. Salzberg, *et al.*, 2013 The MaSuRCA genome assembler. Bioinformatics **29**: 2669–2677.