Published on Aug 22, 2024 by Arcadia Science **DOI**: 10.57844/arcadia-ad7f-7a6d

Identifying circular DNA using short-read mapping

This workflow lets you find potential circular DNA in your organism of interest using short-read, whole-genome sequencing data and a reference genome. We applied it to parasitoid wasps and some other parasites and found putative circular DNA.

Contributors (A-Z)

Audrey Bell, Adair L. Borges, Feridun Mert Celebi, Reilly O. Cooper, Megan L. Hochstrasser, Elizabeth A. McDaniel, Erin McGeever

Version 2 · Apr 01, 2025

Purpose

We developed a computational method to identify circular DNA using short-read DNA sequencing data and reference genomes. We previously identified capsid-like proteins in some venomous and parasitic organisms **[1]**. Inspired by this work, we wanted to search across a broad range of parasitic organisms for circularized (and thus potentially packaged) DNA cargo that parasites might deliver to their hosts.

We figured that by mapping paired short reads to reference genomes and searching for unusually large apparent distances between them, we could find putative circular DNA. To test this approach, we used our workflow to find putative packaged circular DNA in parasitoid wasps and then applied it to a set of species that includes human-associated parasites. We identified clear patterns of large mapped distances and high

coverage in parasitoid wasps. We also found putative circular DNA regions of interest in many human-associated parasite species in our example dataset, showcasing a use case for the workflow.

This method should be broadly applicable for circular DNA searches across organisms using standard short-read sequencing libraries, providing a fully computational, simple way to work with these data. It can also be a supplementary approach to current wet-lab sample processing methods, which often require time-consuming sequencing library enrichment steps. We hope that researchers looking for circular DNA in any organism will be able to apply this workflow as an early screening step.

- The **Nextflow pipeline**, **Python tools**, and **example use cases** are in this GitHub repository.
- **Data** for our two example results are available here.

We've put this effort on ice! ☒

#TranslationalMismatch #DeadEnd

We found interesting patterns suggesting that several parasite species circularize double-stranded DNA cargo for packaging into viral-like particles, but multiple barriers prevent us from pursuing this further. Disentangling putative circular DNA from false positives requires time-intensive manual checks, and validating if and how organisms of interest deliver DNA to their hosts would require significant additional research. In combination with the smaller market opportunity for dsDNA delivery modalities as compared to established areas like RNA delivery, these hurdles led us to discontinue this project.

Learn more about the Icebox and the different reasons we ice projects.

Background and goals

Parasitoid wasps deliver dsDNA-encoded virulence factors genes to their insect and arachnid hosts using endogenized viruses **[2]**. In past work, we looked across venomous species to see if they have endogenized viral capsids that may let them deliver cargo to the organisms they're biting **[1]**, finding putative capsids across several parasitic species. However, we didn't have a clear way to identify the cargo these species deliver, if any. Because we were most interested in finding novel nucleic acid delivery systems for gene therapy applications and the parasitoid wasps that inspired this work circularize DNA cargo to package into capsids, we developed a method to identify circular DNA in sequencing data. We imagine that other organisms might use a similar approach to deliver genes to their hosts, making circularized DNA a potential hallmark of this host manipulation strategy.

We realized that our method might be broadly useful for researchers interested in exploring circular DNA in their organisms of interest. In this pub, we're sharing the workflow we used to search for circular DNA in short-read DNA sequencing data and provide two example datasets where we've applied it.

The problem

We needed a way to search for circular DNA cargo that parasitic organisms might deliver to their hosts. In earlier work identifying DNA cargo in parasitoid wasps, researchers filtered, purified, and sequenced virus-like particles containing the cargo, a time-consuming and tedious process **[3]**. We wanted to take a computational approach that would let us search for potential DNA cargo across all sorts of parasitic organisms.

Several computational tools and workflows to detect circular DNA — in particular, extrachromosomal circular DNA (eccDNA) — already exist, like Circle-Map **[4]**, ecc_finder **[5]**, circlehunter **[6]**, circdna **[7]**, and eccDNA-pipe **[8]**. However, these tools primarily focus on human circular DNA and aren't as well-suited for finding circular DNA across organisms, partially due to their sample preparation preferences. Some of these tools rely on samples pre-enriched for circular DNA (e.g., Circle-seq **[9]**, Circulome-seq **[10]**), or are built specifically for long-read **[11]** or ATAC-seq **[12]** data. Generating these kinds of data is time-consuming and far less common for non-model

organisms, so we wanted to create a method that works with short-read sequencing data and allows for rapid data exploration.

Our strategy

We developed a workflow that lets scientists explore short-read sequencing data for putative circular DNA without needing special library preparation or sequencing methods.

We split this approach into two distinct steps:

1. **Finding circular DNA**: A Nextflow pipeline downloads short-read DNA sequencing data and reference genomes, maps the reads, and extracts mapped reads with insert sizes > 1 kb. Additional files marking regions of high coverage depth are produced.
2. **Learning about each circular DNA sequence**: Python code directly takes the workflow's output and parses it into filtered mapped reads, coverage depth data, and gene annotation data that you can use to investigate the putative circular DNA segments.

DNA can be circularized at specific junctions, and some forward and reverse reads from a read pair might span the junction. When those reads are mapped back to the linear genome, the distance between the paired reads will be much larger than the expected distance for paired short reads. We chose our approach to rapidly scan for a signal of larger-than-expected insert sizes, with Python functions for further downstream exploration of coverage and gene annotations.

The method

We developed an approach to systematically identify positions in eukaryote genomes where paired short reads have a consistently larger mapped distance than expected, a hallmark of circularized DNA. This approach is usable as a Nextflow pipeline, and we've also created some Python tools to explore the putative circularized DNA.

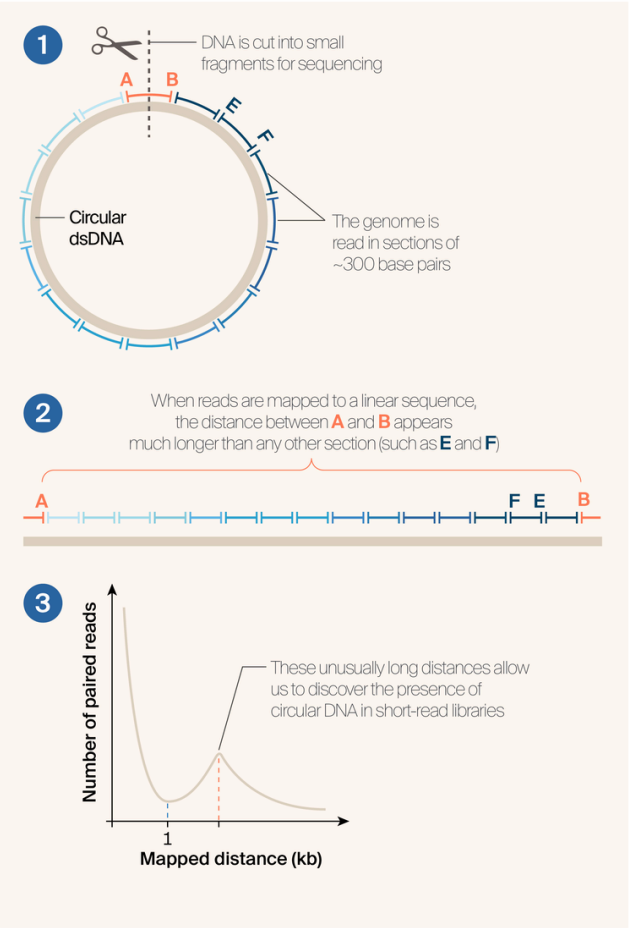


Figure 1

Strategy for identifying circular DNA using short-read DNA libraries.

(1) DNA from samples gets broken down into smaller fragments for short-read sequencing, including circular DNA. Size selection restricts the range of DNA fragment size to those amenable for sequencing, usually 200–600 bp.

(2) Sequenced reads are mapped back to a reference genome. When reads are mapped, if a paired read straddled a circularization junction in circular DNA, the mapped distance between the forward and

reverse reads is far longer than the expected ~300 bp.

(3) If many copies of mapped reads on circular DNA are present, counting the number of reads with mapping distances larger than the expected size range (> 1 kb, a threshold exceeding the expected range of 200–600 bp) will reveal peaks in the data distribution corresponding to the circular DNA's size.

Overarching strategy

We thought we might be able to identify circularized DNA by examining mapped reads for unusually large mapped distances (i.e., the insert sizes between the forward and reverse reads when mapped to the reference genome) between the forward and reverse reads of standard short-read DNA sequencing libraries (Figure 1). Typically, when preparing short-read DNA libraries, at least one size selection step ensures that sequenced segments are mostly between ~200–600 bp, centered around ~300 bp (Figure 1, step 1). However, if any reads straddle a recombination site in circular DNA, mapping those reads back to the linear genome will result in an apparent mapped distance of > 600 bp (Figure 1, step 2). In these cases, the mapped length will correspond to approximately the size of the dsDNA circle. Additionally, circularized DNA is generally present at a higher copy number than linear chromosomal DNA, which could appear as high coverage depth in read-mapping results **[9]**. We hypothesized that circles would be detectable through irregular distance distributions from reads mapped to the genome, and those segments would also likely have higher coverage depth than surrounding DNA (Figure 1, step 3).

To summarize, in scaffolds that don't produce circular DNA, we'd expect a power law distribution (many read pairs with small mapped distances and relatively few with large mapped distances); in scaffolds that do, we'd expect peaks in the distribution at mapped distances that correspond to the length of the circular segment. This turns out to work fairly well, especially for organisms known to produce circular DNA (jump to "Example results..." to see the method in action). We've formalized the read-mapping approach and downstream filtering steps for coverage depth and annotation filtering

into a Nextflow pipeline, which we deployed on Nextflow Tower using AWS Batch spot EC2 instances.

Mapping short reads to find circularized dsDNA

We structured the workflow to take in a sample sheet of reference genomes and corresponding short-read sequencing experiment accessions, structured as a three-column CSV file with a genome accession, the NCBI FTP path, and SRA run accession per line (Figure 2). It handles downloading all genomes and short-read sequencing files with `wget` or `fasterq-dump` from the SRA toolkit (version 3.07 **[13]**) and filtering reads with fastp (version 0.23.4 **[14]**). Next, the workflow maps short reads against the corresponding genome using minimap2 (version 2.28-r1209 **[15]**) and converts from SAM files into sorted BAM files with SAMtools (version 1.20 **[16]**). An `awk` command filters mapped read pairs to only those with mapped distances ≥ 1 kb. Coverage depth is calculated across every position using `samtools depth`, average coverage depth is calculated per scaffold, and positions on each scaffold with coverage depth ≥ 100× the average scaffold coverage depth are identified using `awk`. Then, using BEDTools (version 2.31.1 **[17]**), we merged positions within 100 bases of each other to pinpoint regions of extremely high coverage depth.

The workflow outputs several files usable for downstream analysis:

- ***.sorted.bam**: A sorted BAM file of all mapped reads
- ***.large_inserts.bam**: A filtered BAM file of only reads with mapped distances ≥ 1 kb
- ***.coverage.txt**: A tab-delimited file of coverage depth for each position in the genome
- ***.average_coverage.txt**: A tab-delimited file of per-scaffold average coverage depth
- ***.high_coverage_regions.txt**: A tab-delimited file of positions with coverage depth ≥ 100× average scaffold coverage. Also available in BED format
- ***.high_coverage_region_sizes.txt**: A tab-delimited file of high-coverage depth regions (≥ 100× average scaffold coverage depth, with positions within 100 bp merged into a region), with columns corresponding to scaffold name, start position, end position, and total region size (bp)

- ***.filtered_coverage.txt**: A tab-delimited file of coverage depth only at positions where we identified mapped distances ≥ 1 kb

Considerations for applying the workflow

You may want to consider a more stringent short-read mapping algorithm like BWA **[18]**, especially if you're working with human data, since minimap2 **[15]** isn't as sensitive to small variants and deals with repetitive sequence alignment differently. We used minimap2 primarily for its low-memory overhead and ease of use. Our Nextflow workflow is modular, so it's pretty easy to swap in different programs based on your specific use case.

Importantly, if you suspect your circular DNA comes from multiple different segments or chromosomes of a genome (like in cancer-associated extrachromosomal circular DNAs), this tool won't accurately detect those sequences. We wouldn't recommend using our method in those situations!

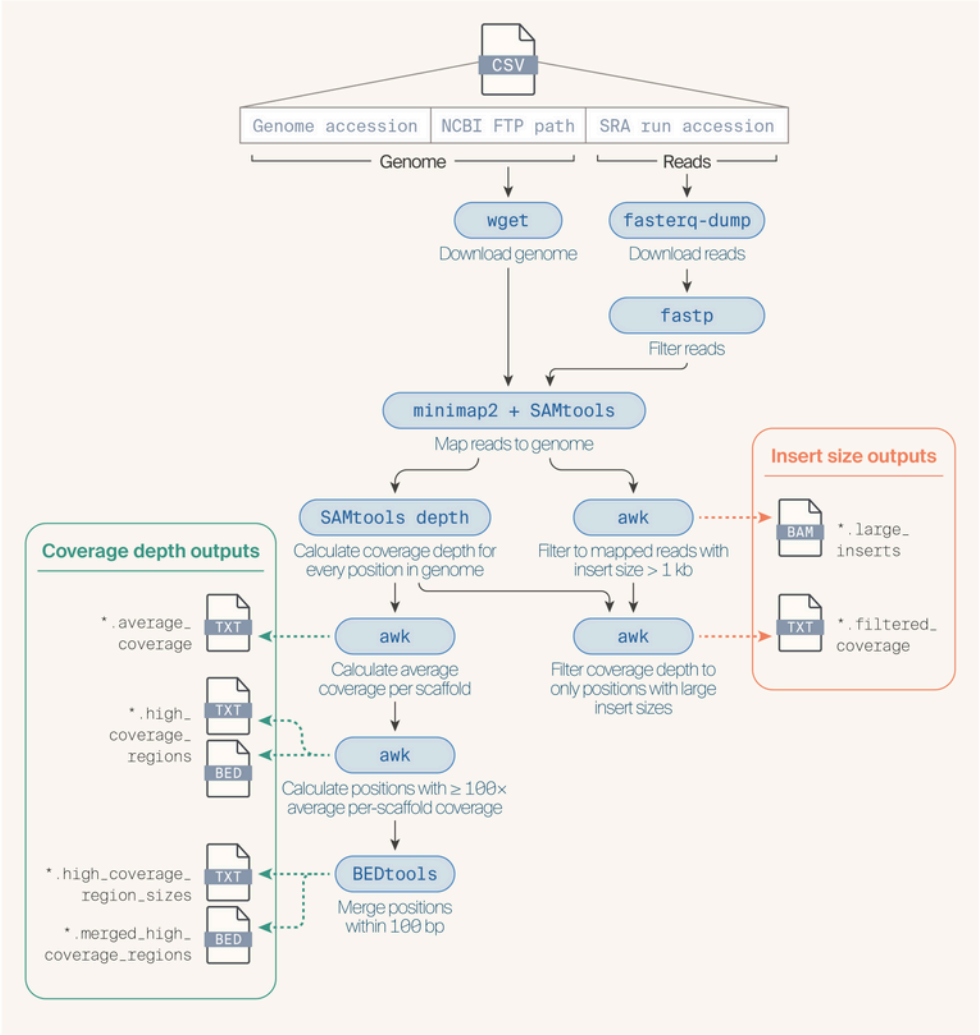


Figure 2

The Nextflow workflow to map and identify putative circular DNA.

Users can specify genomes and associated short-read DNA data using a three-column CSV sample sheet. The workflow maps reads to the genome, and then multiple steps filter the mapped reads and coverage depth data to find regions of interest. Users receive several output files (described above), including mapped reads with unusually large distances and coverage depth information for high-coverage regions.

Parsing the mapped read outputs

When organisms don't have high-quality genomes, the number of scaffolds and inserts this approach identifies is large. We decided to use the Pareto principle to filter scaffolds and inserts for a given genome of interest. The Pareto principle suggests that only a few contributors cause many outputs. In our case, we filter scaffolds to only those contributing to ≥ 80% of the mapped distance data. We summarize all mapped distances (rounded to the nearest kilobase and filtered below a maximum size threshold) and their positions (also rounded to the nearest kilobase) from the filtered scaffolds. We count the number of inserts per position and calculate z-scores to identify statistical outliers. In this case, statistical outliers are regions with many inserts at a given position. Additionally, we extract positions within 10 kb of the inserts. We use the outlier data and extended range to filter each genome's associated coverage depth and annotation data (if present) for easier downstream analysis.

This approach is flexible and usable as the `GenomeInfo` Python class, which uses the Polars package to efficiently parse large datasets. Users must provide the large insert BAM file and filtered coverage file as input to `GenomeInfo`, and gene annotations can optionally be provided in GFF format. Methods in the function will automatically load and parse the mapped distance data (`.load_bam()`), the coverage depth data (`.load_coverage()`), and the annotation data (`.load_gff()`) if provided. Users can provide a list of scaffold names to examine instead of relying on the Pareto principle for filtering or can adjust the Pareto cutoff from its default value of 0.8. Users can then:

- Generate the mapped distance summary with a provided maximum mapped distance threshold to filter with (`.generate_insert_summary(maximum_size_threshold = 100000)`)
- Generate the extended range surrounding inserts with a user-provided base pair width to extend (`.generate_insert_range(bp_width = 20000)`)
- Deduplicate the insert data (`.deduplicate_insert_summary(z_score_threshold = 3)`)

Finally, users can generate filtered coverage depth and annotation files using `.filter_coverage()` and `.filter_gff()`. We hope this framework is usable for genomes and short-read datasets from various organisms.

The **Nextflow pipeline**, **Python tools**, and **code for example use cases** are in this GitHub repository (DOI: 10.5281/zenodo.13363124).

Additional methodology

To identify proviral and non-proviral chromosomes in *Microplitis demolitor*, we downloaded and searched the latest genome assembly's (iyMicDemo2.1a) annotations with the proviral genes annotated in Burke et al., 2018 **[19]**. For *Hyposoter didymator*, we filtered scaffolds for visualization to those with > 10,000 inserts since the genome was more fragmented. To visualize the example results, we used the R data.table (version 1.15.4 **[20]**) package for some preprocessing and the ggplot2 package (version 3.5.0 **[21]**) for visualization. We used the arcadiathemeR package (version 0.1.1) **[22]** to style our visualizations.

We used ChatGPT to help write and clean up code. We also used GitHub Copilot to help write code. Additionally, we used Grammarly Premium to reformat text according to a style guide, streamline and clarify text that we wrote, and suggest wording ideas from which we chose small phrases or sentence structures to use.

Example results from the method

We validated our read-mapping method in parasitoid wasps, organisms that we know deliver circular DNA to their hosts. We then applied the method to some human-associated parasites to demonstrate a use case for species that users of the workflow might be interested in.

SHOW ME THE DATA: data from these example results are available on Zenodo (DOI: 10.5281/zenodo.13362362).

Validating detection of circularized DNA in parasitoid wasps

We wanted to test our read-mapping method with organisms known to make circularized DNA. Parasitoid wasps in the Braconidae and Ichneumonidae families have co-opted viral machinery to manipulate their insect hosts **[2]**. They use integrated viral machinery from polydnaviruses to package circular double-stranded DNA inside of virus-like particles, which they then inject into hosts alongside their eggs. The genetic material inside the virus-like particles compromises host immune responses and significantly increases juvenile survival.

To make and circularize DNA, specific regions of the wasp genome known as “proviral regions” are massively amplified from the wasp genome. Within proviral regions, distinct replication units are individually amplified. Segments in the replication units are then excised, processed by integrase-mediated recombination to produce circularized segments, and packaged **[23]**. In parasitoid wasps, it’s relatively easy to identify replication units by examining mapped reads for regions of extraordinarily high coverage depth (> 100–20,000× the average coverage depth of the wasp genome) **[24][25]**, which are signals of amplification. We developed a method to identify the segments within those replication units that are excised, circularized, and packaged, as described in the prior section. By looking at mapped distance distributions, we hypothesized that we could find unusually large distances in segments within regions of high coverage depth, a pattern indicative of circular DNA.

We analyzed four parasitoid wasp species to check that a distance pattern was only present in wasps that produce circular DNA. Only female wasps of post-reproductive maturity create circular DNA and virus-like particles, so we ran multiple samples per species (when available) to account for samples prepared from males, females, or mixtures.

We first focused on *Microplitis demolitor*, a well-studied braconid wasp with a high-quality genome (iyMicDemo2.1a) and male- and female-only short-read libraries (SRR1565751, SRR2011474), to validate our approach. Not every *M. demolitor* chromosome has proviral segments, so we expected to find peaks in the mapped distance distributions only in the chromosomes with those segments. Moreover, we wanted to make sure that the read mapping approach was identifying circularized DNA, not just proviral segments. Male wasps have the proviral segments in their

genome but don’t circularize and package it; consequently, only mapped reads from female samples should show distance distribution peaks, while distributions from males should look similar to the non-proviral chromosome distance distributions.

Because we were able to use known *M. demolitor* bracovirus genes to identify proviral chromosomes, we didn't filter scaffolds using a Pareto cutoff since this would filter out the non-proviral chromosomes (see “Parsing the mapped read outputs” for more on this cutoff). In female wasps, we observed that chromosomes with known proviral segments had noticeable peaks in mapped distance distributions (Figure 3, A). The peaks are dissimilar across proviral chromosomes, matching the size of circularized DNA from different proviral segments. Except for one peak (corresponding to an unannotated region), we don’t see similar mapped distance distributions in mapped reads from male wasps. Overall, these results suggest we can indeed identify dsDNA circles in parasitoid wasps using this read-mapping approach.

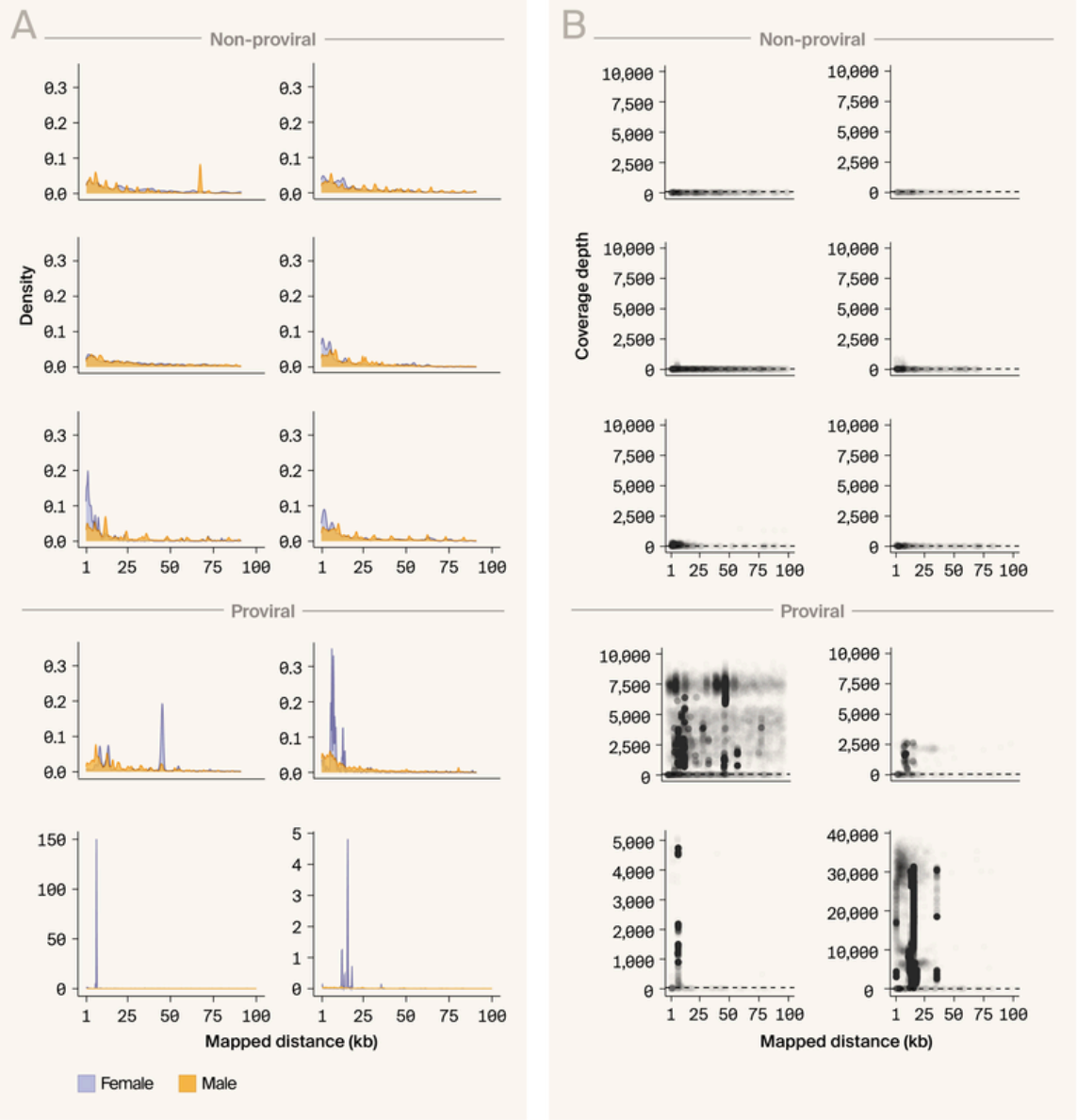


Figure 3

Mapped distance and coverage depth distributions across *Microplitis demolitor* chromosomes.

We filtered out reads with mapped distances < 1 kb to highlight irregular peaks in the distance distributions.

(A) We mapped one female short-read DNA library (blue) and one male (amber) to the *M. demolitor* genome. After filtering the mapped reads to only include those with mapped distances 1–100 kb (approximately the largest replication unit size for this wasp species), we examined mapped distance (x-axis, in kilobases) distributions across chromosomes (individual subplots, with variable y-axis for

two chromosomes on the bottom depending on distribution density). We identified chromosomes as containing proviral segments by matching annotations from the genome assembly against known proviral genes from **[19]**. We removed chromosome names for simplicity, but you can find the underlying data here.

(B) We merged mapped distance data (x-axis) with coverage data [y-axis, variable for the two chromosomes as in A at matching positions per chromosome (individual subplots)]. The dashed line represents average genome coverage.

Next, we wanted to verify that the peaks in mapped distance occurred within regions of high coverage. We merged *M. demolitor*'s read mapping data with the coverage depth data by chromosome position and examined if large distance peaks occurred at high coverage (Figure 3, B). We observe that high coverage corresponds with peaks in mapped distances, further supporting our ability to identify circular DNA from short-read sequencing libraries. Additionally, coverage depth differs across segments of large mapped distance, supporting the pattern of non-equimolar abundance of unique dsDNA circles found in parasitoid wasp virus-like particles **[26]**.

To test our approach with a set of controls, we examined libraries from three other parasitoid wasps. As a negative control (no expectation of large mapped distances), we used a sample of *Cotesia congregata* male wasps, which has a similar bracovirus to that of *Microplitis demolitor* and well-defined replication units **[27]**. As a second negative control, we used a library from *Venturia canescens*, an ichneumonid wasp that has more recently acquired an ichnovirus that incorporates virulence proteins in its virus-like particles rather than DNA **[28]**. Finally, to test if we could find proviral sequences in an ichneumonid wasp, we looked at two female libraries from *Hyposoter didymator*, which has a dsDNA-encoding ichnovirus **[29]** and should exhibit the large mapped distance distribution pattern.

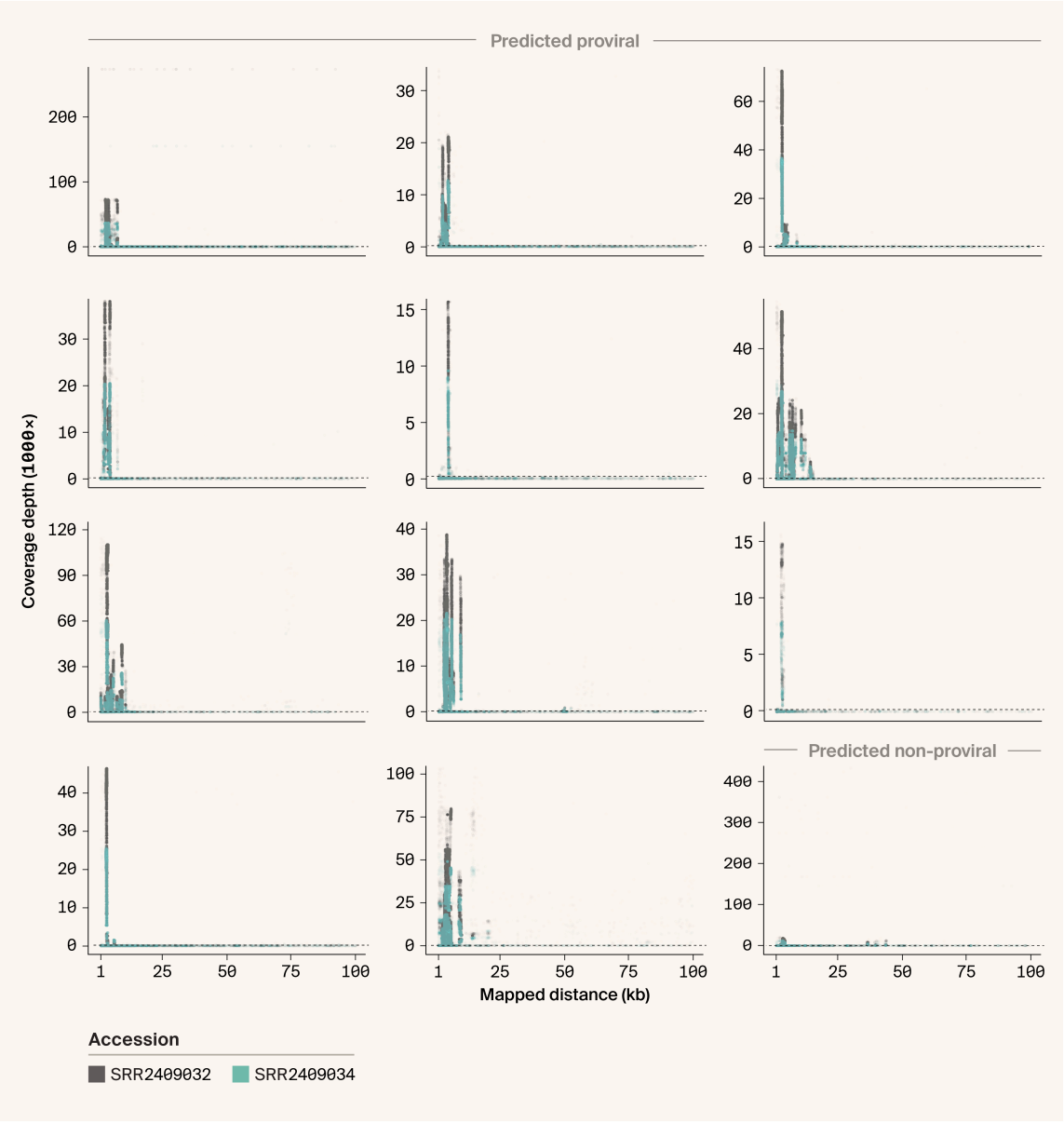


Figure 4

Mapped distance and coverage depth distributions across *Hyposoter didymator* chromosomes.

We mapped two female short-read libraries to the *H. didymator* genome and filtered to mapped distance < 100 kb as in *Microplitis demolitor*. We found mapped distance and coverage patterns indicating the presence of proviral segments across most of the scaffolds except one (bottom right). The dashed line represents average *H. didymator* genome coverage. You can find the underlying data, including chromosome names, here.

Our results were confirmatory: we found a clear mapped distance and coverage pattern in two short-read libraries from *H. didymator* (Figure 4) but no clear signature of circular DNA in either of the other species we'd included as negative controls. Since the latest version of *H. didymator*'s genome didn't have publicly available annotations, we couldn't confirm if the inserts we detected were proviral by leveraging existing annotation data as we'd done with *M. demolitor*. Instead, we first extracted the sequence from the most common mapped distance and position from each scaffold (12 total), plus 2 kb in each direction to capture flanking segments. For each, we then manually performed a BLASTx search against the NR database (June 2024). The top hits for 10 of the 12 segments were *Hyposoter* ichnovirus virulence-associated proteins (nine *H. didymator*, one *H. fugitivus*) and the other two homologs of another parasitoid wasp's proteins (Table 1). These results are promising for detecting genes within inserts — even though *H. didymator*'s ichnovirus has been described by sequencing its virus-like particles **[30]**, we were able to identify ichnovirus genes just by looking for locations with large inserts. In total, these results indicate that we can reliably identify circular DNA from parasitoid wasps using this read-mapping strategy.

Hyposoter didymator scaffold	BLASTx hit description	Hit scientific name	E-value	Percent Identity	Ac
JBAJMU01000001.1	Vinx1	*Hyposoter didymator* ichnovirus	0.0	96.02	
JBAJMU01000002.1	Repeat element protein-d7.3	Ichnoviriform fugitivi	6e−110	76.57	
JBAJMU01000003.1	Vinx1	*Hyposoter didymator* ichnovirus	0.0	99.73	
JBAJMU01000004.1	Vacuolar protein sorting-associated protein 18 homolog isoform X2	*Venturia canescens*	1e−98	62.16	
JBAJMU01000005.1	Cys1	*Hyposoter didymator* ichnovirus	1e−63	89.92	
JBAJMU01000006.1	FSU2	*Hyposoter didymator* ichnovirus	1e−73	98.39	
JBAJMU01000007.1	1-acyl-sn-glycerol-3-phosphate acyltransferase gamma-like	*Venturia canescens*	3e−83	95.58	
JBAJMU01000008.1	N-gene1	*Hyposoter didymator* ichnovirus	0.0	99.09	
JBAJMU01000009.1	Rep1	*Hyposoter didymator* ichnovirus	2e−149	97.01	
JBAJMU01000010.1	PRRP3	*Hyposoter didymator* ichnovirus	3e−28	92.31	
JBAJMU01000011.1	Cys4	*Hyposoter didymator* ichnovirus	3e−140	89.39	

Hyposoter didymator scaffold	BLASTx hit description	Hit scientific name	E-value	Percent Identity	Ac
JBAJMU01000012.1	Rep1	*Hyposoter didymator* ichnovirus	6e–156	99.57	

Table 1

Top BLASTx hits from the most common insert size and position of each *Hyposoter didymator* scaffold.

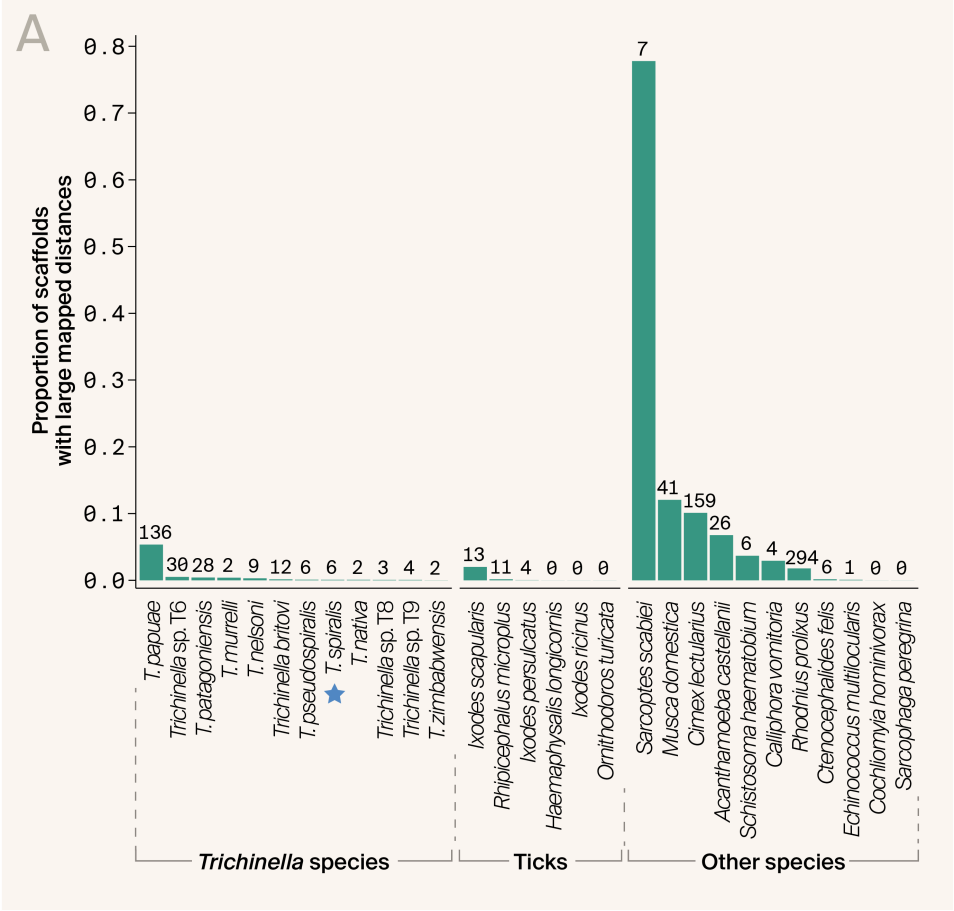
Searching for circularized DNA in human-associated parasites and related species

We wanted our method to be broadly helpful in finding and exploring circular DNA in diverse organisms, so we tested this by applying the Nextflow pipeline to 58 samples from 29 human-associated parasites and related species (12 *Trichinella* species, 6 tick species, and 11 other species, including kissing bugs, parasitic flies, non-parasitic flies, and tapeworms). The sample sheet we used for this example is provided on GitHub. We used multiple samples per species when possible to account for any differences in sex or reproductive maturity, even though we weren't sure how relevant these traits were beyond parasitoid wasps for determining the presence of circular DNA.

Across the 29 species in our dataset, our workflow identified 24 with putative circular DNA (Figure 5, A). We briefly examined the available annotations of scaffolds with large inserts to find genes that might be present within the putative circular DNA segments using the methods provided in `GenomeInfo`. We found some genes implicated in host-parasite interactions **[31]** within the most common large insert in *Trichinella spiralis* (Figure 5, B). We checked that our workflow wasn't just flagging this insert due to multiple gene copies by examining the count of other mapped distances > 1 kb in the same region. If reads were randomly mapping to different copies of the same gene within the segment, we'd expect relatively similar counts of mapped distances between other copies of the genes in the region; instead, we only found evidence of this insert. Additionally, the nucleotide sequences of this multi-copy gene are relatively different within the insert, and as such, we wouldn't expect to find reads mapping indiscriminately to different copies.

However, it's important to note that we also found many false positives within these annotations, including retrotransposons, ribosomal RNAs, and mitochondrial genes. Because this workflow looks for large mapped distances between forward and reverse reads, areas of the target organism's genome with multiple copies of genes with near-identical sequences or repetitive elements could be detected. Users should carefully, manually inspect the flagged inserts from this workflow and validate with orthogonal methods, or consider implementing a supplementary mapping filter in the read-mapping step. Additionally, fragmented genome assemblies will be more difficult to analyze since small scaffolds with few inserts may still be considered outliers.

> All **output files**, including the large insert BAM files, filtered coverage and annotation files (where available), and mapped distance and position summaries for the parasite dataset, are available on Zenodo.



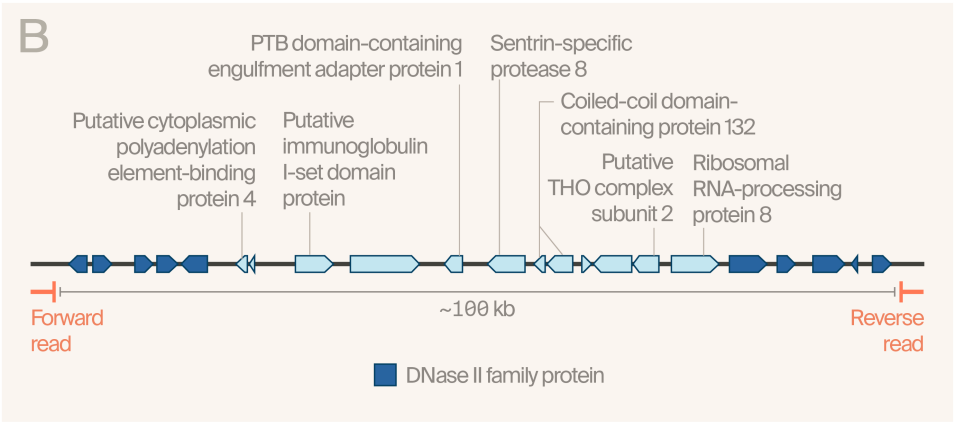


Figure 5

We detect genomes with large apparent inserts across parasite and non-parasite species.

(A) We applied our Nextflow pipeline to a collection of species, including parasites, and then processed the mapped distance and coverage outputs. For each parasite species (x-axis), we show the proportion of total scaffolds in the genome assembly with large inserts that we detected with our pipeline (y-axis) and the total

number of scaffolds with large inserts (numbers above bars), as genome fragmentation varied substantially across species. This figure highlights the complexity of identifying putative circular DNA signals in fragmented genomes and isn't meant to be comparative across groups of species. We've marked *Trichinella spiralis* with a star to indicate that this is the species we focus on in B.

(B) Representative annotations from a large insert we detected in *T. spiralis* showing multiple copies of DNase II, a gene family involved in parasite–host interactions. We've indicated known functional annotations, and we show coding regions with only hypothetical functions without text labels. We found many pairs of forward and reverse reads that span the ends of the region depicted here and show one such example in orange. Finding spanning read pairs and large mapped distances suggests that this ~100 kb element is circular.

Next steps

We've decided not to use this approach to pursue identifying dsDNA cargo in virus-like particles any further. Still, we believe our workflow is generally helpful for finding circular DNA across organisms. For researchers studying parasitoid wasps, whether for use as pest control or for more general ecological and evolutionary biology research, our method appears to reliably identify circularized and packaged DNA without needing to sequence virus-like particles. For scientists studying circular DNA more broadly or in specific target organisms, this workflow could be implemented to look for similar patterns of larger-than-expected mapped distances between paired reads. If you decide to use this workflow, we'd enjoy hearing how it goes here or on GitHub.

References

1 Borges AL, Bigge BM, Chou S, McDaniel EA, Poskanzer KE, York R. (2024). Identification of capsid-like proteins in venomous and parasitic animals. https://doi.org/10.57844/ARCADIA-14B2-6F27

2 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. https://doi.org/10.1098/rstb.2013.0051

3 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. https://doi.org/10.1126/science.1103066

4 10.1186%2fs12859-019-3160-3

5 Zhang P, Peng H, Llauro C, Bucher E, Mirouze M. (2021). ecc_finder: A Robust and Accurate Tool for Detecting Extrachromosomal Circular DNA From Sequencing Data. https://doi.org/10.3389/fpls.2021.743742

6 Yang M, Zhang S, Jiang R, Chen S, Huang M. (2023). Circlehunter: a tool to identify extrachromosomal circular DNA from ATAC-Seq data. https://doi.org/10.1038/s41389-023-00476-0

7 nf-core. (2024). circdna. https://github.com/nf-core/circdna/tree/1.0.1

8 Fang M, Fang J, Luo S, Liu K, Yu Q, Yang J, Zhou Y, Li Z, Sun R, Guo C, Qu K. (2024). eccDNA-pipe: an integrated pipeline for identification, analysis and visualization of extrachromosomal circular DNA from high-throughput sequencing data. https://doi.org/10.1093/bib/bbae034

9 Møller HD, Parsons L, Jørgensen TS, Botstein D, Regenberg B. (2015). Extrachromosomal circular DNA is common in yeast. https://doi.org/10.1073/pnas.1508825112

10 Shoura MJ, Gabdank I, Hansen L, Merker J, Gotlib J, Levene SD, Fire AZ. (2017). Intricate and Cell Type-Specific Populations of Endogenous Circular DNA (eccDNA) in *Caenorhabditis elegans* and *Homo sapiens*. https://doi.org/10.1534/g3.117.300141

11 Li F, Ming W, Lu W, Wang Y, Li X, Dong X, Bai Y. (2023). FLED: a full-length eccDNA detector for long-reads sequencing data. https://doi.org/10.1093/bib/bbad388

12 Kumar P, Kiran S, Saha S, Su Z, Paulsen T, Chatrath A, Shibata Y, Shibata E, Dutta A. (2020). ATAC-seq identifies thousands of extrachromosomal circular DNA in

cancer and cell lines. https://doi.org/10.1126/sciadv.aba2489

13 NCBI. (2023). sra-tools. https://github.com/ncbi/sra-tools/tree/3.0.7

14 Chen S. (2023). Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. https://doi.org/10.1002/imt2.107

15 Li H. (2018). Minimap2: pairwise alignment for nucleotide sequences. https://doi.org/10.1093/bioinformatics/bty191

16 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. (2009). The Sequence Alignment/Map format and SAMtools. https://doi.org/10.1093/bioinformatics/btp352

17 Quinlan AR, Hall IM. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. https://doi.org/10.1093/bioinformatics/btq033

18 Li H, Durbin R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. https://doi.org/10.1093/bioinformatics/btp324

19 Burke GR, Walden KKO, Whitfield JB, Robertson HM, Strand MR. (2018). Whole Genome Sequence of the Parasitoid Wasp *Microplitis demolitor* That Harbors an Endogenous Virus Mutualist. https://doi.org/10.1534/g3.118.200308

20 Rdatatable. (2024). data.table. https://github.com/Rdatatable/data.table/tree/1.15.4

21 tidyverse. (2024). ggplot2. https://github.com/tidyverse/ggplot2/tree/v3.5.0

22 Arcadia Science. (2024). arcadiathemeR. https://github.com/Arcadia-Science/arcadiathemeR

23 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. https://doi.org/10.1128/jvi.01388-15

24 Lorenzi A, Legeai F, Jouan V, Girard P-A, Strand MR, Ravallec M, Eychenne M, Bretaudeau A, Robin S, Rochefort J, Villegas M, Burke GR, Rebollo R, Nègre N, Volkoff A-N. (2024). Identification of a viral gene essential for the genome replication of a domesticated endogenous virus in ichneumonid parasitoid wasps. https://doi.org/10.1371/journal.ppat.1011980

25 Krell PJ, Summers MD, Vinson SB. (1982). Virus with a Multipartite Superhelical DNA Genome from the Ichneumonid Parasitoid *Campoletis sonorensis*. https://doi.org/10.1128/jvi.43.3.859-870.1982

26 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. https://doi.org/10.1098/rstb.2013.0047

27 Leobold M, Bézier A, Pichon A, Herniou EA, Volkoff A-N, Drezen J-M. (2018). The Domestication of a Large DNA Virus by the Wasp Venturia canescens Involves Targeted Genome Reduction through Pseudogenization. https://doi.org/10.1093/gbe/evy127

28 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. https://doi.org/10.1186/s12915-020-00822-3

29 Dorémus T, Cousserans F, Gyapay G, Jouan V, Milano P, Wajnberg E, Darboux I, Cônsoli FL, Volkoff A-N. (2014). Extensive Transcription Analysis of the Hyposoter didymator Ichnovirus Genome in Permissive and Non-Permissive Lepidopteran Host Species. https://doi.org/10.1371/journal.pone.0104072

30 Qi X, Yue X, Han Y, Jiang P, Yang F, Lei JJ, Liu RD, Zhang X, Wang ZQ, Cui J. (2018). Characterization of Two Trichinella spiralis Adult-Specific DNase II and Their Capacity to Induce Protective Immunity. https://doi.org/10.3389/fmicb.2018.02504