Genome-wide patterns of bracovirus chromosomal integration into multiple host tissues during parasitism

Héloïse Muller1, Mohamed Amine Chebbi2,3, Clémence Bouzar1, George Péritquet2, T’aiadjana Fortuna1, Paul-André Calatayud1,4, Bruno Le Ru1,4, Julius Obonyo4, Laure Kaiser1, Jean-Michel Drezen2, Elisabeth Huguet2##, Clément Gilbert1##

1 Université Paris-Saclay, CNRS, IRD, UMR Évolution, Génomes, Comportement et Écologie, 91198, 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 ViroScan3D SAS, Lyon 69008, France
4 icipe (International Center of Insect Physiology and Ecology), IRD team, P.O. Box 30772-00100, Nairobi, Kenya

#Elisabeth Huguet and Clément Gilbert are equal senior authors.

##Corresponding authors: Clément Gilbert, clement.gilbert@egce.cnrs-gif.fr, and Elisabeth Huguet, elisabeth.huguet@univ-tours.fr.

Copyright © 2021 Muller et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license.
Abstract

Bracoviruses are domesticated viruses found in parasitic wasp genomes. They are composed of genes of nudiviral origin involved in particle production and proviral segments encoding virulence genes 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 together with their eggs into host larvae. Bracovirus circles of two wasp species were reported to undergo chromosomal integration in parasitized host hemocytes, through a conserved sequence named 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) seven 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 has occurred at least six times accidentally in the wasp germline 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 non-targeted 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 host integration motif (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 interaction, our study supports HIM-mediated chromosomal integration of bracovirus as a possible source of wasp-to-host horizontal transfer with long term evolutionary consequences.
Introduction

Polydnavirus genomes within parasitoid wasps (Hymenoptera) are composed of domesticated viral genes and genes of different origin involved in virulence (1–3). The domesticated viruses encode viral particles akin to gene transfer agents that are injected during oviposition into the lepidopteran hosts of parasitoid wasps and are necessary for successful development of the wasp larvae. Polydnaviruses result from endogenization events of large double stranded DNA viruses that took place during the course of parasitoid wasp evolution (1, 4–12). Polydnaviruses of the ichnovirus genus (IV) identified in the genomes of certain Ichneumonidae Campopleginae and Banchinae wasps, are believed 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 (Myrs) ago in the ancestor of the microgastroid complex of braconid wasps (7, 13, 14), a hyper-diversified 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 during 100 Myrs. All microgastroid species studied so far express nudivirus-derived bracovirus genes in specialized cells located in a region of the ovaries named “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 (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 “macrolocus” spanning two megabases and including two-thirds of proviral segments. Chromosome-scale assembly of the genome of Cotesia congregata (Microgastrinae) revealed that it contains 10 proviral loci (PL), each made of one to eighteen proviral segments (PL2: 18 segments (3)). Comparisons between Cotesia species and Microplitis demolitor, showed that the synteny of these
proviral loci is well-conserved along the phylogeny of Braconidae since ~53 million years of evolution, suggesting strong evolutionary constraints associated to the function of the segments.

Proviral loci 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 (DRJ) 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 in 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 under 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 two circles in the hemocytes of the host of the Microplitis demolitor wasp (Microgastrinae) via a motif called Host Integration Motif (HIM) conserved in all MdBV circles (26). Another study using primer extension capture followed by high-throughput sequencing unveiled several thousands of chromosomal integrations for eight circles of Cotesia congregata into the hemocytes of its host, the tobacco hornworm Manduca sexta (30). The eight 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 Junction1 (J1) and Junction2 (J2), located within the HIM. J1 and J2 correspond to sequences which 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 has also been recently described for ichnoviruses. This suggests that this phenomenon plays an important role in parasitism success of parasitoid wasps and 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 PDV circle persistence and integration in the host-parasitoid system involving the Cotesia typhae wasp (Microgastrinae, Braconidae) and its natural host, the Mediterranean corn borer (Sesamia nonagrioides, Noctuidae). Cotesia typhae is a recently described species among the Cotesia flavipes species-complex and it is native to sub-Saharan eastern 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 and as such, 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 high number of circle sequences (at least six) 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 in M. sexta parasitized by C. congregata (30). We further uncovered that integration is not required for the persistence of circles during parasitism as the quantity of non-integrated circles is similar to that of most integrated circles in all host tissues seven days after parasitism. Interestingly, high levels of bracovirus integration were also detected in the host’s genome even when parasitism failed.

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 which has been reared in the EGCE laboratory (Gif-sur-Yvette, France) since 2015, from a strain reared at the icipe (Nairobi, Kenya) since 2013 when it was initially collected from the Kobodo locality in Kenya (0.679S,
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 the Buffer Set IV from Macherey Nagel, following the manufacturer's protocol. We obtained 26 ug of DNA, quantified with QuBit (ThermoFisher Scientific). The integrity of DNA was checked on an agarose gel, and a nanodrop measure was performed to confirm the absence of proteins and other contaminants. For whole *C. typhae* genome sequencing, we sub-contracted national sequencing center (Genoscope, Evry) to prepare two types of DNA libraries according to the requirements of Illumina and Oxford Nanopore Technologies (ONT). 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 (Table S1). Paired-end reads were clipped for adapters and low-quality bases and then merged into single reads using the BBMerge tool (34). For the Nanopore sequencing, preparation of libraries was carried out with a 1D genomic DNA ligation protocol (SQK-LSK109, ONT) and sequenced using R9.4.1 flowcells on both MinION and PromethION sequencers (ONT) (Table S1).

**Assembly of the *C. typhae* genome**

The genome size was first estimated from a preliminary assembly obtained from Illumina reads with ABYSS 2.0 (35) using a k-mer length of 96. The genome assembly was then performed de novo with Flye 2.5 (36) using 30X longest ONT reads (Table S1). The resulting Nanopore assembly was polished using Racon v1.4.7 (37) after mapping 100X longest raw ONT reads (Table S1) with Minimap2 v2.17-r941 (38) and then Pilon v.1.23 (39) using the merged Illumina reads mapped with BBMap v.37.62 (40). The completeness of the genome assembly was assessed by searching for similarities against highly conserved genes among insects. For this purpose, we ran BUSCO (version 3.0.1) in “genome mode” specifying a profile library of 1658 single-copy core genes (April 2019 release) (41). Finally, scaffolds were checked for potential contamination by sequences from other organisms by visualizing them with Blobtools (version 1,1,1) using taxon-annotated-GC-coverage plots. Blobtools assigns
scaffolds to taxonomic ranks depending on their homologies, using both blastn (here nt NCBI database downloaded in November 2019) and blastx (here 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 assignation.

**Annotation of the *C. typhae* genome**

Transposable Elements (TEs) were *de novo* identified and annotated in genomic sequences using respectively TEdenovo and TEannot pipelines included in the REPET package v2.5 (42, 43). To construct a *de novo* repeat library, repeats were first screened using Recon (44), Grouper (43) and Piler (45). Consensus repeats were then classified into families using PASTEClassifier and filtered out for all potential wasp genes corresponding to multigenic families. The TE library built by Tedenovo (42, 43) was then applied to perform a homology-based repeat search in the genome using Teannot (42, 43).

Gene annotation was then performed on the repeat-masked assembly by running two iterations of Maker (version 2.31.10) (46). The first iteration of Maker used alignments of *C. vestalis* transcriptome assembly (est2genome=1) and both reviewed Hexapod and Polydnaviridae UniProt-SwissProt proteins (release January 2020) (protein2genome=1) as sources of evidence for homology-based gene prediction. The resulting gene prediction was then used to train SNAP (version 2006-07-28) (47) and AUGUSTUS (version 3.3.2) (48) in order to construct *ab-initio* gene models. The second run of Maker allowed refining all these gene models in a GFF3 output file. Predicted genes were functionally annotated with InterProScan (version 5.39-77.0) (49) using the PfamA database (version 32.0) (43) and BlastP (version 2.7.1+) (50) using the UniProt-SwissProt database (release january 2020). Finally, obtained functional annotations were integrated in the final GFF3 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, 51). 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, which is no longer functional (ps34 for pseudo segment 34) (20, 52). *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 of seven 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* (53) were aligned on *C. typhae*’s genome using blastn to identify genes of each segment. DRJs of each *C. congregata* segment (File S1) were then aligned using blastn on each homologous candidate segments in *C. typhae* to determine precisely the segment coordinates. The coding regions and DRJ of ten segments present in *C. congregata* but not in *C. sesamiae* (Segments S37new in Gauthier et al. (3), S3, S9, S19, S22, S29 and S31 in the macrolocus, plus S10, S11, S21 and ps34 in dispersed loci (52)) were also aligned on *C. typhae*’s genome. The synteny between segments and some other genes flanking the segments also helped to solve 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 chromosomal scale genome available of *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: (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 HIMs/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 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, so we filtered such cases with an 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-duplications for five segments: 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 for which the DRJ5’ and DRJ3’ of a same segment align next to each other (and not at the extremities of the segments, contrary to proviral segments), as expected for HIM-mediated integrations (30).

Sequencing S. nonagrioides larvae parasitized by C. typhae

C. typhae individuals used for this experiment were taken from the strain of Kobodo coming from icipe rearing (see section “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 of the southwest of 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 into its host, generally
ranging between 70 and 110 (54). Larval development typically takes about 14 days in laboratory conditions until wasp larvae emerge from their host and pupate (32). 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 five millimeters 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 respectively six, three, nine and one 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 PBS. DNA was then extracted from a pool of each tissue (except fat body) using the DNeasy Blood & Tissue Kit (Qiagen). We also extracted DNA from one of the six whole larvae in which we were unable to find any wasp embryo. We sub-contracted Novogen to build a paired-end library (2 x 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 in Muller et al. (2021) (55) and available on NCBI (Accession number: JADWQK000000000). In brief, the genome was assembled using short Illumina and long Oxford Nanopore Technology reads using the MaSurCA assembler (56) followed by a run of the purge_dup pipeline (57) to remove scaffolds with low coverage, partial overlaps and haplotigs. The resulting assembly is made of 2,253 scaffolds with an N50 of 1,105 kbp and a total size of 1,021 Mbp. It contains 96% of Lepidoptera BUSCO genes, 2.7% of which are duplicated (55).
Adapters were removed and reads were quality-trimmed with Trimmomatic v0.38 (options LEADING:20, TRAILING:20, SLIDINGWINDOW:4:15 and MINLEN:36) (58). Raw and trimmed read quality was assessed using FastQC v0.11.8 (59). To obtain statistics on sequencing depth, we aligned trimmed paired-end reads from the five samples using Bowtie2 v2.3.4.2 in end-to-end mode separately on the wasp and the moth genomes (60). 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 C. typhae bracovirus 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 onto C. typhae’s 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 onto S. nonagrioides’s 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 S. nonagrioides’s genome. For that, we searched for sequencing reads for which a portion aligns on S. nonagrioides genome only and the other portion aligns on C. typhae BV proviral segments only. Such chimeric reads were identified using an R pipeline previously used to identify recombination events within a single genome that we slightly adapted to our study (61). 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 on neither reference genomes, (iii) no more than 20 bases can align simultaneously on both reference genomes and (iv) no more than five 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 a microhomology between both CtBV DNA circles and host genome at the integration point, whereas the separation should correspond to non-templated addition of nucleotides at this integration point (61, 62). To check whether microhomology lengths at integration points were consistent with those expected by chance, we simulated expected distributions following the approach described in Peccoud et al. (61). Briefly, considering the sequences of the CtBV circles and *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 *S. nonagrioides*’s 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 blasted against the sequences from which they were generated and the blast outputs were submitted to the same analysis as that performed on real data.

### Localization of chimeric reads in *C. typhae* bracovirus circles

Chimeric reads mapping to *C. typhae* bracovirus circles were assigned to three categories depending on the position of the wasp-host junction: (i) chimeric reads for which the CtBV-host junction falls within HIMs, (ii) those for which the junction falls in circles devoid of HIMs or outside of HIMs in circles containing HIMs and (iii) those 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 position supported by the most chimeric reads, plus the positions around this point until a position is supported by less than two 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 chimera falling outside HIMs.

### Data availability

The assembly and annotation of *C. typhae*’s genome are available in GenBank under the accession number JAAOIC000000000.2 and in BIPAA at https://bipaa.genouest.org/. The raw sequencing reads of the five samples of *S. nonagrioides* parasitized by *C. typhae* are available in NCBI under BioProject number PRJNA718433.

### Results

#### Assembly and annotation of *C. typhae*’s genome

The genome of *C. typhae* was sequenced at about 45X depth by short paired-end reads (Illumina) and 350X depth by long reads (ONT) (Table S1). 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 (Table S2). This assembly was made of only 72 scaffolds and had a N50 size of 6.81 Mb (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 (54).

The completeness of the assembly was assessed using BUSCO (41). The results revealed that 1,639 of 1,658 (98.9%) conserved insect genes were present in the final assembly (Table S3). Assembly visualization by Blobtools (63) using taxon-annotated-GC-coverage plots showed a majority assignation to Polydnaviridae family (131 Mb), which is due to the presence of bracovirus sequences dispersed in the wasp genome (3): the majority of large scaffolds are identified as containing a bracovirus sequence (Figure S1). Our automatic annotation revealed that 58.6% of the *C. typhae* genome is made of transposable elements. The most numerous TEs are LARD and TIR elements,
which represent respectively 35 and 28% of the classified TEs (Figure S2). We automatically annotated a total of 8,591 genes in the genome of *C. typhae* (Table S4). More than 90% of the predicted genes had over 50% of their exons supported by RNAseq from the species *C. vestalis*, the closest species for which RNAseq data were available. Genes exhibited a mean of five exons per transcript (Table S4).

The joint functional annotation procedure with InterProScan (49) and BlastP (50) 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 among which =12,000 validated by *C. congregata* RNAseq data) probably in part because of the divergence between *C. typhae* and *C. vestalis* RNA sequences used for annotation.

The annotated genome sequence of *C. typhae* is available in DDBJ/ENA/GenBank under accession number JAAOIC00000000.

**Annotation of *C. typhae* bracovirus proviral segments**

In order to annotate the proviral segments of *C. typhae*, we used the 26 segments of *C. sesamiae*, and the ten segments specific to *C. congregata*, as queries to perform similarity searches on the *C. typhae* genome. 27 out of a total of 38 segments described in *C. congregata* and/or *C. sesamiae* were clearly identified in *C. typhae* (Figure 1 and Table S5). Among the 11 segments not found in *C. typhae*, nine are specific to *C. congregata*, which means that they are not present in *C. sesamiae* either. The segment S13, 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 three segments (S10, S11 and S19) present in *C. congregata* but not found so far in *C. sesamiae*. In total, we annotated 27 proviral segments in *C. typhae* (Figure 1 and 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 proviral loci (PL): PL1 and PL2 (17, 20). The nine 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 this proviral locus is specific to C. congregata and recent. While PL9 is present in both C. congregata (two proviral segments) and C. sesamiae (one proviral segment), we did not find it in C. typhae.

**Proviral segment characteristics and HIM identification**

HIMs were previously described in 12 segments out of 36 in C. congregata (30). These HIMs were used here as queries to perform blastn searches on C. typhae segments (File S2). HIMs were found at their expected homologous loci in C. typhae for all but one segment, S15. The lack of HIM in S15 is likely due to the fact that this segment is undergoing degradation in C. typhae. Indeed, contrary to C. congregata in which S15 is 8700 pb and contains five 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 HIM identified in C. typhae against the five 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 these additional segments (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 (File S2).

Interestingly, our annotation identified seven other bracoviral sequences dispersed in the wasp genome. They share high similarities with viral sequences, but do not follow the organization of proviral segments having DRJs at their extremities. Without DRJs at their extremities, these segments cannot
form circles and they are thus not functional (see section “Persistence of non-integrated circles seven
days post-oviposition” 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 (Figure 1).
Another segment has half a 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 six segments is identical to that observed after integration in the host genome (26,
30), we concluded that as observed for three segments in *C. sesamiae* (64), these six *C. typhae*
segments (namely S10_Hdp1, S10_Hdp2, S14_Hdp, S16_Hdp, S26_Hdp1 and S26_Hdp2) originate
from HIM-mediated duplication. The six 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 non-proviral segment is atypical. It is highly similar to *C. sesamiae* segment 13 but
it is not flanked by DRJs or HIMs. Yet, it possesses a single internal DRJ, presumably resulting from
circularization of its parental segment via recombination of the DRJ5’ and DRJ3’. 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
of S13 (S13_Rdp) that, contrary to the above-described duplications, 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, yet S36
and S37 lie at the extremity of two different contigs (Figure 1). In addition, sequencing depth data also
suggest the presence of S13 in *C. typhae* (see section Persistence of non-integrated circles seven days
post-oviposition). In this case, *C. typhae* would have 28 proviral segments in total. It is also noteworthy
that we identified three other segments that we considered as potentially resulting from assembly
errors. These segments are highly similar to segments S1, S14, and S20/33 and could thus 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). We thus decided not to include them in the annotation.

**Annotation of HIM-mediated duplications in other *Cotesia* species**

The finding of six 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*, *C. rubecula*, *C. vestalis* and *C. glomerata*). The highly fragmented nature of these additional genomes prevented us from reaching a high level of confidence in the annotation of all segments. Thus, the results of this search should be considered preliminary. Out of the six HIM-mediated duplications, we were able to investigate whether they are shared by other *Cotesia* species for five of them (S16_Hdp, S10_Hdp1, S10_Hdp2, S26_Hdp1 and S26_Hdp2). Indeed, our approach relies on both J1 and J2 being clearly annotated. We identified four and one segments orthologous to these five 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* only shares 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 conclude about their absence in the fragmented genome of *C. flavipes*. We were not able to find any orthologous HIM-duplications in the other *Cotesia* wasps and we could identify only two candidate *de novo* HIM-mediated duplications, both in *C. sesamiae*. Those involve the two 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 five samples of parasitized *S. nonagrioides* (heads, hemocytes, fat body, ganglionic chain, whole body) by separately mapping trimmed reads on the genome of *S. nonagrioides* and *C. typhae*. We obtained a total of 335 to 595 million trimmed reads depending on samples, 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 71X and 155X depending on the sample (Figure 2). The percentage of reads mapping on *S. nonagrioides*’s genome varied from 73.05% in the hemocytes to 91.94% in the fat body. Mirroring this variation, between 20.89% and 0.31% and of the reads mapped to *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*. In four of the samples, the proportion of *C. typhae*’s genome covered by the reads was high (94-99%) while it dropped to 41% in the fat body. Importantly, the average coverage was higher on proviral segments (68.8X to 289.6X depending on samples) than on the rest of the genome in all samples (2.8X to 110.2X depending on samples) (Figure 2). This is consistent with the presence of a higher proportion of integrated and/or non-integrated wasp bracoviral circles versus other wasp genomic regions in our DNA extracts.

**HIM-mediated integration of CtBV DNA circles into *S. nonagrioides*’s genome**

To identify and quantify integrations of CtBV DNA circles into *S. nonagrioides*’s 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 on all 16 *C. typhae* proviral segments containing a HIM in all five 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 are located in HIMs (Figure S3). In fact, the vast majority of chimeric reads mapped to two short regions located within HIMs and spaced by 41 to 73bp (Example of segment 1 in Figure 3 and Figure S3). Alignment of all 16 HIMs allowed us to identify a conserved motif under each of these regions that correspond to the J1 and J2 junctions previously characterized in *C. congregata* and *M. demolitor* (Figure 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 *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 (Figure 3 and 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 one chance out of four that the base following the junction position in the CtBV circle is the same as that following the junction in the moth genome (Figure 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 (Figure S4b). There is thus one chance out of four 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 one chance out of sixteen that the wasp-host junction will 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 if bracovirus-moth microhomologies at the junctions may somehow foster integration of DNA circles, we compared expected numbers of chimeric reads for each microhomology length to observed ones (see material and methods) (Figure 4). We did this for chimeric reads falling specifically in J1 or J2 and for chimeric reads falling outside of 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 > 3bp. Thus, while certainly affecting the precision of our junction-mapping pipeline, these microhomologies are unlikely to have biological underpinnings. By contrast, the numbers of 0-bp, 1-bp and 2-bp microhomologies differ markedly from what is expected by chance, with 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 (Figure 4a).

Like 3-bp long microhomologies, 1-bp and 2-bp long microhomologies affect the precision of our junction-mapping pipeline. However, their over-representation indicates that they likely have biological underpinnings. About chimeric reads falling outside of J1 and J2, but still in the HIMs, we observed a major under-representation of 0 to 2-bp microhomology (65 versus 478 reads), mirroring a major over-representation of 4 to 13-bp microhomology (478 versus 105 reads). This suggests that when the breakpoint is located further away from the canonical position 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 of HIMs in HIM-containing segments as well as reads mapping to segments that do not contain a HIM. The number of such reads was low. It varied in HIM-containing segments 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). Contrary to chimeric reads mapping to HIMs which are clustered into two regions corresponding to J1 and J2 motifs, reads mapping outside HIMs are dispersed over the circles, with no circle position outside of HIMs mapped by more than one bracovirus-host junction, except for two junctions covered by two reads each. This pattern could suggest that in addition to HIM-mediated integration, circles could integrate into *S. nonagrioides*’s genome through other mechanisms, possibly involving host DNA-repair pathways, as suggested by Wang et al. (31) for 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 material and methods). However, given the low 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 to circle gene content, we compared gene family content between integrated versus non-integrated circles (Figure 5). This comparison was done for all gene families with known predicted domains and containing more than two genes. Overall, it appears that the integration of a circle is associated to its content in gene families (Fisher exact test p-value < 0.01). Three gene families present on at least three segments seem to explain this observation: VANK, Ser_rich and PTP. The gene families contain respectively five, four and 24 genes distributed over four,
three and seven segments, all found integrated in *S. nonagrioides* genome. This observation suggests that integration of these three gene families is important for parasitism success.

### Quantification of integrated bracovirus circles into *S. nonagrioides*’s 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 fall in the J1 and J2 motifs of the HIMs, that is between 730 and 3126 chimeric reads per sample. We found that the vast majority of integrations in the moth’s genome (6784 out of 6940 IE: 98%) were supported by one chimeric read only. Among integrations supported by more than one read (2%), three are supported by three chimeric reads and the rest by two chimeric reads. This pattern indicates that most chimeric reads correspond to independent IEs. Thus, among the host cells we sequenced, almost no cells harbor a shared IE that would originate from a cell division. Figure 6 shows the number of integrations events (IE) we inferred per segment and per sample by counting only once each integration position. This led to 714 to 3064 IE, depending on samples and segments. In order to be able to compare the number of integration events (IE) for each segment between samples, we turned absolute numbers of IE into relative numbers normalized for one million reads mapping to the genome of *S. nonagrioides* (Table S6). Considering all samples together, S1 is the segment with the most integrations, with 6.64 IE Per Million reads mapping on the Host (*S. nonagrioides*) (IPMH), followed by S7, with 3.38 IPMH and then by S26 with 1.76 IPMH (Figure 6a). All the other segments have less than 1.45 IPMH. For all segments, hemocytes is the sample with the most chimeric reads (Figure 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 fat body (Figure 6b). Given the haploid genome size of *S. nonagrioides* (1,021 Mbp (55)), the read size (150 bp) and the number of IPMH, we estimate the average number of IE per genome as follows: (IPMH / read size) * genome size in Mbp. 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 under their integrated versus circularized forms**

We assessed how many of the HIM-containing CtBV circles we sequenced are integrated into *S. nonagrioides*’s genome versus how many there are 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 numbers of IEs per circle are strongly correlated to sequencing depths for all samples (0.7 < Spearman Rho < 0.9; p-value < 0.01) (Figure 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 versus integrated varies depending on the circles, these variations being similar among samples. For example, circle 1 is characterized by the lowest ratio in three out of four samples while circle 16 has the highest ratio in all samples (Figure 7). This likely reflects variation in the efficiency of integration among circles. In addition, it suggests that a significant part of the circles remain non-integrated, 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 also 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 non-integrated circles seven days post-oviposition

We then used 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 (Figure 8; Figure S5). The sequencing depth of segment 15 is close to the average depth on *C. typhae*’s genome, suggesting that this segment is only present under the form of proviral segments in *C. typhae*’s cells present at varying 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 pseudo-segment (Ps15 on Figure 1). All the other segments display higher coverage than *C. typhae*’s genome, suggesting that in addition to their proviral form present in *C. typhae*’s cells, they are also present in the circular form and/or in the integrated form. We found that to the exception of circle 1 and 7 which are characterized by very high sequencing depths and large numbers of IEs, the range of sequencing depths are similar between HIM-containing circles (from 314 to 946) and other circles (from 311 to 937; leaving Ps15 aside) (Figure 8 and Figure S5). Thus, the number of integrating circles found in host tissues seven days post parasitism is similar to that of non-integrating 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 (Figure 8). This observation supports the presence of the parental S13 in *C. typhae*, that 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 was in our assembly instead map to the very similar S13_Rdp. Importantly, the sequencing depth of all other Rdp and Hdp segments is in the range of the sequencing depth of the other regions of the *C. typhae* genome, supporting that they are only present as proviral segments in teratocytes and other residual wasp cells. This is in line with the non-functional nature of these duplicated segments, which are not expected to generate circles (Figure 8).
Distribution of wasp segments throughout the genome of *S. nonagrioides*

We investigated whether DNA circles integrate randomly along *S. nonagrioides*’s 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 9,121 100,000-bp windows covered 89.3% of the genome and bore 88.3% to 90.7% of IE, depending on samples. Under the null hypothesis that segments integrate at random in *S. nonagrioides*’s genome, we assumed that the number of windows bearing integrations follows a Poisson distribution. We compared observed versus expected numbers of IE 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 six cases (p-values < 0.001) (Figure S6). In the case of the pooled dataset, we observed an excess of windows with no IE and with three or more IE and we observed a deficit of windows with one or two IE. Under a Poisson distribution, we do not expect any windows with more than seven IE. However, some windows have up to 23 IE. This result suggests that segments do not integrate entirely randomly into the genome, as observed for CeBV in the genome of *M. sexta* (30). Interestingly, two windows have at least two IE in all five samples, with a maximum of eight IE. These IE come from nine and 11 segments respectively, so the over-representation of IE in these two windows is not due to one 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 IE along the *S. nonagrioides*’s genome. We found that none of these variables was strongly correlated to the density in IE (not shown).
Discussion

HIM-mediated duplications of CtBV in the wasp genome

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

These duplications imply that after circularization, a segment can integrate into the genome of wasp germline cells, as suggested by (64), who identified three 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 to ectopic expression of wasp factors required for integration in these cells. This scenario appears unlikely as it would imply that the entire set of complex processes leading to particle production would accidentally occur in the germline. Second, circle-containing bracoviral particles could sometimes enter accidentally in germ cells of the wasp that produced them. This scenario also seems unlikely as virus particles are released in the calyx lumen located in the posterior part of the ovaries, whereas germline 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-larvae containing a high density of wasp embryos, as observed for *M. croceipes* (65). 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 (66). 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 already parasitized by another wasp. In this respect, it is noteworthy that we only found HIM-mediated duplications in *Cotesia* wasps parasitizing...
stem borers (C. typhae, C. sesamiae, C. flavipes). No such duplications were found in the three other Cotesia species (C. rubecula, C. glomerata, C. congregata) known to parasitize lepidopteran hosts dwelling on plant leaves (67). Whether the type of host and its habitat has an impact on the likelihood of HIM-mediated duplications will have to be reappraised when higher quality genomes will be 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 (64). 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 DRJ5' and DRJ3' are required for circularization) and none of the regulatory sequences allowing bracovirus DNA amplification, which are located at the extremities of the amplified regions (replication units) 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 PDV DNA circles focused only on one tissue type (the 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 body, 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, *C. typhae* genome has been sequenced deeply in the hemolymph (110X), 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 per wasp embryo (68). Teratocytes play various roles during parasitism, including host immunosuppression, production of anti-microbial peptides, and nutritional functions (69, 70). They undergo physiological and morphological changes during development of the wasp embryos, including an increase in ploidy level (71, 72), which likely explains the high sequencing depth obtained over the *C. typhae* genome in the hemolymph. In the other tissues however, sequencing depth on *C. typhae* is much lower (down to 3X in the fat body), indicating that few if any, wasp teratocytes have been sequenced in these 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 non-hemolymph tissues are *bona fide* chromosomal integrations in cells constituting those tissues. Though we did not perform replicates for each tissue, our study seemingly indicates that the number of circle IE is higher in hemocytes than in other tissues, which is in line with the role of hemocytes in immunity (73, 74) and the known effect of circle-borne virulence genes in thwarting host immunity responses (2). This could also reflect that hemocytes are preferentially infected by CtBV resulting in higher abundance of viral circles in these cells as already suggested in Beck *et al.* (75). Given the multiple effects of bracovirus on host physiology, development and behaviour (29), it is likely that integration in a wide range of tissues contributes to parasitism success and is not merely a byproduct of the capacity of bracoviral particles to enter into many cell types.

**Persistence of integrated versus non-integrated CtBV circles**

It has been traditionally assumed that integration of PDV circles was beneficial to the wasp because it allowed persistence and expression of these circles throughout the entire duration of wasp embryonic
and larval development, which can last between 7 to 14 days in laboratory conditions depending on the
species considered (26, 30, 76). Here, we have estimated that seven days post-oviposition 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 one 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
seven days post parasitism as a result of successive divisions of the original IE-bearing cell. However,
given the range of haploid genome sequencing depths (70 to 155X depending on samples), we
estimated that we sequenced a very low number of *S. nonagrioides* cells (maximum of 35 to 77 cells),
so the probability to sequence two cells with the same IE was very low. Therefore, the fact that we find
very few IEs supported by more than one 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 persistence of circles throughout parasitism is that to the
exception of circles 1 and 7, the 14 other integrating circles are not present in a higher quantity than
non-integrating circles, a trend which holds for all tissues (Figure 8 and Figure S5). Thus, it appears
that similar quantities of integrating and non-integrating circles can persist over at least seven days
during parasitism. It follows that integration is not a requirement for persistence during at least half of
the duration of *C. typhae* embryonic development. Interestingly, our data confirm that integrating and
non-integrating circles clearly differ in terms of gene content, with genes such as VANK and PTP being
exclusively present on integrating circles (Figure 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.

29
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, 77) and delivered to the host, and were shown by RNA interference experiments to be involved in circle excision (78). Interestingly, our study of microhomologies between wasp and moth sequences at bracovirus-moth junctions reveals variation in the mechanism of integration. On the 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 position of 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 out of 7746: 8%). 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 (79).

Evolution of HIM in braconid and ichneumonid wasps

Irrespective of whether chromosomal integration of wasp circles involves a single 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 BV-containing microgastroid wasp belonging to another sub-family (Cheloninae).
It was thus proposed that HIMs might have been acquired by the ancestor of Microgastrinae about 741 Myrs ago, independently and well after the domestication of the nudivirus shared by all microgastrid wasps (30). The recent finding that ichnovirus circles from the ichneumonid wasp Diadegma semiclausum undergo chromosomal integration into their host via HIM-like motifs begs 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 (33-78bp), even though some ichnovirus segments have longer intervening sequences (e.g. DsIV-38 and TrIV-F1, which have 311-bp and 1781-bp long intervening sequences). 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 and in which they were searched (31). Indeed, in addition to D. semiclausum ichnovirus, Wang et al. (2021) found HIMs in Tranosema rostrale and Hyposoter fugitivus ichnoviruses but not in Campoletis sonorensis ichnovirus (31). We believe three evolutionary scenarios could explain the presence of HIMs in both braco- and ichno-virus 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 and episomal form (80)). Nudivirus integration properties might have favored the recurrent domestication of nudiviruses by parasitic wasps (9, 11). However, the mechanism of HzNV1 integration has not been characterized yet. Concerning ichnoviruses, as the ancestor belongs to a virus family possibly extinct, nothing is known on 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 or TEs present in both braconid and ichneumonid wasp genomes. In agreement with this scenario, HIM-like motifs containing inverted terminal repeats and involved in site-specific recombination are common in prokaryotes, yeast and viral genomes and TEs (81). 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 (82). HIM sites correspond fairly well to that description. In eukaryotes, several examples have been reported of recombinases originating from TEs such as the RAG1 protein responsible for shuffling immunoglobulin genes in vertebrates (19) and transposases involved in the maturation of paramecium nuclei (83). Finally, a third scenario would imply that HIMs were acquired only once either by Ichneumonidae or Braconidae wasps and were then transferred between the two PDVs. Such transfer could have been favored by integration properties of PDV circles and by the fact that some wasps from the two families are known to parasitize the same host species (84, 85). However, this seems rather unlikely since it is not sufficient to transfer the HIM sequence to provide a functional mechanism, the recombinase gene needs to be transferred at the same time. Yet, the latter is not present on a bracovirus circle (it is packaged as a protein into PDV particles). The characterization of the different proteins involved in circles integration and of the integrases/recombinases encoded in parasitoid wasp genomes will probably be helpful to further shed light on the evolutionary history of HIMs and PDVs at large.

Possible long-term impact of PDV integration in wasp hosts

Previous studies have uncovered bracovirus circle sequences in the genomes of several species of lepidopterans indicating that such sequences have been horizontally transferred from wasp to lepidopterans at some point during the evolutionary history of these insects (18, 86). Though we have not included host germline tissues in this study, our finding that bracovirus circles can integrate in host
tissues other than hemocytes suggests they may also integrate in host germline cells. In this context, it is remarkable that a fairly large number of bracovirus integrations was found in the “whole body” sample (Figure 6), i.e., a host larva in which no wasp larvae were present seven days post parasitism. The absence of wasp embryos in this, and the other five larvae that we did not sequence, could be due either to an 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 in these larvae (87). Relatively high sequencing depth over the entire C. typhae’s genome in the sequenced larva (Figure 2) is in agreement with a possible presence of teratocytes, in turn suggesting that eggs were indeed injected by the wasp. Though we could not assess whether the sequenced larva would have developed into 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 lab (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.
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. *C. typhae* individuals used in this study originated from the *icipe* under the juridical framework of the Material Transfer Agreement CNRS 072057/IRD 302227/00. We thank Arnaud Lemaître and Jean-Marc Aury from national sequencing center (Genoscope) for obtaining the bulk of *Cotesia typhae* genome raw sequences. We are grateful to Fabrice Legeai and Anthony Bretaud for making the genome of *Cotesia typhae* publicly available on the BIPAA platform. *C. typhae* genome sequencing was funded by French National Research Agency ANR (CoteBio ANR17-CE32-0015-02 to L. Kaiser). The work was also supported by ANR project TransVir ANR-15-CE32-0011-01 to C. Gilbert.

Figure legends

**Figure 1. Map of *C. typhae* bracovirus proviral segments.** Proviral segments are represented by filled rectangles. Segments duplicated after circularization are filled with radial gradient colors. Asterisks indicate HIM-bearing circles found to be integrated into *S. nonagrioides*’s genome, corresponding precisely to all segments originating from RU2-3 part of the macrolocus and isolated loci (PL3, PL4, PL5, PL6, PL7, PL8). Each contig or scaffold in which the segments are located are indicated and a line gathers segments that belong to the same proviral locus (PL). The size of the segments and the space between each of them are to scale, unless when slant bars are present. The colors represent the quality of the annotation. In green, we delimited both extremities of the segments with confidence. In orange and red, one or both extremities (Table S5) have to be taken with caution. Although in green, the DRJs of S37 and S26 are truncated probably due to sequencing or assembly issues.
Figure 2. Average sequencing depths in the five samples. In green, yellow and red respectively are the average sequencing depths over the whole genome of *S. nonagrioides*, the whole genome of *C. typhae*, and the 27 *C. typhae* proviral segments.

Figure 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 5’ DRJ to 3’ DRJ. The white portion represents the HIM, not to scale, near 3’ DRJ. (b) Zoom on the 121-bp HIM, showing two regions with many chimeric reads, called J2 for the left one and J1 for the right one. (c) Sequence logo of J2 and J1 generated with weblogo.berkeley.edu using an alignment of the HIMs of the 16 segments that integrate into *S. nonagrioides*’s genome. For J2, we used the 30 bp upstream the minimum position at which we observed more than two chimeric reads, and for J1 we used the 30 bp downstream the maximum position at which we observed more than two chimeric reads. The highly conserved motif J1 is framed in red and J2 in green.

Figure 4. Distribution of microhomology lengths at wasp-host junctions in chimeric reads. Black bars correspond to numbers of observed chimeric reads for each microhomology length. Red asterisks correspond to 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 of J1 or J2.

Figure 5. Integration capacity of segments containing one or more genes belonging to seven gene families. 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 ones correspond to segments that do not integrate.

Figure 6. Number of integration events (IEs) for each segment and sample. Absolute numbers of IEs and of chimeric reads (in brackets) are shown on top of each bar. (a) Barplot comparing the number of IEs for each segment. (b) Barplot comparing the total number of IEs of all segments in each sample.
Figure 7. Histograms showing the number of chimeric reads and sequencing depth for each HIM-containing segments. Top light gray bars show the number of chimeric reads while bottom dark gray bars show sequencing depth. The ratio of sequencing depth over chimeric reads is indicated at the top of each light gray bar. The rho of Spearman indicates the correlation between sequencing depth and the number of chimeric reads for each sample.

Figure 8. Plot of the sequencing depth over the number of chimeric reads for each of C. typhae’s bracovirus segments. Sequencing depths and numbers of chimeric reads have been summed up for all samples. The same plots are shown for each sample in Figure S5. Blue dots represent proviral segments that do integrate into S. nonagrioides’s genome and red dots represent proviral segments that do not integrate. Green dots represent duplicated segments, i.e., Rdp and Hdp segments. The identification number of the segments are shown near each blue dot. For red dots, only the identification numbers S15, S5 and S20/33 are shown, and for green dots only S13_Rdp is indicated. The yellow dash line shows the average depth on C. typhae’s genome when summing up all samples. The rho of Spearman shows the correlation between sequencing depth and number of IEs for segments that do integrate into S. nonagrioides’s genome.
References


39


