Comparative analysis reveals the expansion of mitochondrial DNA control region containing unusually high G-C tandem repeat arrays in *Nasonia vitripennis*

Zi Jie Lin[1], Xiaozhu Wang[2], Jinbin Wang[3], Yongjun Tan[4], Xueming Tang[3], John H. Werren[5], Dapeng Zhang[4], and Xu Wang[2,6,7,8]

[1]*Department of Chemistry, Columbus State University, Columbus, GA 31909*

[2]*Department of Pathobiology, Auburn University, Auburn, AL 36849*

[3]*Institute of Biotechnology Research, Shanghai Academy of Agricultural Sciences; Key Laboratory of Agricultural Genetics and Breeding, Shanghai 201106, China*

[4]*Department of Biology, College of Arts & Sciences, Saint Louis University, St. Louis MO 63103*

[5]*Department of Biology, University of Rochester, Rochester, NY 14627*

[6]*HudsonAlpha Institute for Biotechnology, Huntsville, AL 35806*

[7]*Alabama Agricultural Experiment Station, Auburn University, Auburn, AL 36849*

[8]*Department of Entomology and Plant Pathology, Auburn University, Auburn, AL 36849*

Corresponding author:

Xu Wang

Phone: (334) 844-7511

Fax: (334) 844-2618

E-mail: xzw0070@auburn.edu

ORCID: 0000-0002-7594-5004

Keywords: mitochondrial genome evolution, inversion, tandem repeats, GC-content, *Nasonia*, *Muscidifurax*

Abstract

Insect mitochondrial DNA (mtDNA) ranges from 14 to 19 Kbp, and the size difference is attributed to the AT-rich control region. Jewel wasps have a parasitoid lifestyle, which may affect mitochondria function and evolution. We sequenced, assembled, and annotated mitochondrial genomes in *Nasonia* and outgroup species. Gene composition and order are conserved within *Nasonia*, but they differ from other parasitoids by two large inversion events that were not reported before. We observed a much higher substitution rate relative to the nuclear genome and mitochondrial introgression between *N. giraulti* and *N. oneida*, which is consistent with previous studies. Most strikingly, *N. vitripennis* mtDNA has an extremely long control region (7,665 bp), containing twenty-nine 217 bp tandem repeats and can fold into a super-cruciform structure. In contrast to tandem repeats commonly found in other mitochondria, these high-copy repeats are highly conserved (98.7% sequence identity), much longer in length (approximately 8 Kb), extremely GC-rich (50.7%), and CpG-rich (percent CpG 19.4% vs. 1.1% in coding region), resulting in a 23 Kbp mtDNA beyond the typical size range in insects. These *N. vitripennis*-specific mitochondrial repeats are not related to any known sequences in insect mitochondria. Their evolutionary origin and functional consequences warrant further investigations.

1. Introduction

Parasitoids are insects that deposit their eggs on or into other arthropods, where the young develop, eventually killing the host. Parasitoids are used extensively in biological control as an alternative to using pesticides to control agricultural pest insect population [1, 2]. *Nasonia* (also known as jewel wasps) are parasitoids of fly pupae that are also widely used as the model system for genetics and parasitoid biology [3, 4]. The *Nasonia* genus consists of four closely-related species: *N. vitripennis, N. giraulti, N. longicornis,* and *N. oneida* [4-6]. They all share a haplodiploidy system of sex determination, in which unfertilized eggs will develop into males while fertilized eggs will develop into females. Four *Nasonia* species are not cross-fertile due to the action of infected *Wolbachia*; however, they can be cross-hybridized to create F1 progenies after curing them with antibiotics [7]. Multiple studies revealed that *Nasonia* utilize DNA methylation, which is absent in the primary genetic model insect species, *Drosophila melanogaster* [8-11].

Nasonia mitochondrial genomes experience an unusually high rate of evolution, approximately 30-40 times faster than that of the nuclear genome in *Nasonia* based on limited sequence data sets available at that time [12]. This finding supports the hypothesis of compensatory feedback in the nuclear genome of *Nasonia*, where the nuclear-encoded genes of the oxidative phosphorylation pathway underwent positive selection to compensate for the deleterious mutations in the mitochondria-encoded proteins due to rapid mutations [12-14]. As expected from this elevated mutation rate, nuclear-mitochondrial incompatibilities play a significant role in hybrid lethality among *Nasonia* species [7, 15-17]. None of the previously-sequenced mitochondrial DNA (mtDNA) of *Nasonia* has been circularized. This is most likely due to its

highly repetitive region, similar to *Pteromalus puparum*, which belongs in the same family of Pteromalidae [18]. Despite the efforts, the circularization of the mitochondrial genomes of *Nasonia* remained a challenge.

The mitochondrial control region, or the 'D-loop region' in animals, is a critical element in regulating transcription and DNA replication [19, 20]. The control region in insects is generally rich in A+T content, with 85% in most insect mitochondrial genomes [21]. It varies in size and nucleotide composition even among species of the same genus [21]. For example, the size of the control region in *Drosophila yakuba* is 1,077 bp, while the one in *D. melanogaster* is 4,601 bp [21]. The size variation is mainly due to the tandem repeats found in the control region. Tandem repeats in the control region not only vary between and within species, but even within individuals. Length heteroplasmy can occur in individuals when the copy number of tandem repeats differs between cells and tissues [21]. The mechanism for the generation of tandem repeat units is still incompletely understood. However, it has been hypothesized that replication slippage could be the root cause of this phenomenon [22-25].

In this work, we presented the assembly and annotation of ten mitochondrial genomes, including eight *Nasonia* strains from four species, the closely related species *Trichomalopsis sarcophagae*, and *Muscidifurax raptorellus*. These are then used in a comparative study of mitochondrial genome evolution in this group of insects that have rapid mitochondrial sequence evolution. All these species occur in the parasitoid family Pteromalidae. We conducted phylogenetic analyses on the protein-coding genes of six wasp species to determine their evolutionary relationship using the mitochondrial genomes and to compare rates of substitution to their nuclear genomes.

We have discovered a series of unusual properties of the mitochondrial genome of *N. vitripennis*, which enrich the current understanding of insect mtDNA structure and evolution.

2. Materials and Methods

2.1. Sample collection and genomic DNA extraction.

All ten wasp strains sequenced in this study are highly inbred and maintained at 25℃ with constant light. For *Nasonia vitripennis* (*Nv*), four strains were sequenced: LabII, a standard laboratory strain originally derived from the Netherlands; AsymCx, a *Wolbachia*-free strain produced from LabII by antibiotic treatment [7]; R5-11, collected in Rochester, New York and V12.1, derived from R5-11 after having lost of one *Wolbachia* strain [26]. In addition, we utilized a strain (R16A) with an *Nv* mtDNA mitochondrial genome in a *Nasonia giraulti* (*Ng*) nuclear genome background. The R16A strain was produced by 16 generations of back-crossing of the *Nv* AsymCx strain approximately 25 years ago [7]. For the other species, we utilized the *Ng* RV2X(u) strain with *Ng* mtDNA [27], *Nasonia longicornis* (*Nl*) IV7(u) strain with *Nl* mtDNA, *Nasonia oneida* (*No*) NONY strain with *No* mtDNA [28], *Trichomalopsis sarcophagae* (*Ts*) Tsarc strain with *Ts* mtDNA, and *Muscidifurax raptorellus* (*Mell*) Chile strain. High molecular weight (HMW) genomic DNA (gDNA) was isolated from each strain using the MagAttract HMW DNA Mini Kit (Qiagen, MD). The quality and size distribution of extracted gDNA were examined on a Qubit 3.0 Fluorometer (Thermo Fisher Scientific, USA) and on Agilent TapeStation 4200 (Agilent Technologies, CA) using the genomic DNA kit.

2.2. Genomic DNA sequencing.

A 10X Genomics® library for each strain was prepared with the Chromium Genome Reagent Kits v2 on the 10X Genomics® Chromium Controller (10X Genomics Inc., CA). HMW gDNA was diluted from its original concentration to 0.8~0.9 ng/µl with EB buffer. The diluted denatured gDNA, sample master mix, and gel beads were loaded to the genomic chip, then ran

on the 10X Genomics® Chromium Controller to create Gel Bead-In-EMulsions (GEMs). The obtained GEMs were used for the subsequent incubation and cleanup. Chromium i7 Sample Index was used as the library barcode. Quality control of post library construction was accessed with Qubit 3.0 Fluorometer for the final library concentration and Agilent TapeStation 4200 for the final library size distribution. The prepared 10X Genomics® library for each strain was sequenced on a HiSeq X or a NovaSeq 6000 sequencer. On average, we obtained 476,368,658 reads per strain, with mean haploid genome coverage of 334.95X (Table S1).

2.3. Assembly and annotation of mitochondrial genome in wasp species.

The 10X Genomics® sequencing reads were first checked for sequence quality using FastQC [29]. The *de novo* assembly of each strain was constructed with customized parameters using Supernova 2.0 [30]. To identify the mitochondrial scaffolds from the draft whole genome assembly of each strain, previous *Nv* mitochondrial draft assemblies (NCBI accession number: EU746609.1 and EU746613.1) were aligned to each *de novo* assembly using BLAT [31]. The scaffolds with identities larger than 85% were kept as potential mitochondrial scaffolds for each species. These scaffolds were then further checked for their read coverage and GC-content. To distinguish true mitochondrial DNA and NuMTs (nuclear mitochondrial DNA), only scaffolds with a coverage greater than 5,000X were kept for the draft mitochondrial genome assembly of each species (Table 1 and Table S1).

The *Nv* mtDNA repeat region and the two palindromic arms were initially characterized using 10X Genomics® data and the Sanger sequencing scaffolds in the *Nv* reference genome [16] AsymCx strain assembly v2.1 (Scaffold2008 and Scaffold1962 in assembly GCA_000002325.2).

The *Nv* AsymCx mtDNA was circularized using sequencing read data from Nv PSR1.1 assembly derived from AsymCx background with NCBI accession number PRJNA575073 [32]. The PacBio and ONT (Oxford Nanopore technology) reads were BLATed against with our mtDNA assembly using 10X Genomics® technology, and 196 reads aligned to the palindromic arms were identified. Among these reads, four were informative to close the mtDNA assembly (Table S2).

The initial draft mitochondrial genome assemblies of *N. giraulti*, *N. longicornis*, *N. oneida*, and *T. sarcophagae* contained a few gap regions with unidentified sequences. Additionally, *T. sarcophagae* had several fragmented/duplicated regions. To further improve the mitochondrial assemblies, problematic regions were extracted for manual improvement. The sequencing reads were aligned to these regions with BWA-MEM [33], and the aligned reads were extracted to generate contigs using MEGAHIT [34]. The mapped reads were used as additional evidence to improve the mitochondrial assemblies of all five species. The final assembly versions were manually curated using the Geneious platform [35]. We used MITOS [36] to annotate the open reading frames (ORFs) in the mitochondrial genome of all five species (Table 2). The NCBI GenBank accession numbers are included in Table S3.

2.4. Phylogenetic analysis of mitochondrial genes.

To detect the evolutionary relationships of mitochondrial genes among ten wasp species, phylogenetic analyses of 13 core mitochondrial genes (*atp6*, *atp8*, *cob*, *cox1*, *cox2*, *cox3*, *nad1*, *nad2*, *nad3*, *nad4*, *nad4l*, *nad5*, and *nad6*) and concatenated nucleotide sequences of these core genes was performed. Another parasitoid species, *Muscidifurax raptorellus* (*Mell*), was selected as an outgroup for this analysis. It is a closely related species, belongs to the family of

Pteromalidae, and similarly, this species is also Dipteran parasitoid. Additionally, the 2-fold and 4-fold degenerate sites were extracted from the alignment of all protein-coding genes as well as the concatenated sequences (Figure S1). The cut-off for the minimum number of 4-fold degenerate sites was 10; therefore, *atp8* was excluded from the 4-fold degenerate site analysis. The coding sequences from the ten strains, including four different *Nv* strains (AsymCx, LabII, V12.1, and R5-11), were aligned using the MUSCLE alignment tool v3.8 [37]. The phylogenetic tree was constructed with coding sequences of mitochondrial genes using the Maximum-Likelihood (ML) method in MEGA version X [38]. The substitution model was selected based on the lowest BIC values produced by MEGA-X. Bootstrap tests with 1,000 replicates were used to evaluate the phylogenetic trees. For specific substitution models used for each gene, please see Table S4. The nuclear genome trees were generated by concatenating 8,133 single-copy orthologs using RAxML v8.2 [39] with the VT protein model (best fit model identified by ProtTest 3) and 1,000 rapid bootstrap replicates [27].

2.5. Identification and characterization of the tandem repeat array in the *Nv* mitochondrial genome.

In the *Nv* mitochondrial genome, a large region containing tandem repetitive sequences was identified. To further characterize this region and determine if this repeat array is *Nv*-specific, we aligned sequencing reads from all wasp species to the assembled *Nv* mitochondrial genome. The mitochondria replacement strain R16A is especially informative, as it has the *Ng* nuclear genome but *Nv* mitochondria background, which was generated through 16 generations of back-crossing [7]. This strain allowed us to separate reads from the mitochondrial and nuclear genomes, including potential nuclear insertions of mitochondrial DNA (NuMTs). Sequence alignments of

the 29 repeat units from the *Nv* AsymCx strain generated a 217 bp consensus sequence, which was further aligned to the consensus sequences of four other strains with *Nv* mtDNA (Table S5). The coding probability of the repeat unit was inferred using CAPT v3.0.2 [40]. Read coverage in the repeat regions was quantified to estimate the repeat unit copy number within the five *Nv* mtDNA strains (AsymCx, LabII, V12.1, R5-11, and R16A). GC-content, CpG percentages and ratios of observed/expected CpGs were calculated for the mtDNA tandem repeats (Table S6) and non-control region. DNA methylation percentages were quantified at CpG positions in the repeat units using adult whole-body whole-genome bisulfite sequencing (WGBS-seq) data [10] (Table S7).

We used the pairwise BLASTn program to examine the sequence similarity pattern of the mitochondrial DNA control region with *Nv* mtDNA repeat array. The dot-matrix plot was generated using the R package v3.8. Potential DNA structures were predicted by using the Mfold DNA folding server [41], in which the folding temperature is set to 25 °C, given that insects are cold-blooded animals.

3. Results

3.1. Mitochondrial genome assembly and annotation in five jewel wasp species.

By using a linked-reads technology, we assembled the mitochondrial (MT) genomes of ten strains of six related parasitoid species, including *N. vitripennis* (*Nv*), *N. giraulti* (*Ng*), *N. longicornis* (*Nl*), *N. oneida* (*No*), and the outgroup species *Trichomalopsis sarcophagae* (*Ts*) and *Muscidifurax raptorellus* (*Mell*) (Table 1). The *Nv* AsymCx strain mtDNA was circularized using additional information from PSR1.1 genome ONT reads (see Materials and Methods). The *Nv* MT genome size is 22,956 bp (GenBank accession number MT635402). The mitochondrial genome size of *Ng*, *No*, *Nl* and *Ts* is 15,310 bp (MT611422), 14,811 bp (MT762278), 14,925 bp (MT755966) and 15,042 bp (MT611423), respectively. The A-T rich control regions were not closed in these four MT genomes, and the size esitmates were based on incomplete control regions (Table 1). It is noted that the GC-content and CpG density of *Nv* mitochondria are significantly higher than other species (Table 1), which was not driven by differences in the genic region.

In all assembled wasp mitochondrial genomes, we successfully identified the two rRNA genes (12S and 16S rRNA) and all 13 protein-coding genes encoding the hydrophobic subunits of respiratory chain complexes, including *atp6*, *atp8*, *cob*, *cox1*, *cox2*, *cox3*, *nad1*, *nad2*, *nad3*, *nad4*, *nad4l*, *nad5*, and *nad6* (Figure 1). In addition, 22 tRNAs were annotated in *Nv*, *Ng*, and *Ts*, as well as 21 tRNAs were identified in *Nl* (missing trnQ) and *No* (missing trnM). The gene coordinates in each mitochondrial genome are listed in Table 2. Gene order and protein length are conserved among these related parasitoid species.

3.2. Significant mitochondrial genome rearrangement in *Nasonia* and *Trichomalopsis* compared to other insects.

Although the MT gene content is conserved in insect species, there is substantial gene order rearrangement in different insect orders. Parasitoid wasps share a large inversion event encompassing six protein-coding genes (*nad3-cox3-atp6-atp8-cox2-cox1*) [18] compared to the ancestral insect MT genome [42] (Figure 1A). With the newly assembled MT genomes, we discovered two additional major MT gene order rearrangements in *Ts* and the *Nasonia* genus compared to other Hymenopteran clades (Figure 1B-D). Previous comparative studies assumed the gene order is the same in all jewel wasps [18], because the *Nasonia* MT genome was assembled as two fragments. In this study, the complete *Nasonia* MT genome assembly revealed the major *Nasonia-Ts*-specific inversion event involving 12 of the 13 protein-coding genes compared to Chalcidoidea (Figure 1D-E), followed by another inversion of the *nad2* gene and four neighboring tRNAs (Figure 1E-F). Gene order analysis in two outgroup species *Mell* and *Pteromalus puparum* (*Ppup*) [18] suggested this event occurred after the 6-gene inversion in all parasitoids, and it is *Nasonia-Ts*-specific assuming parsimony (Figure 1F-J). The rapid gene organization evolution we observed in the *Nasonia* MT genome can be explained by intrachromosomal recombination.

3.3. Phylogenetic analysis of protein-coding genes in five mitochondrial genomes.

We concatenated the coding sequences of all 13 protein-coding genes in each species and performed phylogenetic analyses to investigate the evolutionary relationship of mitochondrial genomes of these five wasp species. The phylogenetic tree of concatenated sequences confirms a similar evolutionary relationship of the five wasp species to the phylogeny based on nuclear

genes, but with a much faster evolutionary rate (Figure 2) [9]. We also conducted phylogenetic analyses on each individual gene, and three tree topologies were discovered (Figure S1 and Table S4). Five genes, *cox2*, *nad4*, *nad2*, *nad5*, and *cob*, show similar patterns as the concatenated mitochondrial sequences (Figure S1A). Six MT genes, *atp6*, *atp8*, *cox3*, *nad3*, *cox1*, and *nad6*, have a tree topology with *Nv* and *Ts* on the same branch (Figure S1B). Lastly, *nad4l* and *nad1* have *Nv* placed farther from the three *Nasonia* species than *Ts*. In these trees, *Ts* was more closely related to *Ng*, *No*, and *Nl* (Figure S1C). A prior nuclear gene-based phylogenetic analysis of the genera used [43] here to support the finding that *T. sarcophagae* is closely related to the *Nasonia* clade, and provide an estimated divergence time of approximately 2.6 million years ago, and 4.9 million years ago for *Muscidifurax*. Discrepancies in topology for the different genes will require future analysis of the pattern of selection and divergence in the different gene sets.

3.4. Characterization of palindromic tandem repeat array in the *Nv* mitochondrial genome.

An intriguing finding in *Nv* mitochondrial genome is an unusually large control region containing two extremely GC-rich and highly similar tandem repeat arrays (repeatR and repeatF, see Figure 3A). The individual repeat unit is 217 bp in length. The 3,767 bp repeatR array is upstream of the 12S rRNA gene, containing 16 complete and two partial repeat units (Figure 3B). The 2,317 bp repeatF array is downstream the *cob* gene with nine full-length and two partial repeat units in it (Figure 3B). Immediately adjacent to the two repeat arrays are two identical 533 bp A-T rich palindromic arms, which are connected by a sequence of 515 bp consisting of (AC)n and (AT)n simple dinucleotide repeats (Figure 3).

The *Nv* MT tandem repeats were not found in any of the other closely related *Nasonia* species, nor in *Ts* or *Mell*, with zero reads mapped to the repeat region (Figure 4A), despite the high sequencing depth in all species (Figure 4B). BLASTn search against the NCBI NR database did not retrieve any other sequences than *Nv* MT itself at an E-value cut-off of $1x10^{-2}$. The coding probability of these tandem repeat units is 0.047, and it was inferred to be noncoding. BLASTn identified a 645 bp ORF, which encodes a hypothetical protein LOC107981453 annotated as *cob*-like by the NCBI automated eukaryotic genome annotation pipeline. However, this prediction was labeled as low quality, and close examination found that this misannotation was due to partial sequence from the *cob* gene immediately upstream of the repeats. Therefore, these repeat units have minimal protein-coding ability.

To confirm these repeats are truly specific to the *Nv* MT genome, we compared the repeat coverage profile in *Nv*, *Ng*, and an MT replacement strain R16A (*Ng* nuclear and *Nv* MT background). *Ng* has zero repeat read depth, indicating the absence of these repeats in the *Ng* nuclear genome or *Ng* MT genome (Figure 4A), despite extensive genome sequencing depth in all parasitoid genomes (Figure 4B). R16A has similar repeat coverage, which indicates that the repeats are only present in its MT genome (Figure 4A), which is the MT genome from *Nv*. Thus, we could exclude the possibility that the aligned reads may come from the nuclear genome. The 8 Kb species-specific repeat region only exists in the *Nasonia vitripennis,* and the sister species of *Nasonia* and *Ts* lack such features in their MT genomes.

To quantify the repeat copy number and relative abundance, we calculated the read counts in three separate regions in the MT genome, the non-control region, forward repeat region

(repeatF), and reverse repeat region (repeatR). Four of the five *Nv* strains, AsymCx, LabII, V12.1, and R16A, have 702,006~943,354 reads mapped to the repeat regions, and the estimated total MT repeat length is 308-547 Kb, which is 0.15%-0.27% of the entire nuclear genome length (Table S1). The R5-11 strain has 161 Kb MT repeats per haploid nuclear genome, about half compared to other *Nv* strains (Table S1). The variations in total repeat length are not due to copy number variations within the MT genome. The repeatR array, repeatF array, and the non-control region have similar coverage within strains, indicating that the repeat copy number is conserved within species (Figure 4C). One exception is the LabII strain, in which the repeat region coverage is three times higher than the non-control region, suggesting potential recent copy number expansion (Figure 4C). This is also not found in the AsymCx strain, which was derived from LabII approximately 30 years ago or approximately 780 generations [44], suggesting that the expansion occurred after that time.

3.5. The *Nv* MT repeats are extremely GC-rich and CpG rich but lack DNA methylation.

Overall, *Nv* has a much higher GC-content (25.8%) and percent CpG (5.9%) compared to the mtDNA of other parasitoid species examined in this study (Table 1), and this is driven entirely by the repeat arrays. The non-repetitive region of the *Nv* mitochondrial genome has a low GC-content (16.5%) and low percent CpG (1.1%), which is similar to that of the other mitochondrial genomes (Table 1). In contrast, the repeat unit is extremely GC-rich (50.7%), which is even higher than the *Nv* nuclear genome (41.7%). Furthermore, the average ratio of observed/expected CpGs of the whole repetitive region is 1.61 (Table S6), which is much higher than that of the non-repetitive region (0.80). The high ratio of observed/expected CpGs is caused by the large number of CpGs in the repeats. In *Nasonia*, CpG clusters tend to be methylated in the nuclear

genome, but genes with high observed/expected CpG ratio are often non-methylated [10]. We checked the methylation status at all CpG sites in these repeats and found no methylation on the region (Table S7).

3.6. Sequence and structural characteristics of the *Nv* MT control region.

We next sought to examine the sequence and structural characteristics of this *Nv* MT repeat region. Pairwise BLASTn comparisons of the region revealed a striking similarity pattern with two blocks of tandem repeats, as indicated by diagonal lines and reverse diagonal lines (Figure 5A). This pattern is a hallmark of an ultra-sized palindromic DNA sequence of 7,665 base pairs. A palindrome of a double-stranded DNA/RNA sequence is defined by containing complementary strands being palindromic of each other. One of its important characteristics is the potential of forming single-stranded stem-loop cruciform structures [45]. Indeed, we confirmed the presence of such super-cruciform structures in the *Nv* MT repeat arrays using Mfold [41]. The two strands in the normal double-stranded state can separate to form a super-cruciform structure, with a new double helix structure formed by the same strand from the upstream portion (armF+repeatF) and the downstream portion (armR+repeatR) (Figure 5B). The un-paired central regions will become single-stranded loops. Additionally, as there are uneven numbers of repeatF (11 units) and repeatR (18 units), the pairing pattern between repeat units could vary, which would result in up to seven repeat units from the downstream portion being unpaired (Figure 5B). As single-stranded repeats are not stable, we wanted to know whether it can be further folded into stable secondary structures. Indeed, by using the minimum energy folding prediction, we found the single repeat unit can also self fold into a stable hairpin (Figure 5C), as supported by the top three thermostable structures. Thus, we show that the control region

of *Nv* MT is an ultra-sized DNA palindrome, with the potential to form a super-cruciform structure. In this structure, the stems are made from the repeat regions (palindromic arm and repeats), and the loops consist of the unpaired central (AC)n and (AT)n simple dinucleotide repeats, with many unpaired repeat units protruding as standalone hairpins (Figure 5B).

3.7. Homogenization among the *Nv* MT repeats.

From the 29 tandem repeat units in *Nv* AsymCx laboratory strain, we identified 217 bp consensus sequence and compared it to other lab strains V12.1 and R16A mitochondrial genomes, as well as their original field strains LabII and R5-11, through sequencing read alignments (see Materials and Methods). R16A and LabII have an identical consensus as AsymCx (Figure 6), which is consistent with the fact that AsymCx and R16A are both derived from LabII with the Netherlands origin. The North American strains R5-11 and V12.1 have four substitutions in the consensus sequence, which is highly similar to the AsymCx strain (98.2% identity). We also investigated the 29 individual repeat units AsymCx, and the sequence identity is very high (98.6%-100%), indicating the high sequence homogenization among these tandem repeats (Figure 6). There are three singletons and four shared segregating polymorphisms among different tandem repeat units (Figure 6). Tandem arrays undergo homogenization by various processes, including unequal recombination and rolling circle replication [46-48]. In addition, palindromic arms tend to have abundant gene conversion events, which can repair the point mutations and lower the divergence between the palindromic sequences [51, 52]. Indeed, we discovered potential recombination-associated sequence motifs within the *Nv* MT repeat (Figure S2), similar to the *E. coli* Chi recombination hotspot sequence [53], and known sequences associated with recombination in the human nuclear genome [54] and mtDNA [55]. The two

palindromic repeat arrays (repeatF and repeatR) can pair into a super-cruciform structure, allowing the gene conversion events to occur (Figure 5B). We discovered compensatory variations in the repeat monomers near the loop structure (units 1-2-3-4 and 29-28-27-26) (Figure 5B and Figure 6), providing evidence for repair after gene conversions.

4. Discussion

4.1. Major inversion events in *Nasonia* MT genomes compared to other wasp species.

The importance of mitochondria in cellular energy production makes it a revealing research topic in biology. As the mitochondrion has its own independent genome, an assembled MT genome may provide novel insights into the biological function, development, and evolution of mitochondrion. The *Nv* reference genome paper published in 2010 [9] reported two MT fragments; however, the complete MT genome assembly is still missing. In this study, we reported high-quality assemblies of MT genomes in five jewel wasp species, including *N. vitripennis*, *N. giraulti*, *N. longicornis*, *N. oneida*, and *Trichomalopsis sarcophagae*. We also revised the *Nasonia* genus' gene order, which was incorrectly inferred in a previous assembly of the *Pteromalus puparum* MT genome due to incomplete *Nv* MT genome [18]. Comparing to the insect ancestral MT genome, the parasitoid *P. puparum* and *M. raptorellus* genomes underwent a six-gene inversion event. Interestingly, we discovered a major subsequent inversion involving 12 protein-coding genes (from *cob* to *nad2*) in *Nasonia* and *Trichomalopsis*, which occurred after they diverged from *Muscidifurax*. This is unusual because most MT gene order alterations in insects do not involve *nad2*, located next to the control region. In the *Nasonia* genus, the protein-coding gene at the coding-control region boundary is *cob*. Orientations of *nad2* and surrounding tRNA genes underwent a subsequent inversion event, which has led to this genus-specific gene order. Interestingly, all inversions occurred in the noncoding regions, either within a tRNA gene cluster or near the control region boundary (Figure 1), suggesting that secondary structures played a role in mediating these events. Similar to the repeats. Our findings indicate a rapid MT genome reorganization in *Nasonia* and *Trichomalopsis*.

4.2. Phylogenetic relationships of mitochondrial genomes in the *Nasonia* genus.

The phylogenetic analysis of concatenated protein-coding sequences in five wasp species indicated that the mitochondrial gene tree and nuclear gene tree have the same topology, in which *Ng* and *No* are the closest sister species with *Nl* in the near branch and *Nv* in the further away branch [9]. Interestingly, the *Ts* node for 8 MT gene trees is located in between *Nv* and other *Nasonia* species, and the bootstrap support is only 81 for *Ts* being the outgroup in the MT concatenated tree (Figure 2). In the nuclear gene tree, the support of *Ts* as an outgroup was not significant (bootstrap value less than 50). The finding is consistent with what is known in the previous studies, in which *Ts* is closely related to the *Nasonia* genus with a divergence time of approximately 2.6 million years ago [43, 56].

For the sister species pair *Ng* and *No*, their nuclear genome divergence is about half of the *Ng*-*Nl* divergence. In contrast, *Ng* and *No* MT genomes are highly similar, and the divergence is only 10% of *Ng*-*Nl* MT divergence (Figure 2). This discrepancy suggests recent introgression between *Ng* and *No* MT genomes, presumably through recent hybridization. This has been reported previously [6], but we provide genome-level evidence here. We found rapid evolution of *atp8* in *Nv*, based on the fact that *Ts* is more closely grouped with other *Nasonia* species in its MT gene tree, which is also consistent with previous research [12].

4.3. A unique GC-rich tandem repeat array exists in the *Nasonia vitripennis* MT genome.

The *Nv* MT genome is unique in many ways. The genome size is 23 Kb, an outlier relative to the typical insect genome range. The GC-content is 25.8%, which is significantly higher than other closely related wasp genomes (16.9% on average). These striking differences are not due to the

genic regions, but the presence of tandem repeat arrays in the mtDNA control region. The control regions are known to contain tandem repeats in several species, whose nucleotide composition of this region was extremely AT-rich [21]. These tandem repeats vary widely in length from 4~82 bp in a moth species *Epirrita autumnata* [57], to 196 bp in bees [58], and 0.8~2.0 Kb in bark weevils [59]. Recently, an even larger control region was found in the mitochondrial genome of *Pteromalus puparum* in the same Pteromalidae family as *Nasonia*, which also contains a palindromic tandem repeat region [18]. However, the *Nv* control region repeats are 2.5 times larger than those in *P. puparum.*

4.4. The origin of the tandem repeat arrays in the *Nasonia vitripennis* MT genome.

The GC-rich MT repeats we identified in *N. vitripennis* MT genome has yet to be observed in any other bilaterians, including closely related *Nasonia* species or *T. sarcophagae.* Although the repeat region is specific to *Nv*, MT coding region is relatively conserved among *Nasonia* species and *Ts*, in terms of protein sequence, gene order, and GC-content. This suggests that the *Nv* repeat arrays were acquired by *N. vitripennis* after its speciation from the common ancestors of the other three *Nasonia* species, which was estimated to be 1.6 million years ago [43]. Five *Nv* strains isolated from different geographic locations share similar repeat copy numbers, suggesting the repeat arrays originated and fixed in the *Nv* population before the global radiation.

Since we cannot find any other similar repeats beyond *Nv* species, the exact source of the repeats was unclear. To search for a potential source of these repeats, we aligned the repeat unit against all assembled genomes in NCBI and identified a 291 bp scaffold (accession number AERW01000434) in the *Nv Wolbachia* endosymbiont *w*VitB assembly, which is a B supergroup

Wolbachia strain in *Nv* and contains 1.34 repeat units with 98.97% sequence identity [60]. The *Wolbachia pipientis* genome is enriched for repetitive sequences, including palindromic repeats [61]. All *Nv* laboratory strains had *Wolbachia* infections when they were isolated from the field. Transfer of *Wolbachia* repeat sequences into mtDNA could be mediated by *Wolbachia* plasmid, which was discovered recently [62]. One plausible explanation is that the *Nv* MT came from the *Wolbachia* endosymbiont genome or *Wolbachia* plasmid genome. However, we cannot exclude the possibility that the *w*VitB genome assembly has some MT sequence contamination during the genomic DNA purification [63, 64]. Otherwise, the source of this GC-rich insertion remains a mystery.

4.5. Biased gene conversion may result in the extremely high GC-content in the Nv MT repeats.

The MT tandem repeats we discovered in this research are extremely GC-enrich (50.7%) compared to *P. puparum* ones (16.5% GC), which is even higher than the nuclear genome GC-content (40.6%) [16]. A possible explanation for the enriched GC-content is GC-biased gene conversion (gBCG) [65, 66]. GC-biased gene conversion arises from the fact that the repair mechanism for the double-strand breaks within the palindromic sequences preferentially uses GC nucleotides as the template. Long palindromic sequences can induce double-strand breaks, which will be resolved by recombination or gene conversion [67]. gBCG frequently occurs within the palindromic sequences with a higher chance of recombination, such as mammalian Y-chromosome and plastome palindromes [68-70]. In these cases, the palindromic sequences have been observed with higher than normal GC-content. The *Nv* MT repeat array can fold into large palindromic structures (Figure 5), which is prone to gBCG. This will not only result in the

homogenization of the inverted repeats but also the increased GC-content. Therefore, we propose that GC-enrichment evolved from a low GC sequence context through GC-biased gene conversion. However, we could not exclude an alternative possibility, which is the repeat sequences came from an unknown source of origin with high GC-content at the time of acquisition.

4.6. Potential function and evolutionary fixiation of the repeat array in the *Nv* MT genome.

Here we proposed two potential explanations for the evolutionary fixation of the large repeat arrays causing a 50% increase in the MT genome size, which is a considerable burden to an A-T rich MT genome. One possibility is that the repeat array has converted the ancestral replication origin (*ori*) in *Nv* into a more efficient version, which facilitates the repeat array-containing MT genomes to outcompete the original control region and rapidly reached fixation in the population. Compared to other *Nasonia* species (*Ng*, *No*, and *Nl*), which have specialized hosts in nature with much smaller habitat in the US, *Nv* is a cosmopolitan species found worldwide with broader host preferences. The fitness advantages of the repeat-containing MT genome may explain the fixation of the repeat arrays in the *Nv* MT genome. Another plausible explanation is that the MT repeat array may function as a molecular driver [71, 72] to convert the MT genomes through an unknown mechanism. These two mechanisms are not mutually exclusive. The repeat array can serve as a strong molecular driver with a fitness advantage to *Nv*, resulting in rapid fixation and maintenance of this repeat array in the MT genome.

4.7. Possible mechanisms for copy number expansion in the *Nv* MT repeat arrays.

Replication slippage, rolling-circle replication, and unequal sister chromatid exchange are considered to be the primary mechanism for the expansion and maintenance of tandem repeats in the genome [21, 47, 73]. However, slippage replication typically occurs in the case of short tandem repeats [74]. It has been predicted that rolling-circle replication can lead to the accumulation of "relic" sequences at the ends of tandem repeat arrays and greater homogenization within the central region [47, 49, 50]. We tested this hypothesis by examining the distribution of polymorphisms distributed across the repeat arrays and did not observed significant differences in divergence. Therefore, it is likely the third mechanism, unequal exchanges in recombination [55], have contributed to the expansion and maintenance of this unusual repeat array. The *Nv* MT repeat unit contains sequence motifs that are similar to known sequences associated with recombination in *E. coli* [53], human mtDNA [55], and human nuclear genome [54] sequences (Figure S2). The hairpin and cruciform structure formed by palindromic sequences could often induce double-strand breaks [75], and the subsequent repair mechanism can lead to gene conversion and homologous recombination [54, 76]. The intra-chromosomal homologous recombination between the two arms of the palindromic repeat regions can facilitate unequal crossovers within the mitochondrial chromosome, causing the repeat array expansion. The hypothesis is supported by the compensatory nucleotide variations in the repeats (Figure 6), which could provide an explanation for how the control region in *Nv* has expanded to such a length that it has now increased MT genome size by 50%. However, the evolutionary origin, maintenance, and potential functions of the *Nv*-specific repeat arrays still warrant further study.

Acknowledgments

This project is supported by an Auburn University Intramural Grant Program Award to X.W. (AUIGP 180271) and the USDA National Institute of Food and Agriculture Hatch project 1018100. X.W. is supported by National Science Foundation EPSCoR RII Track-4 Research Fellowship (NSF OIA 1928770), an Alabama Agriculture Experiment Station (AAES) ARES Agriculture Research Enhancement, Exploration and Development (AgR-SEED) award, and a laboratory start-up fund from Auburn University College Veterinary Medicine. Z.J.L. is supported by an NSF-REU award 1560115. J.W. is supported by Shanghai Academy of Agricultural Sciences Program for Excellent Research Team visiting scholars. Y.T. and D.Z. are supported by the Saint Louis University start-up fund. The authors thank the Auburn Hopper supercomputer clusters for computational support.

Reference

[1] M.R. Strand, Editorial overview: Parasites/parasitoids/biological control: The study of parasitoid physiology begins to mature, Curr Opin Insect Sci 6 (2014) vi-ix.

[2] A. Kapranas, A. Tena, Encyrtid parasitoids of soft scale insects: biology, behavior, and their use in biological control, Annu Rev Entomol 60 (2015) 195-211.

[3] J.A. Lynch, The Expanding Genetic Toolbox of the Wasp Nasonia vitripennis and Its Relatives, Genetics 199(4) (2015) 897-904.

[4] J.H. Werren, D.W. Loehlin, The parasitoid wasp Nasonia: an emerging model system with haploid male genetics, Cold Spring Harb Protoc 2009(10) (2009) pdb emo134.

[5] L. Beukeboom, C. Desplan, Nasonia, Curr Biol 13(22) (2003) R860-R860.

[6] R. Raychoudhury, C.A. Desjardins, J. Buellesbach, D.W. Loehlin, B.K. Grillenberger, L. Beukeboom, T. Schmitt, J.H. Werren, Behavioral and genetic characteristics of a new species of Nasonia, Heredity 104(3) (2010) 278-288.

[7] J.A.J. Breeuwer, J.H. Werren, Hybrid Breakdown between Two Haplodiploid Species: The Role of Nuclear and Cytoplasmic Genes, Evolution 49(4) (1995) 705-717.

[8] X. Wang, J.H. Werren, A.G. Clark, Genetic and epigenetic architecture of sex-biased expression in the jewel wasps Nasonia vitripennis and giraulti, P Natl Acad Sci USA 112(27) (2015) E3545-E3554.

[9] L. Viljakainen, D.C. Oliveira, J.H. Werren, S.K. Behura, Transfers of mitochondrial DNA to the nuclear genome in the wasp Nasonia vitripennis, Insect molecular biology 19 Suppl 1 (2010) 27-35.

[10] X. Wang, D. Wheeler, A. Avery, A. Rago, J.H. Choi, J.K. Colbourne, A.G. Clark, J.H. Werren, Function and Evolution of DNA Methylation in Nasonia vitripennis, Plos Genet 9(10) (2013).

[11] X. Wang, J.H. Werren, A.G. Clark, Allele-Specific Transcriptome and Methylome Analysis Reveals Stable Inheritance and Cis-Regulation of DNA Methylation in Nasonia, PLoS Biol 14(7) (2016) e1002500.

[12] D.C.S.G. Oliveira, R. Raychoudhury, D.V. Lavrov, J.H. Werren, Rapidly evolving mitochondrial genome and directional selection in mitochondrial genes in the parasitic wasp Nasonia (Hymenoptera : Pteromalidae), Mol Biol Evol 25(10) (2008) 2167-2180.

[13] D.M. Rand, R.A. Haney, A.J. Fry, Cytonuclear coevolution: the genomics of cooperation, Trends Ecol Evol 19(12) (2004) 645-653.

[14] Z.C. Yan, G.Y. Ye, J.H. Werren, Evolutionary Rate Correlation between Mitochondrial-Encoded and Mitochondria-Associated Nuclear-Encoded Proteins in Insects, Mol Biol Evol 36(5) (2019) 1022-1036.

[15] C.K. Ellison, O. Niehuis, J. Gadau, Hybrid breakdown and mitochondrial dysfunction in

hybrids of Nasonia parasitoid wasps, J Evol Biol 21(6) (2008) 1844-51.

[16] J.H. Werren, S. Richards, C.A. Desjardins, O. Niehuis, J. Gadau, J.K. Colbourne, G. Nasonia Genome Working, J.H. Werren, S. Richards, C.A. Desjardins, O. Niehuis, J. Gadau, J.K. Colbourne, L.W. Beukeboom, C. Desplan, C.G. Elsik, C.J. Grimmelikhuijzen, P. Kitts, J.A. Lynch, T. Murphy, D.C. Oliveira, C.D. Smith, L. van de Zande, K.C. Worley, E.M. Zdobnov, M. Aerts, S. Albert, V.H. Anaya, J.M. Anzola, A.R. Barchuk, S.K. Behura, A.N. Bera, M.R. Berenbaum, R.C. Bertossa, M.M. Bitondi, S.R. Bordenstein, P. Bork, E. Bornberg-Bauer, M. Brunain, G. Cazzamali, L. Chaboub, J. Chacko, D. Chavez, C.P. Childers, J.H. Choi, M.E. Clark, C. Claudianos, R.A. Clinton, A.G. Cree, A.S. Cristino, P.M. Dang, A.C. Darby, D.C. de Graaf, B. Devreese, H.H. Dinh, R. Edwards, N. Elango, E. Elhaik, O. Ermolaeva, J.D. Evans, S. Foret, G.R. Fowler, D. Gerlach, J.D. Gibson, D.G. Gilbert, D. Graur, S. Grunder, D.E. Hagen, Y. Han, F. Hauser, D. Hultmark, H.C.t. Hunter, G.D. Hurst, S.N. Jhangian, H. Jiang, R.M. Johnson, A.K. Jones, T. Junier, T. Kadowaki, A. Kamping, Y. Kapustin, B. Kechavarzi, J. Kim, J. Kim, B. Kiryutin, T. Koevoets, C.L. Kovar, E.V. Kriventseva, R. Kucharski, H. Lee, S.L. Lee, K. Lees, L.R. Lewis, D.W. Loehlin, J.M. Logsdon, Jr., J.A. Lopez, R.J. Lozado, D. Maglott, R. Maleszka, A. Mayampurath, D.J. Mazur, M.A. McClure, A.D. Moore, M.B. Morgan, J. Muller, M.C. Munoz-Torres, D.M. Muzny, L.V. Nazareth, S. Neupert, N.B. Nguyen, F.M. Nunes, J.G. Oakeshott, G.O. Okwuonu, B.A. Pannebakker, V.R. Pejaver, Z. Peng, S.C. Pratt, R. Predel, L.L. Pu, H. Ranson, R. Raychoudhury, A. Rechtsteiner, J.T. Reese, J.G. Reid, M. Riddle, H.M. Robertson, J. Romero-Severson, M. Rosenberg, T.B. Sackton, D.B. Sattelle, H. Schluns, T. Schmitt, M. Schneider, A. Schuler, A.M. Schurko, D.M. Shuker, Z.L. Simoes, S. Sinha, Z. Smith, V. Solovyev, A. Souvorov, A. Springauf, E. Stafflinger, D.E. Stage, M. Stanke, Y. Tanaka, A. Telschow, C. Trent, S. Vattathil, E.C. Verhulst, L. Viljakainen, K.W. Wanner, R.M. Waterhouse, J.B. Whitfield, T.E. Wilkes, M. Williamson, J.H. Willis, F. Wolschin, S. Wyder, T. Yamada, S.V. Yi, C.N. Zecher, L. Zhang, R.A. Gibbs, Functional and evolutionary insights from the genomes of three parasitoid Nasonia species, Science 327(5963) (2010) 343-8.

[17] J.D. Gibson, O. Niehuis, B.R. Peirson, E.I. Cash, J. Gadau, Genetic and developmental basis of F2 hybrid breakdown in Nasonia parasitoid wasps, Evolution 67(7) (2013) 2124-32.

[18] Z.C. Yan, Q. Fan, Y. Tian, F. Wang, X.X. Chen, J.H. Werren, G.Y. Ye, Mitochondrial DNA and their nuclear copies in the parasitic wasp Pteromalus puparum: A comparative analysis in Chalcidoidea, Int J Biol Macromol 121 (2019) 572-579.

[19] D.A. Clayton, Replication of Animal Mitochondrial-DNA, Cell 28(4) (1982) 693-705.

[20] D.A. Clayton, Replication and Transcription of Vertebrate Mitochondrial-DNA, Annu Rev Cell Biol 7 (1991) 453-478.

[21] D.X. Zhang, G.M. Hewitt, Insect mitochondrial control region: A review of its structure, evolution and usefulness in evolutionary studies, Biochem Syst Ecol 25(2) (1997) 99-120.

[22] R.E. Broughton, T.E. Dowling, Length Variation in Mitochondrial-DNA of the Minnow Cyprinella-Spiloptera, Genetics 138(1) (1994) 179-190.

[23] N.E. Buroker, J.R. Brown, T.A. Gilbert, P.J. Ohara, A.T. Beckenbach, W.K. Thomas, M.J. Smith, Length Heteroplasmy of Sturgeon Mitochondrial-DNA - an Illegitimate Elongation Model, Genetics 124(1) (1990) 157-163.

[24] K. Hayasaka, T. Ishida, S. Horai, Heteroplasmy and Polymorphism in the Major Noncoding Region of Mitochondrial-DNA in Japanese Monkeys - Association with Tandemly Repeated Sequences, Mol Biol Evol 8(4) (1991) 399-415.

[25] D.M. Rand, R.G. Harrison, Molecular Population-Genetics of Mtdna Size Variation in Crickets, Genetics 121(3) (1989) 551-569.

[26] M.J. Perrot-Minnot, L.R. Guo, J.H. Werren, Single and double infections with Wolbachia in the parasitic wasp Nasonia vitripennis: effects on compatibility, Genetics 143(2) (1996) 961-72.

[27] X. Wang, Y.D. Kelkar, X. Xiong, E.O. Martinson, J. Lynch, C. Zhang, J.H. Werren, X. Wang, Genome Report: Whole Genome Sequence and Annotation of the Parasitoid Jewel Wasp Nasonia giraulti Laboratory Strain RV2X[u], G3 (Bethesda) 10(8) (2020) 2565-2572.

[28] X. Wang, X. Xiong, W. Cao, C. Zhang, J.H. Werren, X. Wang, Genome Assembly of the A-Group Wolbachia in Nasonia oneida Using Linked-Reads Technology, Genome Biol Evol 11(10) (2019) 3008-3013.

[29] S. Andrews, FastQC: a quality control tool for high throughput sequence data, 2010.

[30] N.I. Weisenfeld, V. Kumar, P. Shah, D.M. Church, D.B. Jaffe, Direct determination of diploid genome sequences (vol 27, pg 757, 2017), Genome Res 28(4) (2018) 606-606.

[31] W.J. Kent, BLAT - The BLAST-like alignment tool, Genome Res 12(4) (2002) 656-664.

[32] E. Dalla Benetta, I. Antoshechkin, T. Yang, H.Q.M. Nguyen, P.M. Ferree, O.S. Akbari, Genome elimination mediated by gene expression from a selfish chromosome, Sci Adv 6(14) (2020) eaaz9808.

[33] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform, Bioinformatics 25(14) (2009) 1754-60.

[34] D. Li, C.M. Liu, R. Luo, K. Sadakane, T.W. Lam, MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph, Bioinformatics 31(10) (2015) 1674-6.

[35] M. Kearse, R. Moir, A. Wilson, S. Stones-Havas, M. Cheung, S. Sturrock, S. Buxton, A. Cooper, S. Markowitz, C. Duran, T. Thierer, B. Ashton, P. Meintjes, A. Drummond, Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data, Bioinformatics 28(12) (2012) 1647-1649.

[36] M. Bernt, A. Donath, F. Juhling, F. Externbrink, C. Florentz, G. Fritzsch, J. Putz, M. Middendorf, P.F. Stadler, MITOS: Improved de novo metazoan mitochondrial genome annotation, Mol Phylogenet Evol 69(2) (2013) 313-319.

[37] R.C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Res 32(5) (2004) 1792-1797.

[38] S. Kumar, G. Stecher, M. Li, C. Knyaz, K. Tamura, MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms, Mol Biol Evol 35(6) (2018) 1547-1549.

[39] A. Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies, Bioinformatics 30(9) (2014) 1312-3.

[40] L. Wang, H.J. Park, S. Dasari, S. Wang, J.P. Kocher, W. Li, CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model, Nucleic Acids Res 41(6) (2013) e74.

[41] M. Zuker, Mfold web server for nucleic acid folding and hybridization prediction, Nucleic Acids Res 31(13) (2003) 3406-15.

[42] S.L. Cameron, Insect mitochondrial genomics: implications for evolution and phylogeny, Annu Rev Entomol 59 (2014) 95-117.

[43] E.O. Martinson, Mrinalini, Y.D. Kelkar, C.H. Chang, J.H. Werren, The Evolution of Venom by Co-option of Single-Copy Genes, Curr Biol 27(13) (2017) 2007-2013 e8.

[44] J.A.J. Breeuwer, J.H. Werren, Microorganisms associated with chromosome destruction and reproductive isolation between two insect species, Nature 346(6284) (1990) 558-560.

[45] K. Mizuuchi, M. Mizuuchi, M. Gellert, Cruciform structures in palindromic DNA are favored by DNA supercoiling, J Mol Biol 156(2) (1982) 229-43.

[46] G.P. Smith, Evolution of repeated DNA sequences by unequal crossover, Science 191(4227) (1976) 528-35.

[47] B. Charlesworth, P. Sniegowski, W. Stephan, The evolutionary dynamics of repetitive DNA in eukaryotes, Nature 371(6494) (1994) 215-20.

[48] G. Dover, Concerted evolution, molecular drive and natural selection, Curr Biol 4(12) (1994) 1165-6.

[49] B.F. McAllister, J.H. Werren, Evolution of tandemly repeated sequences: What happens at the end of an array?, J Mol Evol 48(4) (1999) 469-81.

[50] D.E. Khost, D.G. Eickbush, A.M. Larracuente, Single-molecule sequencing resolves the detailed structure of complex satellite DNA loci in Drosophila melanogaster, Genome Res 27(5) (2017) 709-721.

[51] S. Rozen, H. Skaletsky, J.D. Marszalek, P.J. Minx, H.S. Cordum, R.H. Waterston, R.K. Wilson, D.C. Page, Abundant gene conversion between arms of palindromes in human and ape Y chromosomes, Nature 423(6942) (2003) 873-6.

[52] A. Helgason, A.W. Einarsson, V.B. Guethmundsdottir, A. Sigurethsson, E.D. Gunnarsdottir, A. Jagadeesan, S.S. Ebenesersdottir, A. Kong, K. Stefansson, The Y-chromosome point mutation rate in humans, Nat Genet 47(5) (2015) 453-7.

[53] S.K. Amundsen, G.R. Smith, Chi hotspot activity in Escherichia coli without RecBCD exonuclease activity: implications for the mechanism of recombination, Genetics 175(1) (2007) 41-54.

[54] A.J. Jeffreys, V. Wilson, S.L. Thein, Hypervariable 'minisatellite' regions in human DNA,

Nature 314(6006) (1985) 67-73.

[55] D.R. Johns, S.L. Rutledge, O.C. Stine, O. Hurko, Directly repeated sequences associated with pathogenic mitochondrial DNA deletions, Proc Natl Acad Sci U S A 86(20) (1989) 8059-62.

[56] B.C. Campbell, J.D. Steffen-Campbell, J.H. Werren, Phylogeny of the Nasonia species complex (Hymenoptera: Pteromalidae) inferred from an internal transcribed spacer (ITS2) and 28S rDNA sequences, Insect molecular biology 2(4) (1993) 225-37.

[57] N. Snall, K. Huoponen, M.L. Savontaus, K. Ruohomaki, Tandem repeats and length variation in the mitochondrial DNA control region of Epirrita autumnata (Lepidoptera : Geometridae), Genome 45(5) (2002) 855-861.

[58] J.M. Cornuet, L. Garnery, M. Solignac, Putative Origin and Function of the Intergenic Region between Coi and Coii of Apis-Mellifera L Mitochondrial-DNA, Genetics 128(2) (1991) 393-403.

[59] T.M. Boyce, M.E. Zwick, C.F. Aquadro, Mitochondrial-DNA in the Bark Weevils - Size, Structure and Heteroplasmy, Genetics 123(4) (1989) 825-836.

[60] B.N. Kent, L. Salichos, J.G. Gibbons, A. Rokas, I.L. Newton, M.E. Clark, S.R. Bordenstein, Complete bacteriophage transfer in a bacterial endosymbiont (Wolbachia) determined by targeted genome capture, Genome Biol Evol 3 (2011) 209-18.

[61] H. Ogata, K. Suhre, J.M. Claverie, Discovery of protein-coding palindromic repeats in Wolbachia, Trends Microbiol 13(6) (2005) 253-5.

[62] J. Reveillaud, S.R. Bordenstein, C. Cruaud, A. Shaiber, O.C. Esen, M. Weill, P. Makoundou, K. Lolans, A.R. Watson, I. Rakotoarivony, S.R. Bordenstein, A.M. Eren, The Wolbachia mobilome in Culex pipiens includes a putative plasmid, Nat Commun 10(1) (2019) 1051.

[63] J. Pascar, C.H. Chandler, A bioinformatics approach to identifying Wolbachia infections in arthropods, PeerJ 6 (2018) e5486.

[64] X. Wang, X. Xiong, W. Cao, C. Zhang, J.H. Werren, X. Wang, Phylogenomic analysis of Wolbachia strains reveals patterns of genome evolution and recombination, Genome Biology and Evolution (2020).

[65] C.F. Mugal, C.C. Weber, H. Ellegren, GC-biased gene conversion links the recombination landscape and demography to genomic base composition: GC-biased gene conversion drives genomic base composition across a wide range of species, Bioessays 37(12) (2015) 1317-26.

[66] E. Pessia, A. Popa, S. Mousset, C. Rezvoy, L. Duret, G.A. Marais, Evidence for widespread GC-biased gene conversion in eukaryotes, Genome Biol Evol 4(7) (2012) 675-82.

[67] F. Nasar, C. Jankowski, D.K. Nag, Long palindromic sequences induce double-strand breaks during meiosis in yeast, Mol Cell Biol 20(10) (2000) 3449-58.

[68] S. Glémin, P.F. Arndt, P.W. Messer, D. Petrov, N. Galtier, L. Duret, Quantification of GC-

biased gene conversion in the human genome, Genome Res 25(8) (2015) 1215-1228.

[69] Z. Niu, Q. Xue, H. Wang, X. Xie, S. Zhu, W. Liu, X. Ding, Mutational Biases and GC-Biased Gene Conversion Affect GC Content in the Plastomes of Dendrobium Genus, Int J Mol Sci 18(11) (2017) 2307.

[70] B. Trombetta, F. Cruciani, Y chromosome palindromes and gene conversion, Human Genetics 136(5) (2017) 605-619.

[71] G. Dover, Molecular drive: a cohesive mode of species evolution, Nature 299(5879) (1982) 111-7.

[72] G. Dover, Molecular drive, Trends Genet 18(11) (2002) 587-9.

[73] D.E. Axelrod, K.A. Baggerly, M. Kimmel, Gene amplification by unequal sister chromatid exchange: probabilistic modeling and analysis of drug resistance data, J Theor Biol 168(2) (1994) 151-9.

[74] H. Fan, J.Y. Chu, A brief review of short tandem repeat mutation, Genomics Proteomics Bioinformatics 5(1) (2007) 7-14.

[75] D.R. Leach, Long DNA palindromes, cruciform structures, genetic instability and secondary structure repair, Bioessays 16(12) (1994) 893-900.

[76] Z.-H. Zhou, E. Akgün, M. Jasin, Repeat expansion by homologous recombination in the mouse germ line at palindromic sequences, Proceedings of the National Academy of Sciences 98(15) (2001) 8326.

List of tables

Table 1. Mitochondrial genome assemblies of five wasp species.

Table 2. Mitochondrial gene annotations of five jewel wasp species.

Figure legends

Figure 1. Mitochondrial gene annotation and genome organization in seven wasp species and comparison to the ancestral insect mitochondrial genome.
Mitochondrial genome annotations of (A) insect ancestral mtDNA, adapted from Cameron 2014; (B) *Apis mellifera*, recreated from GenBank accession number MT859135.1; (C) *Ppup* (*Pteromalus puparum*), recreated from Yan et al. 2019; (D) *Muscidifurax raptorellus* (*Mell*); (E) a hypothetical ancestral *Nasonia* mtDNA; (F) *T. sarcophagae* (*Ts*); (G) *N. giraulti* (*Ng*); (H) *N. oneida* (*No*); (I) *N. longicornis* (*Nl*); (J) *N. vitripennis* (*Nv*). The *A*. *mellifera* gene order and orientations for protein-coding genes are the same as the ancestral insect mtDNA. Both the *Mell* and *Ppup* have identical gene order, which differs from the ancestral insect mtDNA by having an inversion between the tRNA genes trnE and trnC, and a translocation event of trnC. The exact location of the inversion event is less defined for the tRNAs when including *A. mellifera,* but it is clear that there is an inversion between *cox1* and *nad3* (Event I). The genic inversion is indicated by dark solid lines. (E) is the hypothetical ancestral *Nasonia* mtDNA intermediate between *Mell*/*Ppup* gene order and *Ts* and *Nasonia*. A single inversion event occurred between trnQ and trnM, which involves 12 protein-coding genes (Event II). Using the hypothetical intermediate mtDNA as a reference, there is an additional inversion that have occurred between trnQ and trnR (Event III). A simplified phylogenetic relationship is shown on the right, and the time of three inversion events is labeled.

Figure 2. Comparison of phylogenetic relationship of five jewel wasp species using nuclear genome and mitochondrial genome sequences.
The phylogenetic tree for mitochondrial genes is constructed using MEGA-X with 1000 bootstraps on concatenated protein-coding genes. The nuclear gene tree is generated using RAxML with 1,000 bootstraps on concatenated single-copy orthologs.

Figure 3. *Nasonia vitripennis* mitochondrial DNA complete sequence, gene, and tandem repeat array annotation.
The structure of the *Nasonia vitripennis* mitochondrial genome and zoomed-in and linear view of the control region. (A) The circular mitochondrial genome of *Nv*. Green indicates protein-coding genes; red indicates ribosomal RNAs, and pink indicates tRNAs. GC% of the entire genome is represented as a blue graph beneath the genome annotations. The gray circles serve as a reference for the GC% as 0%, 50%, and 100%. (B) The elaborated and linear view of the control region, including GC percentage. Yellow indicates the individual repeat units, and green indicates the palindromic arms that are complementary to each other.

Figure 4. *Nasonia vitripennis* repeat region coverage in five *Nv* MT strains and five other wasp species.

(A) Repeat read counts per million mapped reads to the *Nv* genome for ten jewel wasp strains. (B) Haploid nuclear genome coverage of all ten wasp strains. (C) Coverage of reads from all jewel wasps in all three regions of the *Nv* mitochondrial genome per 1X haploid nuclear genome coverage.

Figure 5. Sequence and structural features of *N. vitripennis* MT control region.
(A) Dot-matrix plot of pairwise BLASTn comparison analysis indicates the presence of an ultra-sized palindromic DNA sequence in the control region. (B) Models of representative structures of the control region, including the normal double-stranded state and two other potential states with super-cruciform structures. The blue arrows indicate the forward tandem repeats while the light blue arrows are complementary copies. The purple arrow and light purple arrow are the pair of central palindromic sequence repeat. (C) The potential hairpin structure formed by the single repeat unit.

Figure 6. Alignments of 29 tandem repeat units in the *Nasonia vitripennis* (*Nv*) mitochondrial genome indicate high sequence similarity.
(A) The alignment of all 29 repeat units from the *Nv* AsymCx strain mitochondrial control region. The red highlight indicates the substitution of A, blue indicates C, green indicates T, and yellow indicates G. For alignment purposes, the direction of repeat units 18-29 is reversed. The compensatory nucleotide variations in repeats 1-2-3-4 and 26-27-28-29 were labeled in gray boxes. (B) The consensus repeat unit sequences in five *Nv* strains (AsymCx, LabII, R16A, R5-11, and V12.1).

Figure 1

Click here to access/download;Figure;Figure1_v5.pdf

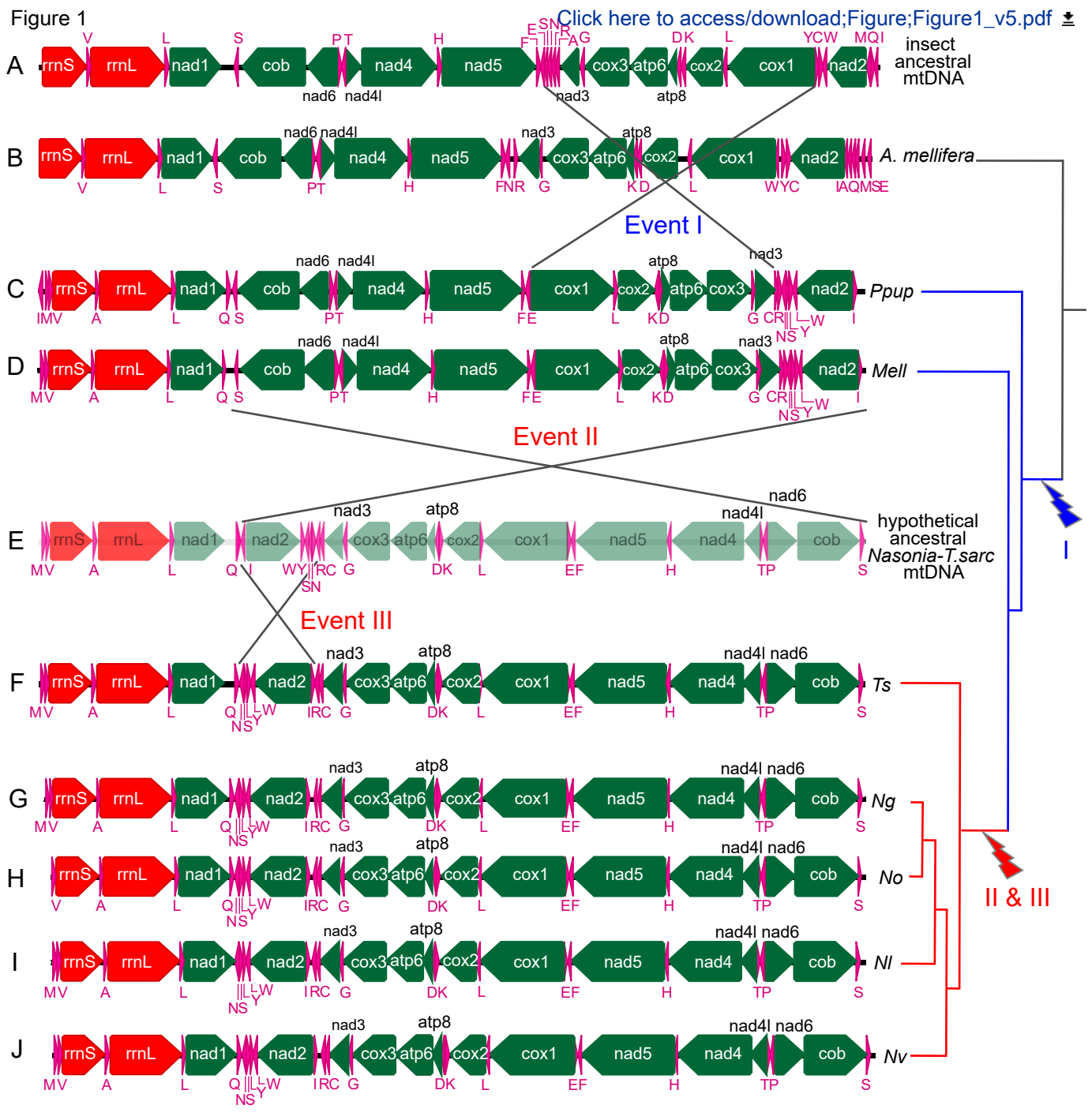
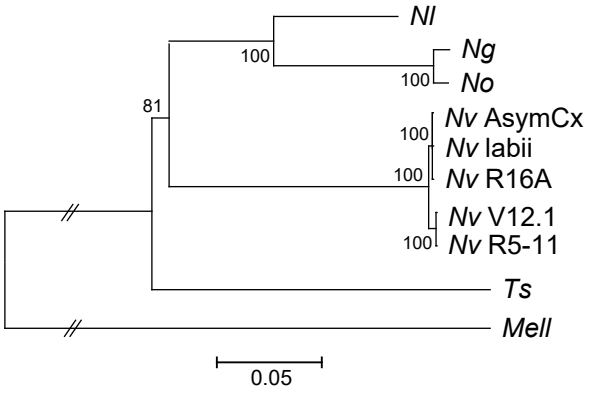


Figure 2



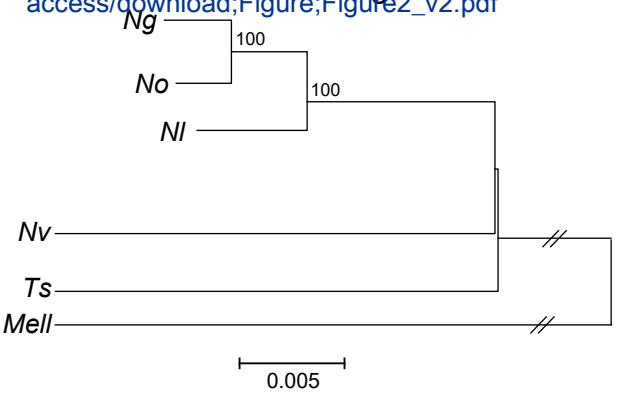


Figure 3

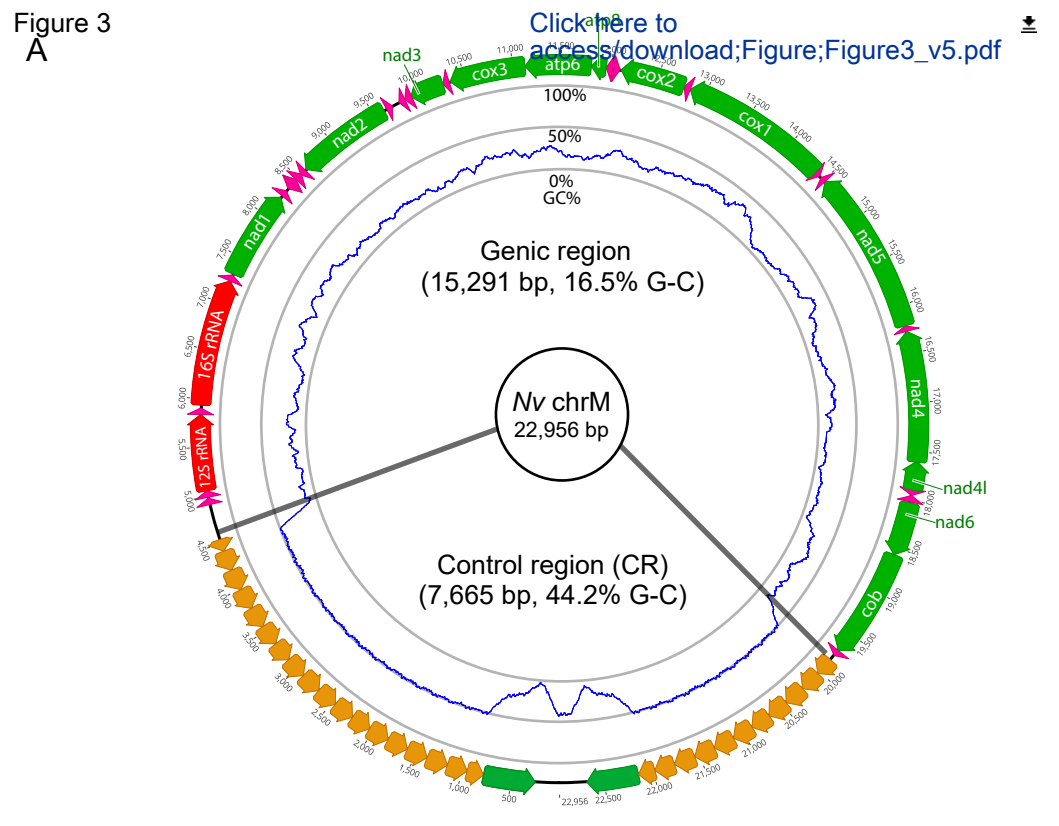


Figure 4

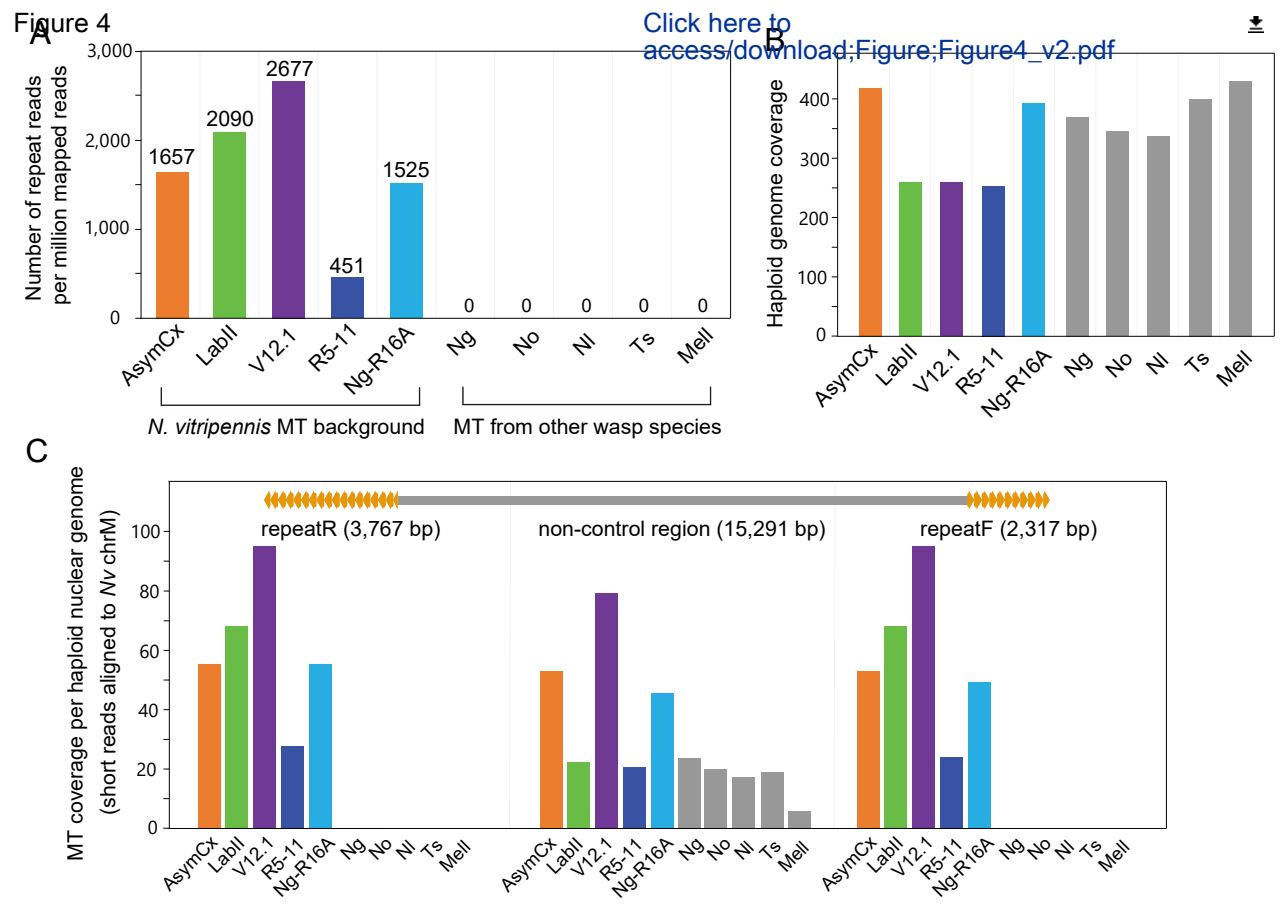


Figure 5

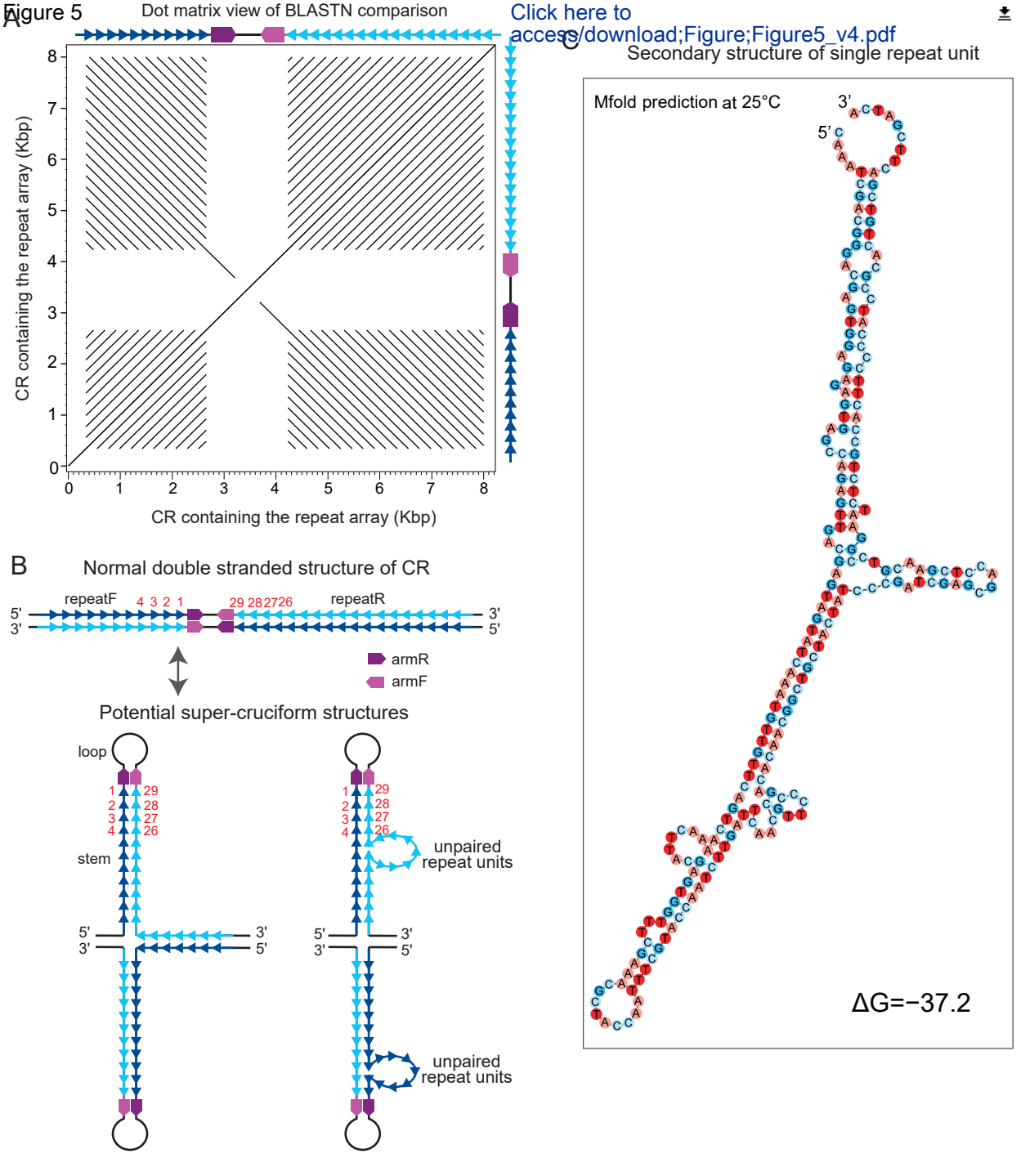


Figure 6

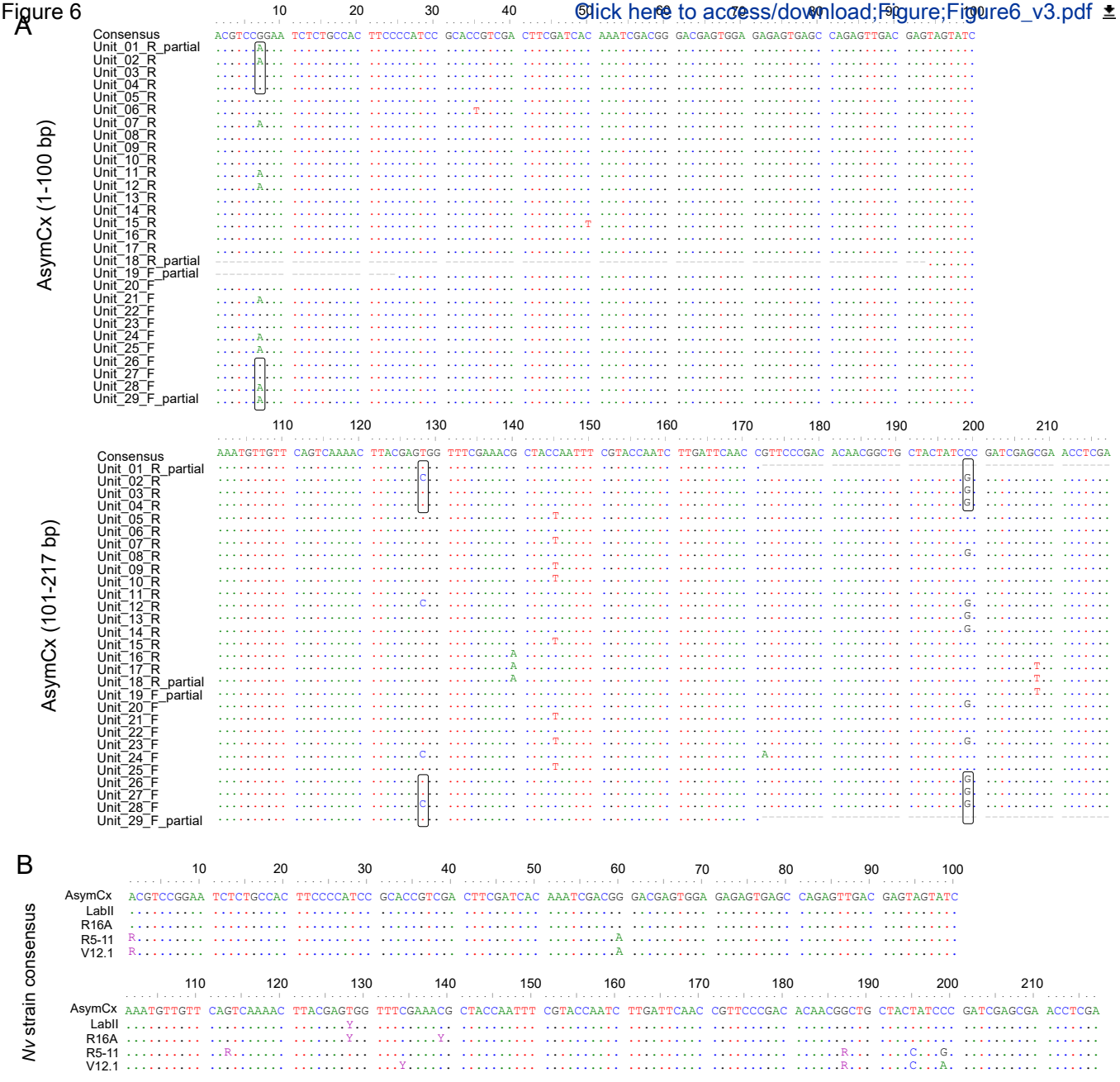


Table 1. Mitochondrial genome assemblies of five wasp species.

mtDNA	Length (bp)	Depth (X)	GC%	CpG%	Completion status
Nv (AsymCx)	22,956	21433.1	25.8%	5.9%	Circularized
Ng	15,310	54221.9	18.0%	1.7%	CR not closed
No	14,811	33091.9	16.5%	1.1%	CR not closed, missing trnM
Nl	14,925	24097.0	15.8%	1.1%	CR not closed, missing trnQ
Ts	15,042	32079.1	17.1%	1.1%	CR not closed

Table 2. Mitochondrial gene annotations of five jewel wasp species.

Type	Gene	*Nv* start	*Nv* end	*Ng* start	*Ng* end	*No* start	*No* end	*Nl* start	*Nl* end	*Ts* start	*Ts* end
rRNA	12 S rRNA	5,053	5,836	395	462	94	871	285	1,063	203	981
	16 S rRNA	5,927	7,234	466	535	953	2,260	1,149	2,506	1,081	2,388
Protein coding genes	nad1	7,308	8,234	536	1,312	2,338	3,264	2,534	3460	2,461	3,387
	nad2	8,612	9,616	1,326	1,389	3,625	4,627	3,734	4,736	3,951	4,950
	nad3	9,923	10,273	1,390	2,700	4,920	5,269	5,010	5,360	5,190	5,539
	cox3	10,344	11,129	2,701	2,771	5,344	6,129	5,430	6,215	5,608	6,393
	atp6	11,132	11,806	2,778	3,704	6,130	6,802	6,216	6,888	6,394	7,066
	atp8	11,800	11,958	3,751	3,822	6,796	6,954	6,882	7,046	7,060	7,218
	cox2	12,105	12,773	3,855	3,921	7,103	7,771	7,192	7,860	7,366	8,034
	cox1	12,837	14,378	3,924	3,985	7,838	9,374	7,927	9,463	8,102	9,638
	nad5	14,520	16,205	3,987	4,054	9,506	11,194	9,594	11,282	9,762	11,448
	nad4	16,271	17,609	4,053	4,118	11,258	12,596	11,347	12,685	11,517	12,855
	nad4l	17,603	17,890	4,119	5,121	12,590	12,877	12,679	12,966	12,849	13,136
	nad6	18,028	18,576	5,149	5,218	13,018	13,566	13,102	13,644	13,276	13,824
	cob	18,578	19,715	5,279	5,348	13,567	14,704	13,645	14,782	13,824	14,961
tRNA	trnM	4,910	4,981	5,354	5,421	-	-	147	215	61	130
	trnV	4,985	5,052	5,422	5,771	24	93	216	284	136	202
	trnA	5,840	5,903	5,772	5,837	885	948	1,072	1,137	987	1,050
	trnL1(UAG)	7,235	7,301	5,846	6,631	2,261	2,331	2,458	2,527	2,389	2,457
	trnQ	8,249	8,320	6,632	7,304	3,267	3,338	-	-	3,592	3,659
	trnN	8,345	8,411	7,298	7,456	3,361	3,427	3,473	3,539	3,682	3,748
	trnS1	8,414	8,473	7,458	7,524	3,430	3,491	3,540	3,602	3,750	3,813
	trnY	8,476	8,543	7,529	7,601	3,493	3,560	3,604	3,669	3,819	3,886
	trnW	8,542	8,608	7,605	8,273	3,559	3,624	3,668	3,733	3,885	3,950
	trnI	9,644	9,704	8,274	8,339	4,655	4,724	4,764	4,832	4,976	5,043
	trnR	9,784	9,854	8,340	9,876	4,777	4,846	4,872	4,942	5,050	5,119
	trnC	9,857	9,923	9,871	9,935	4,852	4,919	4,944	5,010	5,124	5,189
	trnG	10,274	10,341	9,942	10,006	5,270	5,335	5,361	5,427	5,540	5,605
	trnD	11,960	12,024	10,008	11,696	6,956	7,022	7,048	7,113	7,220	7,286
	trnK	12,028	12,101	11,697	11,759	7,027	7,099	7,117	7,189	7,291	7,363
	trnL2(UAA)	12,774	12,841	11,760	13,098	7,772	7,837	7,861	7,926	8,035	8,101
	trnE	14,373	14,442	13,092	13,379	9,369	9,433	9,458	9,523	9,633	9,699
	trnF	14,456	14,520	13,378	13,441	9,440	9,505	9,530	9,593	9,698	9,761
	trnH	16,206	16,270	13,442	13,507	11,195	11,257	11,283	11,346	11,449	11,516
	trnT	17,890	17,954	13,521	14,069	12,877	12,940	12,966	13,030	13,136	13,199
	trnP	17,890	17,954	14,070	15,207	12,941	13,005	13,031	13,096	13,201	13,267
	trnS2	19,716	19,782	15,208	15,275	14,705	14,772	14,783	14,850	14,962	15,029

-: the gene was not found in the current mitochondrial genome assembly.

Author statements

Zi Jie Lin: Methodology, Software, Validation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization. **Xiaozhu Wang**: Software, Investigation, Writing - Original Draft. **Jinbin Wang**: Investigation. **Yongjun Tan**: Methodology, Software, Formal analysis, Data Curation, Writing - Review & Editing. **Xueming Tang**: Resources, Supervision. **John H. Werren**: Conceptualization, Resources, Writing- Reviewing and Editing. **Dapeng Zhang**: Methodology, Software, Formal analysis, Data Curation, Writing - Review & Editing, Visualization, Supervision. **Xu Wang**: Conceptualization, Methodology, Software, Formal analysis, Resources, Data Curation, Writing - Original Draft, Writing- Reviewing and Editing, Visualization, Supervision, Funding acquisition.

Supplementary Figures

Supplementary Tables