Bacterial and fungal symbionts in parasitoid wasp genera *Nasonia* and *Muscidifurax*

by

Xiao Xiong

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
December 1, 2023

Keywords: parasitoid wasps *Nasonia*/*Muscidifurax*, *Wolbachia* endosymbionts, fungal pathogen *Nosema*, genome evolution, comparative genomics

Approved by

Xu Wang, Chair, Associate Professor of Pathobiology, College of Veterinary Medicine
Bruce F. Smith, Professor of Pathobiology, College of Veterinary Medicine
Laurie S. Stevison, Associate Professor in the Department of Biological Sciences
Brian A. Counterman, Associate Professor in the Department of Biological Sciences

Abstract

Jewel wasps in the *Nasonia* genus are ideal models for insect genetics, genomics, epigenetics, development, and evolution. Wasps in the *Muscidifurax* genus have a close evolutionary relationship to the model parasitoid genus *Nasonia*. *Muscidifurax raptorellus* (*M. raptorellus*) has received extensive attention for its potential in biological control against filth flies. I reported the first *de novo Nasonia giraulti* (*Ng*) assembly (259 Mbp) and *M. raptorellus* assembly (314 Mbp) with excellent continuity and completeness using 10× Genomics sequencing and PacBio long-read sequencing technologies.

Nasonia has been an excellent model for intracellular bacteria *Wolbachia* research, with 11 *Wolbachia* strains identified in four *Nasonia* species. Phylogenomic analyses with 210 identified core genes indicated that all 33 *Wolbachia* strains maintained the supergroup relationship determined by multilocus sequence typing (MLST) genes. Fourteen inter-supergroup recombination events (9 A-B events and 5 A-E events) were discovered using an interclade recombination screening method in six genes (2.9%) among 210 single-copy orthologs, which suggested a relatively low frequency of intergroup recombination in *Wolbachia*.

Nosema muscidifuracis (*N. muscidifuracis*) is a microsporidian parasite in the parasitoid wasps *M. zaraptor* and *M. raptor*. I reported a chromosome-level *N. muscidifuracis* genome (14.4 Mbp) with a novel composite 4-bp (TAGG)n and 5-bp (TTAGG)n telomeric repeat motif discovered at the ends of chromosomes. The genome exhibits extensive gene duplications and rearrangements, with high similarity in duplicated genes. *Nosema* titers remained high in *Nosema*-cured parasitoid wasps, suggesting incomplete infection elimination. Comparative phylogenomic

analyses revealed incongruency in *Nosema* and host species trees, indicating a host switch event between parasitoid wasps and bees. A conserved *cis*-regulatory motif ACCC was identified upstream of the start codon. Cytogenetic analyses revealed a substantial *Nosema* load in wasp ovaries, suggesting heritable infection and vertical transmission. The tractability of the parasitoid-*Nosema* system makes it a potential model for Nosemosis studies.

My dissertation work provides novel insights into the genetic architecture, gene regulation, and genome evolution of parasitoid wasps, *Wolbachia* endosymbionts, and fungal pathogen *Nosema*, which will build the foundation for the study of comparative genomics and host-parasite interactions in the parasitoid wasp *Nasonia*/*Muscidifurax* model system, as well as future biocontrol applications.

Acknowledgments

I would like to take this opportunity to express my deepest appreciation to all those who have supported and contributed to the completion of my doctoral dissertation.

First and foremost, I would like to thank my supervisor Dr. Xu Wang, for his invaluable guidance, unwavering support, and exceptional mentorship throughout this challenging academic journey. His continuous encouragement, patience, and insightful feedback have been instrumental in shaping the outcome of my work. I am truly grateful to my committee members, Dr. Smith, Dr. Stevison, and Dr. Counterman, for their diverse perspectives and constructive suggestions that have greatly refined my ideas and strengthened the scientific rigor of my research. I am deeply grateful to my university reader Dr. Zheng, for her suggestions and feedback, which have improved the quality of my dissertation.

I would also like to extend my gratitude to my family, friends, and lab members for their unconditional love, understanding, and support during this demanding period of my life. A special thanks goes to my boyfriend for his belief in my abilities, and words of encouragement, which have been the driving source behind my success. I am truly fortunate to have him by my side.

Thank you all for being a part of my academic journey and for your unwavering support. I do believe that the knowledge and experience acquired during the pursuit of my Ph.D. degree will have a lifelong positive impact on both my academic and professional endeavors.

Table of Contents

Abstract 2

Acknowledgments 4

Table of Contents 5

List of Tables 11

List of Figures 13

Chapter 1 Introduction 17

1.1 The parasitoid wasp *Nasonia*, an emerging model for genetics and evolutionary biology studies 17

1.2 The parasitoid wasp *Muscidifurax*, a biological control agent for filth flies 22

1.3 The associations between parasitoid wasps *Nasonia*/*Muscidifurax* and symbionts 24

1.4 *Wolbachia* genome evolution and host switch 26

1.5 *Nosema* genomes and infections 28

Chapter 2 Genome assembly and evolution of the parasitoid wasps *Nasonia giraulti* and *Muscidifurax raptorellus* 33

2.1 Abstract 33

2.2 Introduction 34

2.3 Materials and methods 35

2.3.1 Sample source and insect rearing 35

2.3.2 Genomic DNA extraction, library preparation, and sequencing 36

2.3.3 Genome assembly, polishing, and assessment 39

2.3.4 RNA-seq data processing and transcriptome assembly 41

2.3.5 Repeat annotation 42

2.3.6 Genome annotation 42

2.3.7 Comparative genomic analysis 44

2.3.8 Phylogenetic analysis 45

2.4 Results 47

2.4.1 Genome assembly and assessment 47

2.4.2 Repeat annotation 51

2.4.3 Gene annotations 54

2.4.4 Genome comparisons 56

2.4.5 Phylogenetic relationships 62

2.5 Discussion 65

2.5.1 High-quality genome assembly of parasitoid wasp *Nasonia giraulti* 65

2.5.2 Potential functions of *Ng*-specific genes 65

2.5.3 The significance of parasitoid wasp *Nasonia* and *Muscidifurax* in genetic research .. 70

2.5.4 The application of parasitoid wasp *Nasonia* and *Muscidifurax* in pest biological control 71

Chapter 3 Phylogenomic analysis of *Wolbachia* strains reveals patterns of genome evolution and recombination 73

3.1 Abstract 73

3.2 Introduction 73

3.3 Materials and methods 76

3.3.1 Phylogenomic analysis of annotated *Wolbachia* genomes 76

3.3.2 Identification of individual gene trees with intergroup recombination events 80

3.3.3 Phylogenetic analysis of *Wolbachia* in *Nasonia* using MLST genes. 81

3.4 Results 82

3.4.1 Phylogenomic analysis of annotated *Wolbachia* genomes 82

3.4.2 Identification of inter-supergroup recombination events 87

3.4.3 Concordance of MLST genes and whole genome divergence 99

3.5 Discussion 102

Chapter 4 Genome evolution and transmission mechanism of the microsporidian pathogen *Nosema muscidifuracis* 105

4.1 Abstract 105

4.2 Introduction 106

4.3 Materials and methods 109

4.3.1 Sample source and insect rearing 109

4.3.2 High molecular weight DNA extraction, PacBio CCS library preparation and sequencing 110

4.3.3 10× Genomics linked-read library construction and Illumina sequencing. 111

4.3.4 Assembly of the *N. muscidifuracis* Genome 112

4.3.5 Assessment of *Nosema* genomes 112

4.3.6 Genome size estimation 113

4.3.7 Telomeric repeat identification 113

4.3.8 Repeat annotation 115

4.3.9 Noncoding RNA annotation 115

4.3.10 *Nosema* inspection and treatment procedures 115

4.3.11 Genomic DNA extraction and *Nosema* titer determination using quantitative PCR116

4.3.12 RNA sample quality control, RNA-seq library preparation and sequencing 119

4.3.13 RNA-seq data processing and gene annotation 119

4.3.14 Comparative genome analysis 121

4.3.15 Confirmation of gene copy number differences using qPCR 121

4.3.16 Functional and pathway annotation of *N. muscidifuracis* proteins..... 122

4.3.17 Motifs prediction in regulatory regions 123

4.3.18 Phylogenetic analysis between *N. muscidifuracis* and other microsporidia..... 123

4.3.19 Cytogenetic analysis of *Nosema* distribution in the ovaries of three *Muscidifurax* species..... 124

4.4 Results..... 125

4.4.1 *Nosema muscidifuracis* genome assembly and statistics..... 125

4.4.2 Assessment of the continuity and completeness of the *N. muscidifuracis* genome... 130

4.4.3 A novel composite 4-bp and 5-bp telomeric repeat motif in *N. muscidifuracis*..... 130

4.4.4 Repeat annotation..... 132

4.4.5 Noncoding RNA annotation identified 57 rDNA clusters located in the middle of *N. muscidifuracis* chromosomes..... 133

4.4.6 Reoccurrence of *Nosema* infection in *M. zaraptor* after cured by a combination of heat treatment and Pasteur method..... 134

4.4.7 Comparative genomic analysis of *N. muscidifuracis* with *N. ceranae* and an outgroup microsporidian *E. cuniculi*..... 137

4.4.8 Genome annotation reveals extensive gene duplication events in *Nosema muscidifuracis* 139

4.4.9 Severe genome reduction and lack of mitochondrial genes in *N. muscidifuracis* 143

4.4.10 A conserved regulatory motif upstream of the translation start sites in *N. muscidifuracis* 147

4.4.11 Phylogenomic analysis with other *Nosema* genomes revealed a host switch event between wasps and bees 150

4.4.12 Cytogenetics of reproductive tissues 152

4.5 Discussion 153

4.5.1 A high-quality genome assembly of *Nosema muscidifuracis* 153

4.5.2 Variation in genome size among *Nosema* species 154

4.5.3 A novel composite form telomere in *Nosema muscidifuracis* 155

4.5.4 Extensive gene duplication events in *Nosema muscidifuracis* 155

4.5.5 Lack of mitochondria 156

4.5.6 Transmission of *Nosema* in parasitoid wasps 157

4.5.7 Toward an ultimate cure for Nosemosis 158

Chapter 5 Conclusions and future directions 160

References 164

List of Tables

Table 1. The summary of *Wolbachia* strains identified in the four *Nasonia* species 28

Table 2. The genome accession numbers and hosts of *Nosema* species..... 30

Table 3. Statistics of the *N. giraulti* genome assembly compared to other wasp species..... 48

Table 4. Summary statistics of the *Muscidifurax raptorellus* genome assemblies..... 50

Table 5. Summary of repetitive element content found in the *N. giraulti* genome assembly 52

Table 6. Summary repeat element classes in *Muscidifurax raptorellus* and *Nasonia vitripennis* genomes 53

Table 7. Alignment length and percentage of *N. giraulti* scaffolds to *N. vitripennis* genome 57

Table 8. The top 10 *N. giraulti* scaffolds covering each of *N. vitripennis* chromosomes 58

Table 9. Summary of *Ng*-specific genes compared to *Nv*..... 67

Table 10. Summary of current sequenced *Wolbachia* genomes..... 74

Table 11. List of six *Wolbachia* genes with interclade recombination events..... 87

Table 12. Interclade recombination events detected in 33 *Wolbachia* genomes between A and B supergroups..... 91

Table 13. Correlation of evolutionary divergence estimates between *Wolbachia* species using 210 core gene set and five MLST genes..... 100

Table 14. Genome assembly statistics for *Nosema muscidifuracis* and comparison with other microsporidian genomes 113

Table 15. Short-read genome sequencing data in *Nosema* species for telomeric repeat motif identification 114

Table 16. The 18S primers used for the quantification of *Nosema* titer in the *Muscidifurax zaraptor* genome 117

Table 17. RNA sequencing sample information, data yield, quality control summary statistics 120

Table 18. Summary PacBio long-read and Illumina (10× Genomics) linked-read sequencing data generated for *Muscidifurax zaraptor* and *Nosema muscidifuracis* genome assembly 126

Table 19. Average PacBio coverage depth against the *Nosema muscidifuracis* genome..... 126

Table 20. Summary of annotated repeat elements in *Nosema muscidifuracis* genome..... 132

Table 21. The annotation of noncoding RNAs in the *Nosema muscidifuracis* genome 133

Table 22. The number of genes in major functional pathways identified in *Nosema muscidifuracis* and *Saccharomyces cerevisiae* 144

Table 23. Motif sequences and significance identified with 449 shared genes in *Nosema* species 149

Table 24. Motif sequences and significance identified with other predicted genes in *Nosema* species 149

List of Figures

Figure 1. The life cycle of *Nasonia* wasps. 18

Figure 2. Phylogenetic relationship between four closely related *Nasonia* species and the outgroup species *Trichomalopsis sarcophagae* (*T. sarcophagae*, *T. sarc*). 19

Figure 3. Image of *N. giraulti* and its geographic distribution in the North America based on Darling & Werren (1990) [6]. 20

Figure 4. Cytoplasmic incompatibility (CI) induced by different *Wolbachia* infection status. 21

Figure 5. Phylogenetic relationship of *Nasonia* and other sequenced genomes. 22

Figure 6. The interactions between parasitoid wasp *Nasonia* and microbiota. 25

Figure 7. The map of the *w*Mel (A-*Wolbachia* parasite of *Drosophila melanogaster*) chromosome shows the location of the five MLST genes (*coxA*, *gatB*, *hcpA*, *ftsZ* and *fbpA*) and *wsp* gene [59]. 28

Figure 8. Scaffold median coverage and GC content in the *N. giraulti* genome assembly. 48

Figure 9. Comparison of the gene length distributions for 3,662 shared 1:1 single copy orthologs in *M. raptorellus* and nine representative Hymenopteran species. 56

Figure 10. Chromosome level alignment between *N. giraulti* scaffolds and *N. vitripennis* chromosomes. 57

Figure 11. Genome comparisons between *Muscidifurax raptorellus* and *Nasonia vitripennis*. 61

Figure 12. Phylogenetic relationships of *N. giraulti* with eight selected arthropod species. 63

Figure 13. Phylogenetic relationship between *M. raptorellus* and nine representative hymenopteran species. 64

Figure 14. Workflow of bioinformatic identification *Ng*-specific genes compared to *Nv.* 67

Figure 15. Functional clustering analysis of *Ng*-specific genes using Blast2GO. 70

Figure 16. Nucleotide ML trees for six *Wolbachia* genes with interclade recombination events. 80

Figure 17. Phylogenetic analysis of 34 *Wolbachia* genomes. 84

Figure 18. Phylogenomic relationships of 33 *Wolbachia* strains. 85

Figure 19. Phylogenetic analysis of 33 *Wolbachia* genomes (*w*Con excluded) from concatenated protein sequence alignment of 210 single-copy orthologous genes. 86

Figure 20. Inter-supergroup recombination events of B supergroup genes *fstH* and *rplU* in A-*Wolbachia* strains. 91

Figure 21. Inter-supergroup recombination events of A supergroup genes *coxB* and *WONE_04820* in B-*Wolbachia* and E-*Wolbachia* strains. 94

Figure 22. Intragenic recombination event between supergroups in *argS* gene. 96

Figure 23. The nucleotide ML tree reveals recombination event where A-*Wolbachia* cluster with E-*Wolbachia* in *dnaK* gene. 98

Figure 24. Correlation of evolutionary divergence estimated by core gene set and the five concatenated MLST genes. 101

Figure 25. PCR results for three primer sets targeting the 18S rDNA gene in *Nosema muscidifuracis*. 118

Figure 26. The plot of GC content and CpG percentage versus average coverage for all scaffolds from the initial assembly suggested that the scaffolds of *N. muscidifuracis* are separated from other scaffolds in *M. zaraptor* genome. 128

Figure 27. Chromosome level genome assembly of *Nosema muscidifuracis* showing GC content and rDNA clusters..... 129

Figure 28. A novel type of telomere in the *Nosema muscidifuracis* genome. 131

Figure 29. Quantification of the *Nosema* titer in AUB and USDA *Muscidifurax zaraptor* colony. 136

Figure 30. Genome comparisons between *Nosema muscidifuracis* and *N. ceranae*. 138

Figure 31. Genome comparisons between *Nosema muscidifuracis* and *Encephalitozoon cuniculi*. 139

Figure 32. Histograms of gene copy number in *Nosema muscidifuracis*, *N. ceranae*, and *E. cuniculi* genomes and self-circos plot of 28 contigs in *N. muscidifuracis*..... 142

Figure 33. Quantification and confirmation of gene copy number in *Nosema muscidifuracis* using quantitative PCR approach. 143

Figure 34. Functional pathway specific genome reduction in *Nosema muscidifuracis*. 146

Figure 35. A motif associated with translation start sites and gene expression levels in *Nosema muscidifuracis*. 148

Figure 36. Phylogenomic analysis revealed a host switch event and conserved sequence motif in *Nosema*..... 151

Figure 37. Cytogenetic analysis of *Nosema muscidifuracis* distribution in developing and mature oocytes of three *Muscidifurax* species. 153

Chapter 1 Introduction

1.1 The parasitoid wasp *Nasonia*, an emerging model for genetics and evolutionary biology studies

As a representative genus of parasitoid wasps, *Nasonia* (Hymenoptera: Pteromalidae) is easy to rear on commercially available fly pupae in the laboratory with a short generation time (about two weeks at the temperature of 25°C and a constant 24-hour light), large family size, and cross-fertile species. The life cycle comprises stages of egg, larva, pupa, and adult (Figure 1) [1]. Pupae and adults can be stored in 4°C refrigeration for several weeks, and larvae exhibit an inducible overwintering diapause, allowing for the maintenance of strains in refrigeration for approximately 18 months [2]. *Nasonia* wasps live a parasitoid lifestyle, where females inject venom into fly pupal hosts and then deposit eggs onto the fly puparium. The venoms induce developmental arrest and changes in host gene expression and metabolism [3-5], with the feeding wasp larvae eventually killing the host. Development during the first three stages occurs within the host, and mature adults emerge from the host by chewing small holes in the host puparium (Figure 1) [1]. Adult wasps can mate within the host pupa or after emergence.

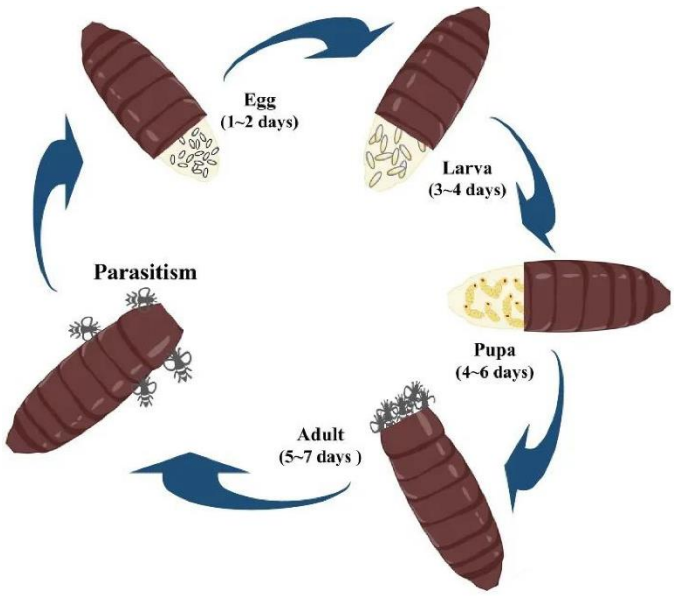


Figure 1. The life cycle of *Nasonia* wasps.

This figure shows the developmental stages of *Nasonia* and the corresponding duration in days for each stage [1].

There are four closely related species in the genus of *Nasonia*, including *N. vitripennis* (*Nv*), *N. giraulti* (*Ng*), *N. oneida* (*No*), and *N. longicornis* (*Nl*) (Figure 2) [6-9]. *Nv* was the first and only species described in the *Nasonia* genus for a long period of time and has a worldwide distribution [8]. *Ng*, *Nl*, and *No* are closely related to *Nv* and have a more restricted North American distribution (Figure 3), where they parasitize blowfly pupae in birds' nests [6, 7]. All *Nasonia* species typically have a haploid karyotype of n = 5, which corresponds to 5 linkage groups. The form of sex determination in *Nasonia* is haplodiploidy (Females are diploid and develop from fertilized eggs, whereas males are haploid and originate from unfertilized eggs), which is shared among the Hymenoptera [8, 10, 11]. The four species in the *Nasonia* genus are interfertile, while reproductive incompatibility due to *Wolbachia*-induced cytoplasmic incompatibility occurs in *Nasonia*, except

for *Ng*/*No* (Figure 4) [12, 13]. However, interspecies interspecific hybrids of *Nasonia* are readily generated after antibiotic curing the wasp of endosymbiont *Wolbachia*.

Figure 2. Phylogenetic relationship between four closely related *Nasonia* species and the outgroup species *Trichomalopsis sarcophagae* (*T. sarcophagae*, *T. sarc*).

The grey box shows the three sequenced *Nasonia* genomes, *N. vitripennis* (*Nv*), *N. girault* (*Ng*), and *N. longicornis* (*Nl*) [9]. All species in the *Nasonia* genus are interfertile after the removal of the intracellular bacterium *Wolbachia*.

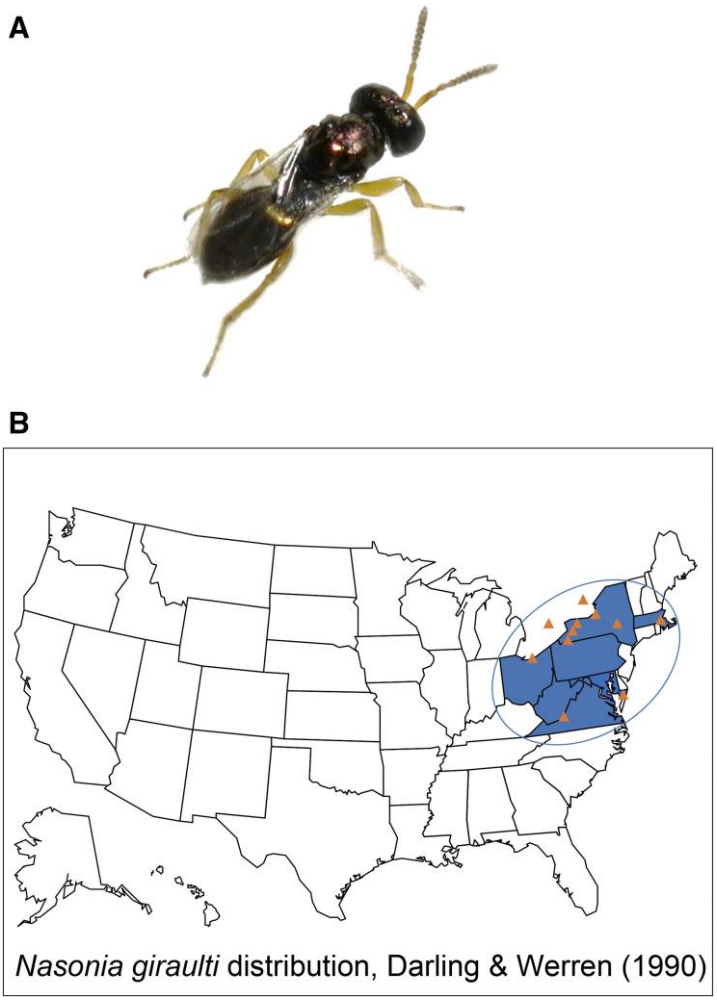


Figure 3. Image of *N. giraulti* and its geographic distribution in North America based on Darling & Werren (1990) [6].

(A) The image of *N. giraulti* wasp (♂) under microscopy. (B) The map of the United States shows the geographic distribution of *N. giraulti* in North America, and orange triangles represent the distribution of *N. giraulti* wasps.

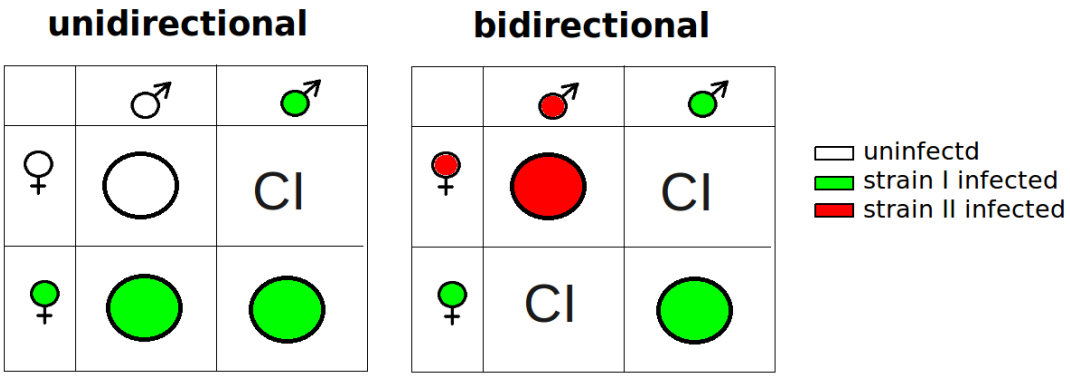


Figure 4. Cytoplasmic incompatibility (CI) induced by different *Wolbachia* infection status.

Unidirectional CI typically occurs when infected males mated with uninfected females, while all other crosses are compatible. Bidirectional CI is expressed in males and females infected with two different *Wolbachia* strains. Circles represent viable offspring, and the infection status is color-coded. CI denotes cases where no viable offspring are produced (Picture credit goes to Hu. Johannes, 2010).

Unlike *Drosophila melanogaster*, which lacks CpG DNA methylation, all three DNA cytosine-5-methyltransferases (Dnmts) are present in *Nasonia* (Figure 5) [9, 14]. Approximately a third of the genes are methylated in the *Nv* and *Ng* genomes, and the methylation is primarily on the gene body [15, 16]. Furthermore, DNA methylation appears to be regulated in *cis* in F1 hybrids, and stable inheritance of methylation status of individual genes has been observed over successive generations in hybrids [17]. The biological features of *Nasonia* wasps include ease of rearing, rapid development, a short generation time, a large number of offspring, interfertile species, visible and molecular markers, and haplodiploid genetics with a full DNA methylation toolkit. These characteristics make *Nasonia* an excellent model organism for studies in genetics, epigenetics,

development, evolution, and behavior [8, 10, 11, 18-20]. Moreover, the microbial diversity in parasitoid wasps develops a model system to investigate host-microbiome interactions [21, 22].

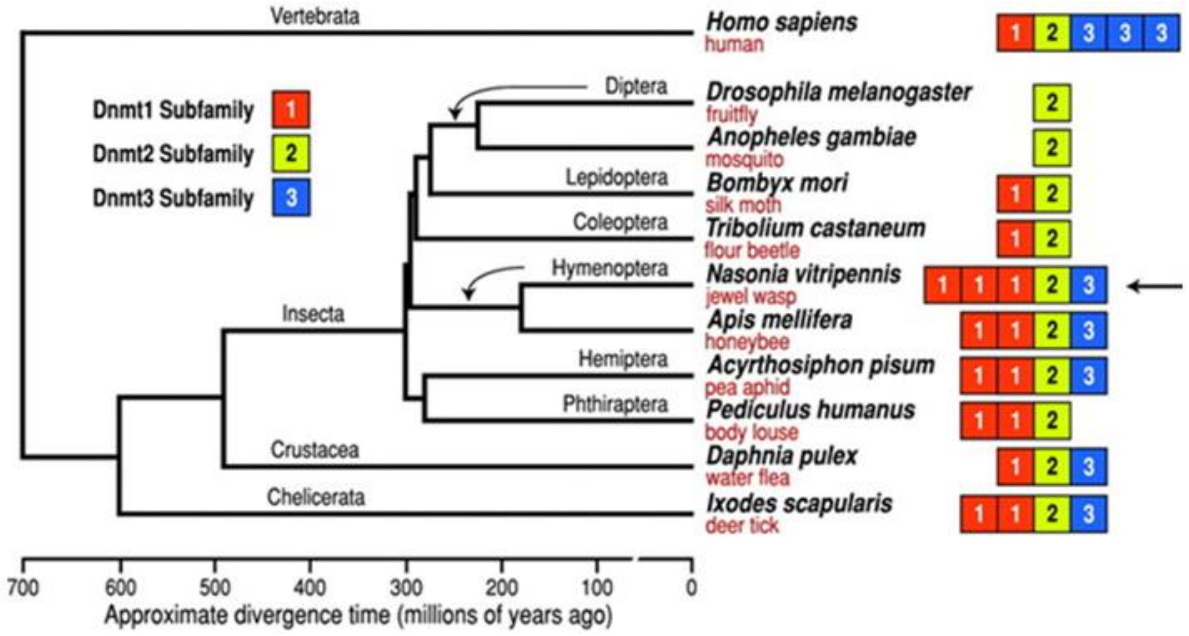


Figure 5. Phylogenetic relationship of *Nasonia* and other sequenced genomes.

On the right, the figure shows the DNA methylation tool kit and the presence of DNA methyltransferase subfamilies (Dnmt1, Dnmt2, Dnmt3) in these taxa [9].

1.2 The parasitoid wasp *Muscidifurax*, a biological control agent for filth flies

Muscidifurax (Hymenoptera: Pteromalidae) is a chalcid wasp genus with nine characterized species, all of which are pupal parasitoids. *Muscidifurax raptor* was the first species described in the genus in 1910 by Girault and Sanders [23]. In 1970, four sibling species were described: *M. zaraptor* Kogan and Legner, collected from the southwestern United States; *M. raptoroides* Kogan and Legner collected from Central America and Mexico; *M. raptorellus* Kogan

and Legner collected from Uruguay and Chile; and a thelytokous species *M. uniraptor* Kogan and Legner collected in the central mountain range of the island of Puerto Rico [24]. Based on the mitochondrial gene sequence alignment in this genus, the most closely related sexual species to the asexual *M. uniraptor* is *M. raptorellus* [25]. Four additional *Muscidifurax* species were identified in China [26].

Muscidifurax raptorellus (Chilean strain) is a gregarious parasitoid typically produces 2-10 offspring per parasitized host pupa [27]. The number of eclosed offspring depends on the host size [28]. A population found in Uruguay is partially gregarious [29]. Females can lay 16-20 eggs per day during their peak ovipositional period and about 150 eggs during their lifetime [30]. In sharp contrast, *M. zaraptor* only deposits one egg per host, and the first larva will eliminate subsequent larvae or eggs deposited by superparasitism [31]. *M. uniraptor* only produces a single female offspring from each host, and the parthenogenesis is caused by the infection of an A-strain *Wolbachia* bacteria [32, 33]. The diverse reproductive strategies make this genus an excellent model system for the study of sexual vs. asexual evolution.

Traditional pest management methods relied upon the application of chemical insecticides. Nevertheless, this approach encountered constraints associated with its high expenses, environmental pollution, and the emergence of resistance genes within pest populations. Parasitoid wasp *M. raptorellus* is an effective biological control agent against dipteran filth flies that is both environmentally friendly and sustainable [27, 30, 34, 35]. Furthermore, wasps in the genus of *Muscidifurax* hold a special place in the realm of comparative genome studies due to their close relationship with the model parasitoid genus *Nasonia,* with an estimated divergence of approximately 15 million years [36].

1.3 The associations between parasitoid wasps *Nasonia*/*Muscidifurax* and symbionts

The associations between parasitoid wasps *Nasonia*/*Muscidifurax* and symbionts are complex and diverse. The symbionts, including intracellular bacteria *Wolbachia*, fungal pathogen *Nosema*, and gut microbes, are revealed to have significant impacts on various aspects of fitness in parasitoid wasps. Previous research has provided insights into the roles of the symbionts within parasitoid wasps *Nasonia* (Figure 6) [1]. These include the influence of intracellular bacteria *Wolbachia* on cytoplasmic incompatibility (CI) [37], the impact of fungal pathogen *Nosema* on development, longevity, and fecundity [38], the effect of microbiota on pesticide resistance [39], and the role of microbiota in nutrient allocation during diapause [40].

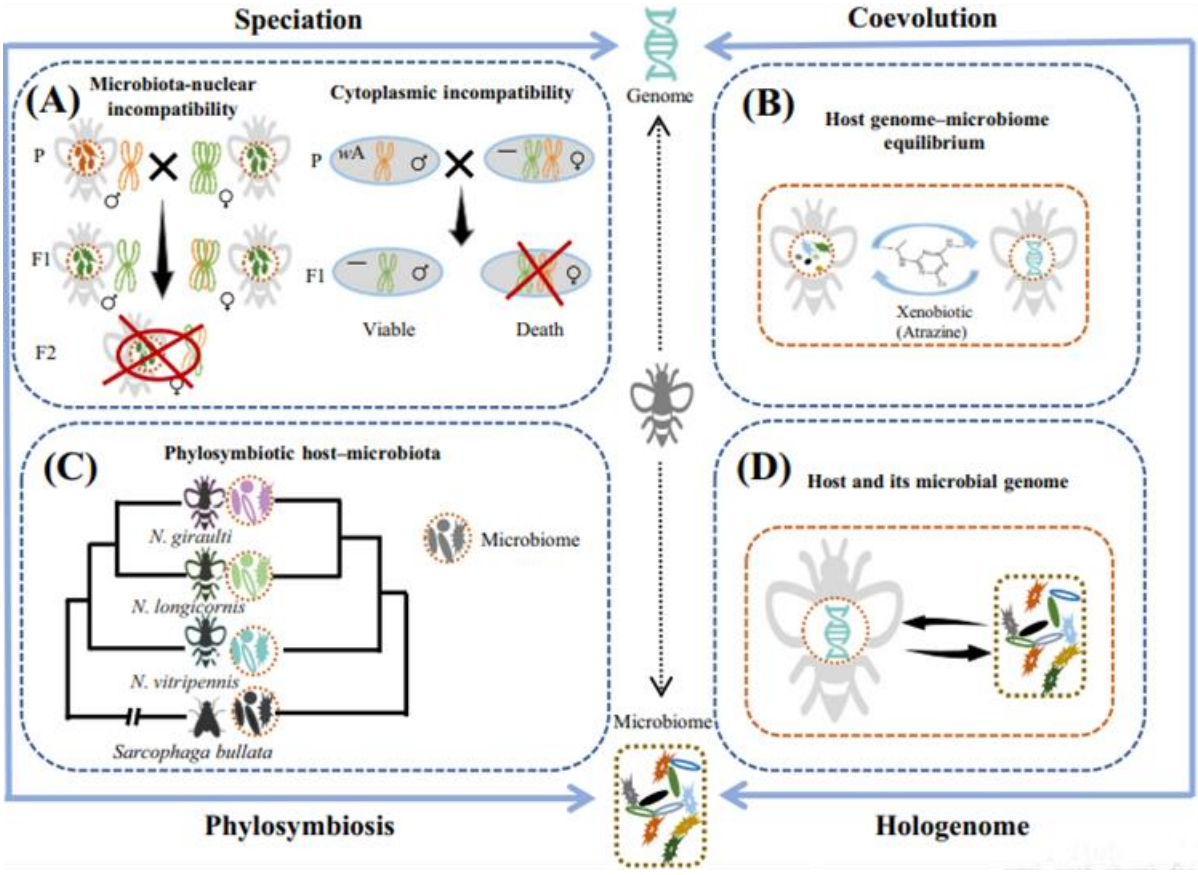


Figure 6. The interactions between parasitoid wasp *Nasonia* and microbiota.

(A) Microbial-nuclear incompatibility and cytoplasmic incompatibility induced by *Wolbachia* contributed to speciation in parasitoid wasps *Nasonia*. (B) Coevolution of the host genome and its microbiota occurred through continuous, multigenerational exposure of *Nasonia* to low-concentration atrazine. (C) The phylogeny of three *Nasonia* species closely resembles that of their microbiota, a phenomenon known as phylosymbiosis. (D) Hologenome represents both the host genome and the genomes of its associated microorganisms [1].

Wolbachia, a widespread intracellular bacterium found in insects, plays a vital role in parasitoid wasps [41]. It can manipulate host reproduction through mechanisms like cytoplasmic incompatibility (CI), parthenogenesis (PI), male-killing (MK), and feminization [37]. In the genus of *Nasonia*, 11 different *Wolbachia* strains from supergroups A and B have been identified in four *Nasonia* species, with the presence of different strains leading to reproductive barriers and facilitating speciation. In the parasitoid wasp genus *Muscidifurax*, an A-*Wolbachia* (*w*Uni) has been proven to be associated with parthenogenesis in *M. uniraptor* [32, 33]. The interactions between parasitoid wasps and *Wolbachia* have profound implications for the evolution of parasitoid wasp species.

Microbiota in *Nasonia* jewel wasps have been linked to pesticide resistance [39]. Exposure to low concentrations of atrazine pesticide led to changes in the microbiota (at least two rare bacteria: *Serratia marcescens* NVIT01 and *Pseudomonas protegens* NVIT02) that conferred resistance to *Nasonia* hosts [39]. The altered microbial communities were found to be heritable and contributed to host adaptation to environmental stress. Isolated bacteria from resistant hosts were also identified as contributors to atrazine resistance, shedding light on the role of microbiota

in pesticide resistance. *Nasonia* wasps undergo diapause, a period of dormancy lasting up to two years. During this period, the microbiota is revealed to play a critical role in nutrient allocation [40]. Microbiota presence or absence influences the levels of glucose, glycerol, triglycerides, and protein in *Nasonia* larvae during diapause [42]. These findings offer insights into the importance of the microbiota in nutritional distribution and metabolism, especially under low-temperature conditions [40].

1.4 *Wolbachia* genome evolution and host switch

The obligate intracellular bacteria *Wolbachia* (alphaproteobacterial endosymbionts) commonly infect arthropods and filarial nematodes [43-45]. In particular, more than half of the arthropod species are infected by *Wolbachia* [46, 47] due to horizontal movement of the bacteria between different host species (host switch), although *Wolbachia* are generally transmitted vertically. The extensive prevalence of *Wolbachia* infections possibly represents a dynamic equilibrium between gain and loss on a global scale [48-50]. The *Wolbachia*-host interaction generally spans a range from reproductive parasitism to mutualism. *Wolbachia* can alter host reproduction to enhance their own transmission in different ways, such as feminization of genetic males, male-killing, parthenogenetic induction, and cytoplasmic incompatibility (Figure 4) [44, 51, 52]. Other effects of *Wolbachia* can include viral suppression [53], and nutritional mutualism [54]. In nematodes, *Wolbachia* appear to have evolved a long-standing mutualistic relationship [43]. *Wolbachia* have been found to move across species boundaries through horizontal transfer and hybrid introgression [55-57].

Wolbachia pipientis have been divided into supergroups (A-H) based on 16S ribosomal RNA sequences and other sequence information, including six supergroups (A,B and E-H)

primarily identified in arthropods and two supergroups (C and D) commonly found in filarial nematodes [44]. However, it has been proposed that supergroup G be decommissioned, as it is based primarily on recombinant *wsp* (*Wolbachia* surface protein) sequences and cluster with A supergroup based on five multi-locus strain typing genes [58, 59], eight supergroups (A- H) are still widely used in the research community. A multi-locus strain typing (MLST) system based on five house-keeping genes (*coxA*, *gatB*, *hcpA*, *ftsZ*, and *fbpA*) has been developed for *Wolbachia* (Figure 7) [59], and is widely used for strain typing and to characterize strain variation within *Wolbachia*. However, the increasing number of genome sequences for *Wolbachia* allows for a more detailed characterization of their diversity, including inter-strain recombination events. The jewel wasp genus of *Nasonia* has been an excellent model for *Wolbachia* research [13, 52, 60-62]. Eleven *Wolbachia* strains have so far been identified in the four species of *Nasonia* (Table 1) [62]. These are often maintained as multiple infections within individual wasps of each species [61] and have diverse evolutionary origins, indicating horizontal transfers from divergent host species.

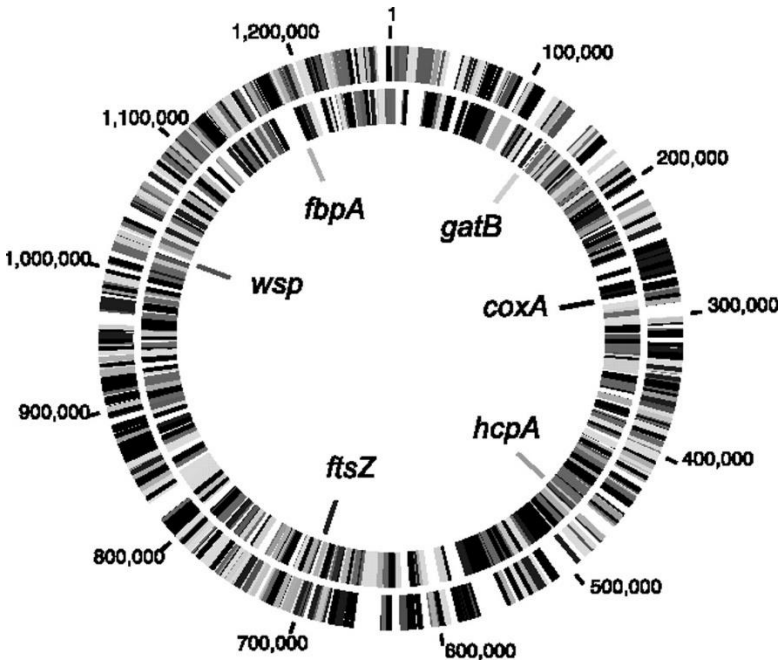


Figure 7. The map of the *w*Mel (A-*Wolbachia* parasite of *Drosophila melanogaster*) chromosome shows the location of the five MLST genes (*coxA, gatB, hcpA, ftsZ* and *fbpA*) and *wsp* gene [59].

Table 1. The summary of *Wolbachia* strains identified in the four *Nasonia* species

***Nasonia* species**	***Nasonia* strain in Wang Lab (w/*Wolbachia*)**	***Wolbachia* strain**
N. vitripennis	-	*w*Nvit A, *w*Nvit B
	NvLabII	*w*Nvit A, *w*Nvit B
	NvV12	*w*Nvit A
N. giraulti	-	*w*Ngir A1, *w*Ngir A2, *w*Ngir B
	NgVA19_008U	*w*Ngir A1, *w*Ngir B
N. oneida	-	*w*None A1, *w*None A2, *w*None B
	NONY11/36	*w*None A1
	NONYAUD108	*w*None A2, *w*None B
N. longicornis	-	*w*Nlon A, *w*Nlon B1, *w*Nlon B2
	NlMTMC13A	*w*Nlon A, *w*Nlon B2

1.5 *Nosema* genomes and infections

Microsporidia are unicellular eukaryotes that comprise a large group of obligate intracellular fungi. They are able to parasitize a wide range of vertebrates and invertebrates, particularly abundant in insects [63, 64]. More than 1,400 microsporidia belonging to over 200

genera have been reported thus far. Compared with other eukaryotes, the genomes of some microsporidia are extremely compact [65, 66]. Microsporidia are highly specialized and have undergone massive reductions in morphology, ultrastructure, metabolism, and gene content, along with other adaptations to their intracellular parasitic life cycle [67, 68]. *Nosema* (Microsporidia: Nosematidae) is one of the most diverse genera of microsporidian parasites, and they are widely distributed in insects and other arthropods [63, 64].

The honey bees are susceptible to the best-known *Nosema* species, *Nosema apis* and *Nosema ceranae*. *N. apis* infects the Western honey bee *Apis mellifera*. The only available *N. apis* genome has an assembly size of 8.6 Mb, and the BUSCO score was only 75.0% [69]. *N. ceranae* is parasitic on the Asian honey bee *Apis cerana*. The latest reported assembly (BRL strain) was in good quality, with a genome size of 8.8 Mb and a BUSCO score of 97.0% [70]. *Nosema* sp. YNPr has the smallest reported genome size of 3.6 Mb and a completeness BUSCO score of 84.5%, which is pathogenic to the cabbage butterfly *Pieris rapae* [71]. *Nosema bombycis* is considered to be a parasite of the silk moth *Bombyx mori*, and the genome size of *N. bombycis* is 15.7 Mb with an 83.0% BUSCO score [72]. The smaller genome size and lower BUSCO score suggested that the genome assembly may be incomplete or loss of some conserved genes in microsporidia. *N. antheraeae* (7.1 Mb) was isolated from the Chinese tussar moth *Antheraea pernyi* [73], and *N. granulosis* (8.9 Mb) infects the Amphipod *Gammarus duebeni* (Table 2) [74]. High-quality genomes are important genetic resources for the studies of comparative genomics and genome evolution of *Nosema* pathogens. Compared to other Microsporidia (*E. cuniculi* GB-M1 strain: the genome size of 2.5 Mb with 100% completeness) [75, 76], the limitation of low GC content makes it hard to achieve complete and high-quality genome assembly.

Table 2. The genome accession numbers and hosts of *Nosema* species

Species (strain)	Host name	Host order	Accession
N. apis (BRL01)	*Apis mellifera*	Hymenoptera	GCA_000447185.1
N. ceranae (BRL)	*Apis cerana*	Hymenoptera	GCA_004919615.1
N. ceranae (PA08)	*Apis cerana*	Hymenoptera	GCA_000988165.1
N. ceranae (BRL01)	*Apis cerana*	Hymenoptera	GCA_000182985.1
Nosema sp. (YNPr)	*Pieris rapae*	Lepidoptera	N/A
N. antheraeae (YY)	*Antheraea pernyi*	Lepidoptera	N/A
N. bombycis (CQ1)	*Bombyx mori*	Lepidoptera	GCA_000383075.1
N. granulosis (Ou3-Ou53)	*Gammarus duebeni*	Amphipoda	GCA_015832245.1
E. cuniculi (GB-M1)	*Rabbit/Human*	Lagomorpha/ Primates	GCA_000091225.2

Nosema infection, a disease known as Nosemosis, occurs throughout the world and leads to agricultural economic losses through detrimental effects on pollinators due to *N. apis* infections found in European honeybee *Apis mellifera* and *N. ceranae* infections originally detected in Asian honeybee *Apis cerana* [69, 77-79]. *Nosema* parasitism has destructive impacts on honey bees, including digestive disorders, poor colony development, and reduction in metabolism and reproduction [80-82]. *N. ceranae* is more prevalent, impairing bee health and declining bee colonies. It damages the physical and immune barriers of honeybees, making them more vulnerable to other pathogenic factors, which are related to colony collapse [83-87]. Despite the low prevalence of *N. apis* compared to *N. ceranae* [88], there is no substantial difference in virulence between the two species and the negative impacts of *N. apis* on the infected colonies

cannot be neglected [89, 90]. *N. apis* was reported to be associated with reduced lifespan and increased winter mortality in infected bees, and its negative effects on colony strength and productivity have been described in several studies [91-94].

Another *Nosema* species, *N. bombycis*, infects the silk moth *Bombyx mori*, causing a lethal disease called Pébrine, which is a serious threat to silk production [72]. The consequences of *Nosema* infection were extensively studied due to its economic importance. *N. bombycis*-infected larvae are inactive and develop slowly, with black spots spreading all over the bodies, eventually leading to death [72]. Currently, there is no effective treatment for the disease.

Nosema also infects other beneficial insects, including the parasitoid wasp species in *Muscidifurax* (Hymenoptera: Pteromalidae) genus, which serve as biological control agents for filth flies. The microsporidium was first found infecting *M. raptor* collected from New York dairy farms in 1990 [95], and was later described as *N. muscidifuracis* [96]. Nosemosis significantly affects the development, longevity, and fecundity of *M. raptor*. *M. raptor* infected by *N. muscidifuracis* takes longer to complete development, lives about half as long, and produces about 10% as many progeny as uninfected individuals [38].

The mechanisms of *Nosema* transmission are directly related to the control of Nosemosis. *N. bombycis* can be transmitted from the mother host to their eggs vertically and gradually invades the whole body of the larvae, including muscles, intestines, silk glands, and Malpighian tubules. *Nosema* transmission in honey bees is primarily through the mouth to uninfected bees. *N. apis* tends to restrict its life cycle to gut epithelial cells after infection, while *N. ceranae* is mainly distributed in the midgut but also spread in other tissues [85, 97]. Due to the limitations of conducting *Nosema* infection experiments in bees, another Hymenoptera-*Nosema* model system

is needed for research in controlled laboratory settings to determine the infection, distribution, transmission, and evolution of *Nosema* and inform the control of nosemosis.

Chapter 2 Genome assembly and evolution of the parasitoid wasps *Nasonia giraulti* and *Muscidifurax raptorellus*

2.1 Abstract

Jewel wasps in the genus of *Nasonia* are excellent models for insect genetics, genomics, epigenetics, development, and evolution. *N. vitripennis* (*Nv*) and *N. giraulti* (*Ng*) are closely related species that can be intercrossed, particularly after removal of the intracellular bacterium *Wolbachia*, serving as a powerful tool to map and positionally clone morphological, behavioral, gene expression and DNA methylation phenotypes. The *Nv* reference genome was assembled using Sanger and PacBio data and annotated with extensive RNA-seq data. In contrast, the *Ng* genome is only available through low-coverage resequencing. Therefore, *de novo* assembly of a high-quality genome is urgently needed to advance uses of this system. Wasps in the genus *Muscidifurax* have a close evolutionary relationship to the model parasitoid genus *Nasonia*. The parasitoid wasp *Muscidifurax raptorellus* (*M. raptorellus*) is a gregarious species that has received extensive attention for its potential in biological pest control against house fly, stable fly, and other filth flies. It has a high reproductive capacity and can be reared easily. However, genome assembly is not available for *M. raptorellus* or any other species in this genus. In this study, I report the first *de novo Ng* assembly using 10× Genomics linked-reads with 600× sequencing depth. The current assembly has a genome size of 259 Mbp in 3,160 scaffolds with a BUSCO completeness score of 98.6% and a 97% perfect mapping rate of RNA-seq reads, indicating high quality in contiguity and completeness. The high-quality genome of *M. raptorellus* was assembled using a combination of long-read (104× genome coverage) and short-read (326× genome coverage) sequencing technologies. The genome size is 314 Mbp in 226 contigs with a 97.9% BUSCO completeness

score and a contig N50 of 4.67 Mb, suggesting excellent continuity of this assembly. My assemblies of *Ng* and *M. raptorellus* build the foundation for comparative and evolutionary genomic analysis in the model of the *Nasonia*/*Muscidifurax* research system and possible future biocontrol applications.

2.2 Introduction

Nasonia has been a good model for genetics, epigenetics, developmental biology, evolutionary biology, and behavioral studies [8, 10, 11]. Whole-genome sequencing efforts have been made in *Nv*, *Ng*, and *Nl* [98]. The *Nv* genome was sequenced with 6× coverage Sanger sequencing to generate a *de novo* assembly, whereas *Ng* and *Nl* genomes were sequenced with 1× coverage supplemented with short-read Illumina sequences, and aligned to the *Nv* assembly for reference-based genomes [98].

Plenty of datasets have been published for the *Nv* genome and transcriptomes after its reference genome was available. Crosses between *Nv* and *Ng* have been extremely successful for mapping and positional cloning of genes involved in species differences [11], in some cases using chromosomal regions of *Ng* introgressed into an *Nv* genetic background [99]. Comparative genomics between *Ng* and *Nv* is informative to investigate many aspects of *Nasonia* biology, such as behavior [100], development [101], pheromones [102], sex determination [103], gene expression [104-106], venom evolution [107], and regulation by DNA methylation [16, 17, 98]. Therefore, a well-assembled reference genome of *Ng* will advance the utility of the system by the research community. In this study, I generated a high-quality reference genome assembly for *N. giraulti*, which will provide essential new genomic tools for *Nasonia* research.

Wasps in the genus *Muscidifurax* are also of interest for comparative genome studies, due to their close relationship to the model parasitoid genus *Nasonia*, which currently has genome assemblies for three species [9, 108], with *Muscidifurax* estimated to be 15 million years divergent [36]. *M. raptorellus* is an effective biological control agent of dipteran filth flies, including house fly (*Musca domestica* L.), stable fly (*Stomoxys calcitrans* [L.]), horn fly (*Hematobia irritans* [L.]), black dump fly (*Hydrotaea aenescens* [Weidemann]), and flesh fly (*Sarcophaga bullata* [Parker]) [27, 30, 34]. The application of insecticide, which is the primary control strategy, is of limited effectiveness due to the evolution of resistant genes in these pests. Parasitoid wasps have great potential as an alternative management strategy that is more environmentally friendly and sustainable [35].

Here, I report the first draft genome assembly of *M. raptorellus* using PacBio long-read sequencing. This well-assembled and annotated genome will provide an essential genetic toolkit for functional and evolutionary genomic studies in *M. raptorellus* and its sibling species. The high-quality reference genome could also inform and facilitate future genome manipulation in parasitoid wasps for more effective biological control strategies [109].

2.3 Materials and Methods

2.3.1 Sample source and insect rearing

The source of *M. raptorellus* used in this study was derived from a colony maintained by Dr. Chris Geden at the Center for Medical, Agricultural, and Veterinary Entomology, USDA Agricultural Research Service (Gainesville, FL). Genomic sequencing samples were collected from two independent colonies, both derived from the same USDA colony: one maintained in at

Auburn University College of Veterinary Medicine in Auburn, Alabama since 2019 (Aub sample), and the other one maintained at Koppert Biological Systems in the Netherlands (Kop sample) since 20 years ago. *M. raptorellus* was originally collected in 1965 from Chile, but referred to as *M. raptor* [110], and was subsequently described as *M. raptorellus* in 1970 [24], and was afterward distributed in North America for biological control efforts. The current colony was originally established from field-collected specimens on a New York poultry farm [111], and maintained in the Geden laboratory on housefly pupae. Samples from the colony were obtained by the Werren laboratory in 2016 and maintained on *Sarcophaga bullata* pupae, and sent to the Wang laboratory in Auburn, Alabama in 2019, and maintained on commercial *Sarcophaga bullata* pupae (flesh fly pupae) and at a constant temperature of 25°C and 24h constant light. The Kop sample was maintained on *Lucilia* spp. pupae for 20 years and was sent to the Verhulst laboratory in 2014 and maintained on *Calliphora* spp. pupae since at 25°C and 18h/6h light/dark. Both the Aub and the Kop samples were from the same fully inbred strain of *M. raptorellus*.

2.3.2 Genomic DNA extraction, library preparation, and sequencing

2.3.2.1 Genomic DNA extraction, library preparation, and sequencing

DNA was extracted from 24-hour male adults of the *N. giraulti* RV2X[u] strain. High molecular weight (HMW) genomic DNA (gDNA) was isolated using MagAttract HMW DNA Mini Kit (Qiagen, MD). The quality of extracted gDNA was examined on a Qubit 3.0 Fluorometer (Thermo Fisher Scientific, USA). The size distribution of the extracted gDNA was accessed using the genomic DNA kit on Agilent TapeStation 4200 (Agilent Technologies, CA).

A 10× Genomic library was prepared with the Chromium Genome Reagent Kits v2 on the 10× Chromium Controller (10× Genomics Inc., CA). In brief, HMW gDNA was diluted from

original concentrations to ~ 0.9 ng/µl with EB buffer. The diluted denatured gDNA, sample master mix and gel beads were loaded to the genomic chip, and then ran on 10× Chromium Controller to create Gel Bead-In-EMulsions (GEMs). After the run, 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 and Agilent TapeStation 4200. The prepared 10× genomic library was sequenced on a HiSeq X sequencer at the Genomic Services Lab at the HudsonAlpha Institute for Biotechnology. An Illumina short-read resequencing library (300 bp insert size) was made from genomic DNA samples extracted from six *N. giraulti* adult males (whole body), using TruSeq DNA Sample Prep Kit. Approximately 50× paired-end sequencing was done using the Illumina HiSeq 2000 platform.

2.3.2.2 Genomic DNA extraction, library preparation, and sequencing

High-molecular-weight (HMW) genomic DNA (gDNA) was extracted from adults of the *M. raptorellus* Aub sample using the Genomic-tip 20/G kit (Qiagen, Catalog No. 10223) with DNA concentration checked on a Qubit 3.0 Fluorometer (Thermo Fisher Scientific, United States). The size distribution and gDNA quality were assessed on an Agilent TapeStation 4200 machine (Agilent Technologies, CA) using the genomics kit (Agilent, Catalog No. 5067-5366). A total of 10 µg high-quality *M. raptorellus* genomic DNA was sheared into 20 Kb fragments, and the end damage was repaired. After sequencing adapter ligation, the DNA fragment was annealed with Sequencing Primer v2 and Sequel II DNA Polymerase and bound to the SMRTbell templates, and the library was constructed following SMRTbell Template Prep Kit v2 following the CCS HiFi library protocol (Pacific Biosciences, CA). The size distribution of the constructed library was assessed using LabChip GX Touch HT (PerkinElmer, MA, United States), and the final library

quantity was examined with a Qubit 3.0 Fluorometer (Thermo Fisher Scientific, United States). The PacBio library was sequenced on a PacBio Sequel II System at the HudsonAlpha Genome Sequencing Center.

HMW genomic DNA was diluted to ~ 0.8 ng/μL with elution buffer for 10× Genomics library preparation using Chromium Genome Reagent Kit v2 (10× Genomics, Inc., CA). The diluted denatured gDNA, sample master mix, and gel beads were loaded to the genomic chip following the protocol and then ran on a 10× Chromium Controller to generate Gel Bead-In-EMulsions (GEMs). The obtained GEMs were used for the subsequent incubation and cleanup. The Chromium i7 Sample Index served as the library barcode to provide linked information. After quality control with a Qubit 3.0 Fluorometer (Thermo Fisher Scientific, MA, United States) and Agilent TapeStation 4200 (Agilent Technologies, CA), the 10× genomic sequencing was performed on an Illumina NovaSeq 6000 machine.

HMW gDNA was extracted from a pool of thirty females of the *M. raptorellus* Kop sample that were collected at the black pupal stage (~16 days after egg-laying), using the Genomic-tip 100/G kit (Qiagen, Catalog No. 10243) combined with the Genomic DNA Buffer Set (Qiagen, Catalog No. 19060). The sample was ground to fine powder in liquid nitrogen by a plastic pestle, and the total DNA was extracted following the protocol provided by the manufacturer. After extraction, genomic DNA was sheared into an 8-30 Kb range by using g-TUBE (Covaris) following the manufacturer's protocol. The quality and quantity of sheared genomic DNA were checked by gel electrophoresis with 1.5% TAE agarose gel stained with Midori Green (NIPPON Genetics) and by spectrophotometry (Nanodrop™ 2000, Thermo Fisher). The genomic DNA was measured, and quality controlled at Novogene Co., Ltd. (Beijing, China). SMRTbell library

templates were prepared for long-read sequencing on the PacBio Sequel system using three flow cells, to generate up to 70 Kb long reads with an average read length of 12-15 Kb. A total of 1.57 million high-quality subreads were obtained, with an estimated read depth of 55.8×.

2.3.3 Genome assembly, polishing, and assessment

2.3.3.1 *N. giraulti* genome assembly and assessment

The raw sequencing reads from both 10× library and the Illumina resequencing library were checked for sequencing quality by FastQC v11.5 [112] before being used for genome assembly. The genome assembly strategy of *N. giraulti* includes the constructions of three draft *de novo* assemblies using different assemblers and a final step to reconcile three draft assemblies into a final high-quality assembly. The first *de novo* assembly of *N. giraulti* genome was performed with the Supernova 2.0 assembler [113] using linked reads from 10× Genomics library. To achieve the best *de novo* assembly result, I examined a grid of barcode subsampling percentage parameters and the maximum number of input reads, including no barcode subsampling with all linked reads. A second *de novo* assembly was conducted by MEGAHIT v1.2.9 [114]. The 10× linked reads were transferred to regular paired-end Illumina sequencing reads by trimming the barcode sequences and potential adaptor sequences with Trimmomatic v0.38 [115]. All trimmed sequencing reads were used for the second *de novo* assembly using MEGAHIT v1.2.9 [114] with all default parameter settings. In addition, a third *de novo* assembly (ngirB_goodCOV) was generated by velvet v1.2.10 [116] using sequencing reads from the Illumina short-read resequencing library.

A final high-quality assembly was generated by merging these three draft assemblies using an assembly reconciliation tool Metassembler v1.5 [117]. All reverse complementary scaffolds

with same length, coverage, A/T/C/G counts, as well as the duplicated scaffolds identified by self-BLAT version 35 [118] were removed from the final assembly. To estimate the contiguity and completeness of my genome assembly, two evaluation pipelines were performed: (1) genome sequencing reads were aligned to my assembly with BWA-MEM aligner version 0.7.17 [119]; (2) transcriptomic data of different developmental stages and sexes were mapped to the current assembly using Tophat v2.1.1 [120]; (3)The BUSCO [121] score of my genome assembly was calculated by aligning to arthropoda_odb9 with a total of 1,066 orthologs.

2.3.3.2 *M. raptorellus* genome assembly, polishing, and assessment

The raw sequencing reads (Aub sample) from both PacBio library and 10× Genomics library were checked for sequencing quality using FastQC [122] before genome assembly. *De novo* genome assembly for *M. raptorellus* Aub sample was performed by Supernova 2.1.1 [123] assembler using 400 million reads subsampled from the total amount of reads generated from the 10× Genomics library. Filtered HiFi PacBio reads were assembled by hifiasm v0.13 [124] and HiCanu v2.1.1 [125], dedicated assemblers using long-read sequencing reads. The Kop PacBio data was assembled using Canu v2.1 [126]. Aub and Kop cultures have identical mitochondrial genomes (100% sequence identity) with only one 11-bp indel. The Aub 10× Genomics reads were aligned to the repeat masked Kop assembly using Longranger v2.1.6 [127] software suite with ALIGN pipeline. 58,350 SNPs were called by UnifiedGenotyper in the Genome Analysis Toolkit (GATK) [128, 129]. SNP positions in repetitive regions, and variants outside the coverage depth threshold (120-500 bp) were filtered out using BEDTools v2.30.0 [130]. A total of 11,523 homozygote SNPs between Aub and Kop were identified, and the percentage of fixed differences in the nuclear genome was estimated to be 0.0038%. To achieve the best assembly, these draft

assemblies with different assemblers from both Aub and Kop samples were merged into a draft assembly using an assembly combination tool quickmerge v0.3.0 [131]. Potential bacterial contaminations were checked and removed using a pipeline described in the previous research [108]. The draft assembly was polished to yield a final high-quality assembly with the Illumina short-reads for indel correction using Pilon v1.23.0 [132]. The final genome assembly was evaluated based on the N50 size of contigs, RNA-seq read mapping percentages, and genome completeness was assessed by BUSCO version 4.0.6 [133]. The BUSCO scores were calculated using arthropoda_odb10 with a total of 1,013 orthologs.

2.3.4 RNA-seq data processing and transcriptome assembly

2.3.4.1 Total RNA extraction, library preparation, and sequencing of developmental stage samples in *N. giraulti*

Male and female *N. giraulti* RV2X[u] strain samples were collected at five developmental stages: 0-10hr early embryo, 14-24hr late embryo, 44-54hr larva, yellow pupa and 1-day adult. *Sarcophaga bullata* (*S. bullata*) pupae were inserted into foam plugs, with only anterior available for oviposition. To obtain the male samples, two host pupae were provided to each virgin female, allowing host feeding for 48 hr. These unmated females lay unfertilized eggs and produce all-male progeny. For female sample collection, mated females will produce more than 90% daughters under the experimental conditions, allowing the expression quantification of mostly female progeny for embryo and larva stages. Six individuals were pooled per stage, except early embryo for which 40 individuals were pooled due to the small size. All samples were homogenized in 1 mL TRIzol and stored at -80°C freezer. Total RNA extractions, quantification, library preparation and sequencing protocol were previously described [134].

2.3.4.2 Total RNA extraction, RNA-seq data processing, and transcriptome assembly

Total RNA was isolated from adult whole-body samples of adult male and female *M. raptorellus* in three biological replicates for each sex from samples collected in the Werren laboratory. The RNA extraction, quantification, library preparation, and sequencing protocol were previously described [107]. A total of 308,475,537 reads were obtained from six samples. FastQC [122] was used for quality control of raw RNA-seq data. The paired-end RNA-seq reads were processed with Trimmomatic v0.38 [115]. After trimming the potential adapter sequences, I performed *de novo* assembly of the *M. raptorellus* transcriptome using Trinity v2.4.0 [135], and pre-aligned transcripts were annotated by Cufflinks v2.2.1 [136].

2.3.5 Repeat annotation

A *de novo M. raptorellus* repeat database was constructed using RepeatModeler v2.0.1 [137] with the default parameters, which employs three complementary computational programs, RECON v1.0.8 [138], RepeatScout v1.0.5 [139], and Tandem Repeats Finder (TRF) [140] to annotate repetitive elements in my genome assembly. RepeatScout is a *de novo* repeat finder to identify highly conserved repetitive elements, while RECON can find less conserved elements. TRF is a program to locate and display tandem repeats. The high-quality library of transposable element (TE) families was then used to mask homologous repeats and low complexity DNA sequences using RepeatMasker v4.0.6 [141] with RMBlast v. 2.10.0 as the default search engine.

2.3.6 Genome annotation

2.3.6.1 Gene annotation for *N. giraulti* genome

The annotation of the *N. giraulti* genome was performed using MAKER version 2.31.9 [142] based on the following pipeline: (1) A custom *N. giraulti* repeat database constructed with RepeatModeler v.1.08 using the default parameter settings, with low complexity repeat regions soft-masked by MAKER; (2) A *de novo* assembly of the *N. giraulti* transcriptomes by Trinity v 2.4.0 [143] and pre-aligned transcripts annotated by Cufflinks [144]. For gene annotation, *ab initio* gene prediction algorithms were trained to predict gene models using protein and transcriptome evidence by EST2GENOME and PROTEIN2GENOME in MAKER. After filtered based on gene length and quality, the predicted genes were then used to train both the SNAP and the AUGUSTUS gene predictors. The results were fed to MAKER to repeat this procedure for another round, to generate the final predicted genes in *N. giraulti* genome.

2.3.6.2 Gene prediction and functional annotation for *M. raptorellus* genome

To annotate the structures and functions of the *M. raptorellus* genome, I integrated *ab initio* and RNA-seq based methods to predict the genes in repeat-masked assembly. For RNA-seq prediction, the trimmed RNA-seq reads were aligned to the repeat-masked genome assembly using Tophat v2.1.1 [145], and then assembled into transcripts using cufflinks v 2.2.1 [136] with default parameters. In addition, *de novo* assembly of *M. raptorellus* transcriptome was achieved by Trinity v2.4.0 [135]. The annotation of the genome assembly was performed using MAKER v2.31.9 [146] annotation pipeline. Gene models were predicted using *ab initio* gene prediction algorithms with protein and transcriptome evidence by EST2GENOME and PROTEIN2GENOME procedures in MAKER. The generated GFF3 file and assembled transcriptome from RNA-seq prediction were provided as expressed sequence tags (ESTs) evidence. The Arthropoda_odb10 dataset served as protein homology evidence. After evaluation and filtering with evidence scores, the predicted

genes were used to train both the SNAP [147] and the AUGUSTUS [148, 149] gene predictors. Two additional iterations were performed to generate the final predicted gene models for *M. raptorellus* genome. A homology-based gene prediction tool, Gene Model Mapper (GeMoMa) [150], was also utilized to annotate the coding genes in *M. raptorellus* using well-annotated *Nasonia vitripennis* OGS2 (official gene set 2) [151] as the protein reference.

2.3.7 Comparative genomic analysis

2.3.7.1 Comparative genomic analysis between *N. giraulti* and *N. vitripennis* genomes

To compare the genome structure between *N. giraulti* and *N. vitripennis* genomes, I conducted whole-genome alignment of my *Ng* assembly and the recent *Nv* genome assembly of [152] using NUCmer in the MUMmer v4.0 program suite with default p parameter settings [153]. The pairwise alignments (match length longer than 500bp) between *Ng* scaffolds and *Nv* chromosomes were visualized using Mummerplot [153].

To identify the candidate *Ng*-specific genes, genes with no assigned orthogroup between *N. giraulti* and *N. vitripennis* were generated using OrthoFinder v2.2.7 [154]. The *Ng* genes identified to have no assigned orthogroup with *Nv* were potential candidates for *Ng*-specific genes. To ensure the absence of these candidates in the *Nv* genome, protein sequences of these candidate *Ng*-specific genes were BLASTed to two *Nv* genome assemblies, including the *Nv* reference genome assembly (GCA_000002325.2) [98] and the newly released *Nv* PacBio genome assembly (GCA_009193385.1) [155] with an E-value cutoff of 1E-5 and protein length larger than 30. Genes with no BLAST hit to the two *Nv* genome assemblies were then aligned to the annotated *Ng* transcripts. The annotated *Ng* transcripts were generated with available *Ng* RNA-seq data from

different developmental stages and sexes (Embryo stage of 0-10 hr, 10-24 hr, 24-36 hr, female and male pupa and adult) using Cufflinks [144]. Genes with support from annotated transcripts were kept as *Ng*-specific candidates. The protein sequences of these genes were aligned to the *N. vitripennis* (Nvit_psr_1.1) [152] and *Trichomalopsis sarcophagae* assemblies using tBLASTn with an E-value cutoff of 1E-5. The final genes were annotated using both Blast2GO and KofamKOALA with an E-value cutoff of 1E-4.

2.3.7.2 Comparative genomic analysis between *M. raptorellus* and *N. vitripennis* genomes

To compare the genome structure between *M. raptorellus* and *N. vitripennis* genomes, the homologous regions in these two genomes were identified using MCScanX [156] with default parameters, which is a Python package for synteny detection and evolutionary analysis. The inferred gene pairs and linked relationships were visualized and placed in the context of whole-genome collinearity using a genomic circle generated by Circos [157]. The chromosome-level genome assembly of *N. vitripennis* (Nvit_psr_1.1) [152] was downloaded at NCBI Assembly with accession number GCA_009193385.2.

2.3.8 Phylogenetic analysis

2.3.8.1 Phylogenetic analysis of *N. giraulti* and eight sequenced arthropod species

I conducted a phylogenomic analysis using my assembled *N. giraulti* genome and eight other sequenced insect genomes, including the fruit fly *Drosophila melanogaster* (GCA_000001215.4) [158], pea aphid *Acyrthosiphon pisum* (GCA_005508785.1) [159], honeybee *Apis mellifera* (GCA_003254395.2) [160], water flea *Daphnia pulex* (GCA_000187875.1) [161], human lice *Pediculus humanus* (GCA_000006295.1) [162], mosquito

Anopheles gambiae (GCA_000005575.1) [163], silk moth *Bombyx mori* (GCA_000151625.1) [164], and jewel wasp *Nasonia vitripennis* (GCA_000002325.2) [98]. Homologous genes among these 9 genomes were identified using OrthoFinder [154, 165] with default settings. The protein sequences of the core single-copy genes shared in all 9 genomes were aligned with MAFFT v7.407 [166]. ProtTest 3 [167] was used to evaluate The best-fit model of protein evolution. The Maximum Likelihood (ML) phylogenetic tree of the concatenated protein sequence was inferred by using RAxML v8.2 [168] with the VT protein model (best fit model identified by ProtTest 3) and 1,000 rapid bootstrap replicates.

2.3.8.1 Phylogenetic analysis of *M. raptorellus* and nine representative Hymenoptera insect species

To investigate the phylogenetic relationship between *M. raptorellus* and other Hymenoptera insect species, nine representative species (jewel wasp *Nasonia vitripennis*, honeybee *Apis mellifera*, turnip sawfly *Athalia rosae*, fig wasp *Ceratosolen solmsi marchali*, Indian jumping ant *Harpegnathos saltator*, Braconid wasp *Microplitis demolitor*, wood wasp *Orussus abietinus*, red paper wasp *Polistes canadensis*, and minute polyphagous wasp *Trichogramma pretiosum*) were selected from 40 Hymenoptera species in OrthoDB v10.1 (https://www.orthodb.org/) [169]. A total of 4,390 1:1 single-copy orthologs among these nine genomes were identified. The protein sequences for *M. raptorellus* were aligned to *N. vitripennis* using BLASTp alignments with a minimum of 60% sequence identity, and 3,662 1:1 orthologs were identified. Subsequently, the protein sequences of the single-copy orthologs in the nine species were extracted from the OrthoDB fasta file, and *M. raptorellus* protein sequences of these genes were extracted from my genome assembly. The protein sequences across the selected

Hymenoptera species and *M. raptorellus* were independently aligned with MAFFT v7.407 [166]. The protein alignments were concatenated for phylogenomic analysis. ProtTest 3 [167] was used to estimate the best protein model of protein evolution. The maximum-likelihood (ML) phylogenetic tree was finally built with the concatenated protein sequence by using RAxML v8.2 [168] with the best JTT protein model. 1,000 rapid bootstrap replicates were applied for evaluation of their branch supports. The tree was displayed by FigTree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/).

2.4 Results

2.4.1 Genome assembly and assessment

2.4.1.1 Genome assembly and assessment for *N. giraulti*

The draft *de novo* assembly was found to contain some artifacts, which was also reported for this assembler in a recent study [170]. I removed all the identical or nearly identical scaffolds as well as reverse complementary scaffolds prior to subsequent analyses. All these three *de novo* assemblies generated from different algorithms were further reconciled using an assembly reconciliation tool Metassembler [117]. To identify the mitochondrial scaffold, I aligned the final assembly to the previously assembled mitochondrial genome of *N. giraulti*. Scaffolds with high identity (>90%) and high coverage (>16,000×) were assigned as mitochondrial scaffolds (Figure 8). The detailed genome statistics of my final assembly of *N. giraulti* and all other available wasp genomes, including previously assembled genomes, are listed in Table 3. The final genome assembly of *N. giraulti* is a total of 259,040,977 bp in 3,160 scaffolds. The contig N50 is 34,917 bp and the scaffold N50 is 545,346 bp, respectively.

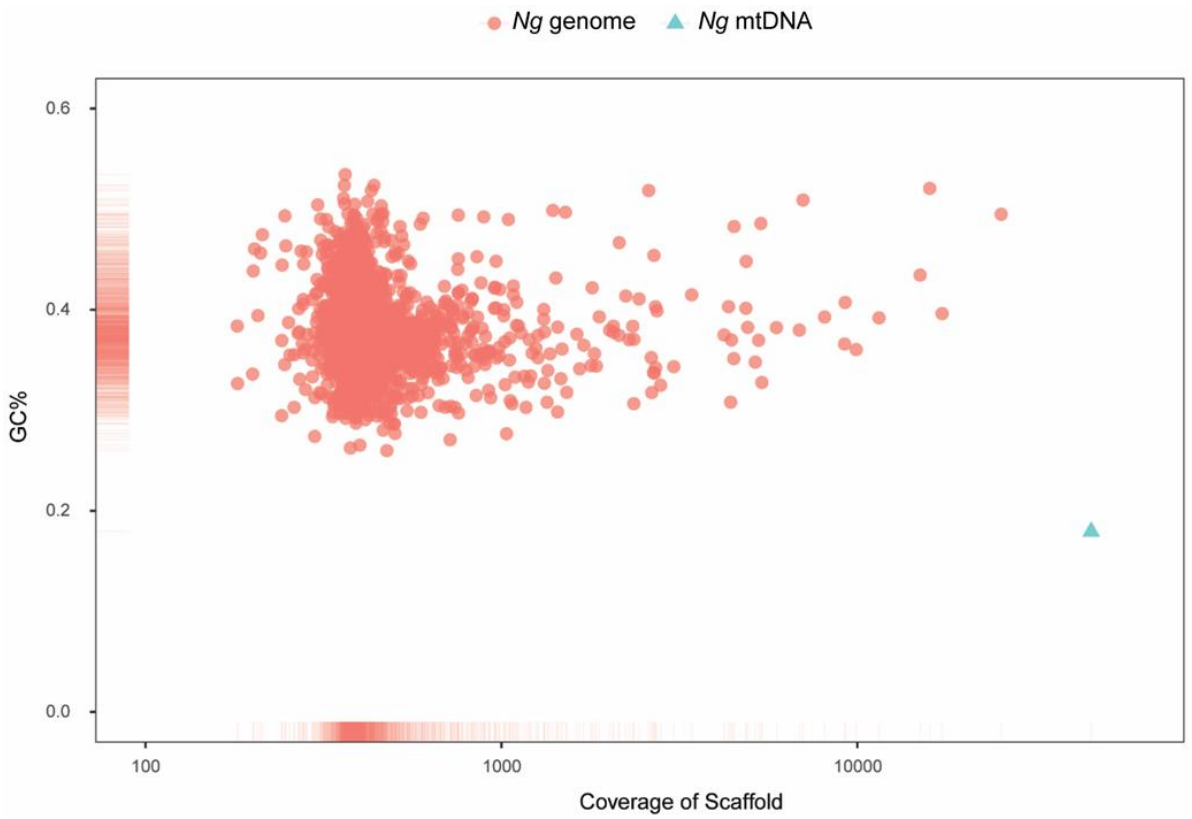


Figure 8. Scaffold median coverage and GC content in the *N. girault*i genome assembly.

Scatterplot of median coverages on the *x*-axis and GC percentage on the *y*-axis for each *N. girault*i scaffold. Red dots represent *Ng* genome scaffolds, and green triangle represents *Ng* mtDNA.

Table 3. Statistics of the *N. girault*i genome assembly compared to other wasp species

Genome assembly	**Ngir_v5**	**Ngir_1.0**	**Nvit_2.1**	**Nlon_1.0**	**Tsac_v1**
Species	*N. girault*i	*N. girault*i	*N. vitripennis*	*N. longicornis*	*T. sarcophagae*
No. of scaffolds	3,160	4,912	6,098	5,214	4,0891
No. of contigs	14,039	373,227	25,484	385,077	57,930
Scaffold length (bp)	259,040,977	283,606,953	295,780,872	285,726,340	236,484,274
Contig length (bp)	255,292,562	178,561,037	238,616,307	181,397,296	235,211,350

Gap percentage	1.5 %	37.0 %	19.3 %	36.5 %	0.5 %
Scaffold N50 (bp)	545,346	759,431	897,131	758,407	22,350
Contig N50 (bp)	34,917	1,973	18,840	1,877	9,957
Scaffold N90 (bp)	46,391	62,470	46,455	59,334	2,779
Contig N90 (bp)	9,262	163	4,180	162	1,943
Scaffold maximum length (bp)	6,445,087	9,412,112	33,571,687	9,412,414	350,161
Contig maximum length (bp)	385,696	35,702	226,699	39,258	140,646
Percentage of scaffold > 50 Kb	89.51	91.30	89.44	91.02	26.39
GC contents	38.05 %	39.40 %	38.33%	39.02 %	40.29 %
BUSCO completeness	98.6%	97.0 %	97.0 %	92.8 %	98.6 %
GenBank assembly accession No.	QLYP00000000	GCA_000004775.1	GCA_000002325.2	GCA_000004795.1	GCA_002249905.1

The 10× Genomics reads were aligned to the final assembly to compute the summary statistics. The average scaffold coverage is 671.87× and the GC-content is 41.4 % (Figure 8). RNA-seq reads from different development stages (see Materials and Methods) in *N. giraulti* samples were also aligned to the final assembly with an average mapping percentage of 97%, indicating a high-quality assembly of *N. giraulti* genome. To assess the completeness of this genome, the BUSCO scores of all five wasp genome assemblies were generated (Table 3). The BUSCO completeness score for the current assembly of *N. giraulti* is 98.6% (N=1,066; Complete: 98.6%; Fragmented:0.4%; Missing:1.0%), indicating a high level of completeness of my genome assembly.

2.4.1.2 Genome assembly and assessment for *M. raptorellus*

Two independent PacBio libraries were constructed for the assembly of *M. raptorellus* genome (see Materials and Methods). The PacBio Sequel II HiFi reads (14,992,520,996 bp) generated from Aub sample were assembled using hifiasm and HiCanu, and the Kop PacBio data (17,675,696,457 bp) was assembled using Canu (see Materials and Methods). The genome size of all three assemblies ranges from 315.7 to 316.9 Mbp (Table 4), which is very close to the estimated size from 10× Genomics data using Supernova based on K-mer profiles (315 Mbp), indicating high confidence in the genome size. The merged genome has significant improvement over individual assemblies, in terms of reduction in the number of contigs (from 527 to 226), increase in contig N50 (from 1.5 Mb to 4.7 Mb) and maximum contig length (from 8.7 Mb to 21.2 Mb), as well as a reduced proportion of duplicated BUSCO (from 2% to 1.1%), without sacrificing the DNA and RNA sequencing mapping rate (Table 4). The final assembled genome is 313,931,273 bp in length with 226 scaffolds (GC content is 40.06%) and a circularized mitochondrial genome (GenBank accession number: MT985329) [171]. The contig N50 is 4,673,378 bp, and the BUSCO completeness score is 97.9% (96.8% single-copy, 1.1% duplicated, 0.5% fragmented, and 1.6% missing). The adult RNA-seq reads were aligned to the *M. raptorellus* assembly using Tophat [120], and 97% of the reads were mapped to the genome. The 10× Genomics short-read data were also mapped to the genome assembly, and the alignment rate was 96.68%. The proportion of the genome with zero depth is 0.06%. The assembly and mapping statistics suggest that the quality of my assembly is high in both genome completeness and continuity (Table 4).

Table 4. Summary statistics of the *Muscidifurax raptorellus* genome assemblies

Genome assembly	Aub_Hifiasm	Aub_HiCanu	Kop_PacBio	Final
Data and coverage				

PacBio sequencing data	15.0 Gb Sequel II CCS reads	15.0 Gb Sequel II CCS reads	17.7 Gb Sequel CLR reads	-
Illumina sequencing data	81.6 Gb	81.6 Gb	20.6 Gb	-
Genome coverage	CCS: 48× Illumina: 260×	CCS: 48× Illumina: 260×	CLS: 56× Illumina: 66×	PacBio: 104× Illumina: 326×
Assembly statistics				
Genome size (bp)	315,727,724	316,569,142	316,926,883	313,931,273
No. of scaffold	489	527	384	226 + chrM
Scaffold N50 (bp)	1,479,014	2,597,351	2,784,708	4,673,378
Contig N50 (bp)	1,479,014	2,597,351	2,784,708	4,673,378
Maximum contig length (bp)	8,668,935	14,498,644	14,510,203	21,163,931
Completeness				
BUSCO completeness	97.90%	97.90%	97.90%	97.90%
single-copy BUSCO	95.90%	95.90%	96.20%	96.80%
duplicated BUSCO	2.00%	2.00%	1.70%	1.10%
fragmented BUSCO	0.50%	0.50%	0.50%	0.50%
missing BUSCO	1.60%	1.60%	1.60%	1.60%
Mapping statistics				
% of gDNAseq reads mapped	96.67%	96.74%	96.71%	96.68%
% of gDNAseq covered positions	99.99%	99.91%	99.78%	99.94%
Adult RNA-seq, all mapped	97.41%	97.49%	97.42%	97.24%
Adult RNA-seq, uniquely mapped	95.02%	94.78%	95.19%	94.56%

2.4.2 Repeat annotation

2.4.2.1 Repeat annotation for *N. giraulti* genome

In my current *N. giraulti* assembly, I have identified a total repeat content of 83,899,561 bp, by using an *Ng*-specific repeat library, consisting of approximately 32.39 % of the genome assembly (Table 5). Among the classified repetitive elements, the top three repeat types are DNA elements (7.58%), LINEs (6.71%), and SINEs (6.68%) (Table 5).

Table 5. Summary of repetitive element content found in the *N. giraulti* genome assembly

Categories	Number of elements	Length occupied (bp)	Percentage occupied (%)
SINEs*	586	99,652	0.04
LINEs*	18,830	17,387,298	6.71
LTR* elements	23,401	17,311,094	6.68
DNA elements	41,707	19,644,786	7.58
Small RNA	25	4,445	0.00
Satellites	1,824	745,156	0.29
Simple repeats	130,377	5,623,109	2.17
Low complexity	8,384	400,042	0.15

*SINEs: Short interspersed nuclear elements

*****LINEs: Long interspersed nuclear elements

*LTR: Long terminal repeat

2.4.2.2 Repeat annotation for *M. raptorellus* genome

Repetitive regions accounted for 40% of the *M. raptorellus* genome with a total length of 126 Mbp based on the *M. raptorellus*-specific repeat database (Table 6). The proportion of repeat

regions is similar to that in *Nasonia vitripennis*, a jewel wasp species in the *Nasonia* genus (40.27%). LINEs (6.0%) and Gypsy (7.2%) elements are the most abundant classes in *M. raptorellus*, both with significantly higher abundance compared to those in *N. vitripennis* (Table 6).

Table 6. Summary repeat element classes in *Muscidifurax raptorellus* and *Nasonia vitripennis* genomes

	Muscidifurax raptorellus		***Nasonia vitripennis***	
	# of elements	Length (%)	# of elements	Length (%)
Retroelements				
Penelope	913	327,347 (0.1%)	1,065	317,344 (0.11%)
LINEs*	17,663	18,752,397 (5.97%)	14,783	14,534,403 (5.07%)
L2/CR1/Rex	7,925	7,800,224 (2.48%)	6,577	5,837,873 (2.03%)
R1/LOA/Jockey	6,013	6,346,905 (2.02%)	4,008	3,617,626 (1.26%)
R2/R4/NeSL	0	0 (0%)	151	406,741 (0.14%)
LTR* elements				
BEL/Pao	1,496	1,683,607 (0.54%)	883	993,809 (0.35%)
Ty1/Copia	1,992	1,660,186 (0.53%)	2,396	2,624,950 (0.91%)
Gypsy/DIRS1	17,473	22,464,754 (7.16%)	9,516	9,681,184 (3.37%)
DNA transposons				
hobo-Activator	5,99	261,203 (0.08%)	646	248,902 (0.09%)
Tc1-IS630-Pogo	5,037	2,550,239 (0.81%)	3,453	4,340,897 (1.51%)
PiggyBac	443	257,410 (0.08%)	549	293,323 (0.1%)

Tourist/Harbinger	115	54,423 (0.02%)	61	35,468 (0.01%)
Rolling-circles	2,391	1,701,169 (0.54%)	5,841	2,970,560 (1.04%)
Unclassified	136,718	55,208,650 (17.59%)	136,074	63,582,769 (22.16%)
Simple repeats	150,695	6,103,642 (1.94%)	132,857	5,673,959 (1.98%)
Low complexity	10,350	497,537 (0.16%)	8,588	400,956 (0.14%)
Total	359,224	125,669,693 (40.03%)	327,448	115,560,764 (40.27%)

*****LINEs: Long interspersed nuclear elements

*LTR: Long terminal repeat

2.4.3 Gene annotations

2.4.3.1 Gene annotation for *N. giraulti* genome

After all the repeat regions were soft-masked by MAKER; the final annotation resulted in 14,777 protein coding genes. When comparing the annotated genes in *Ng* and *Nv*, there are 10,640 one-to-one orthologous genes shared between *Ng* and *Nv*, and 83.7% *Ng* genes were assigned in orthogroups between *Ng* and *Nv*.

2.4.3.2 Gene annotation for *M. raptorellus* genome

After repeat regions were soft-masked, the first-round MAKER annotation based on Trinity output generated 18,392 gene models. Subsequent MAKER iterations resulted in 10,362 protein-coding genes supported by both RNA-seq and gene prediction algorithms. Among them, 7,520 single-copy orthologs were identified between *M. raptorellus* and *N. vitripennis*. To evaluate the completeness and quality of predicted genes, I compared the gene length distributions of the 7,520 orthologs and found an average CDS length of 1,008 bp in *M. raptorellus* (standard deviation

= 1,585) and 1,035 bp in *N. vitripennis* (standard deviation = 1,631). The 3,662 single-copy 1:1 orthologs between *M. raptorellus* and nine other Hymenopteran species also have similar CDS length distributions (Figure 9), indicating good gene model quality for these orthologs in *M. raptorellus*. To perform the gene annotation using an independent approach, 9,520 protein-coding genes (with 20,493 transcript isoforms) were annotated using the homology-based gene predictor GeMoMa [150]. 417 tRNA (transfer RNA) genes and 83 rRNA (ribosomal RNA) gene clusters were also annotated in the genome.

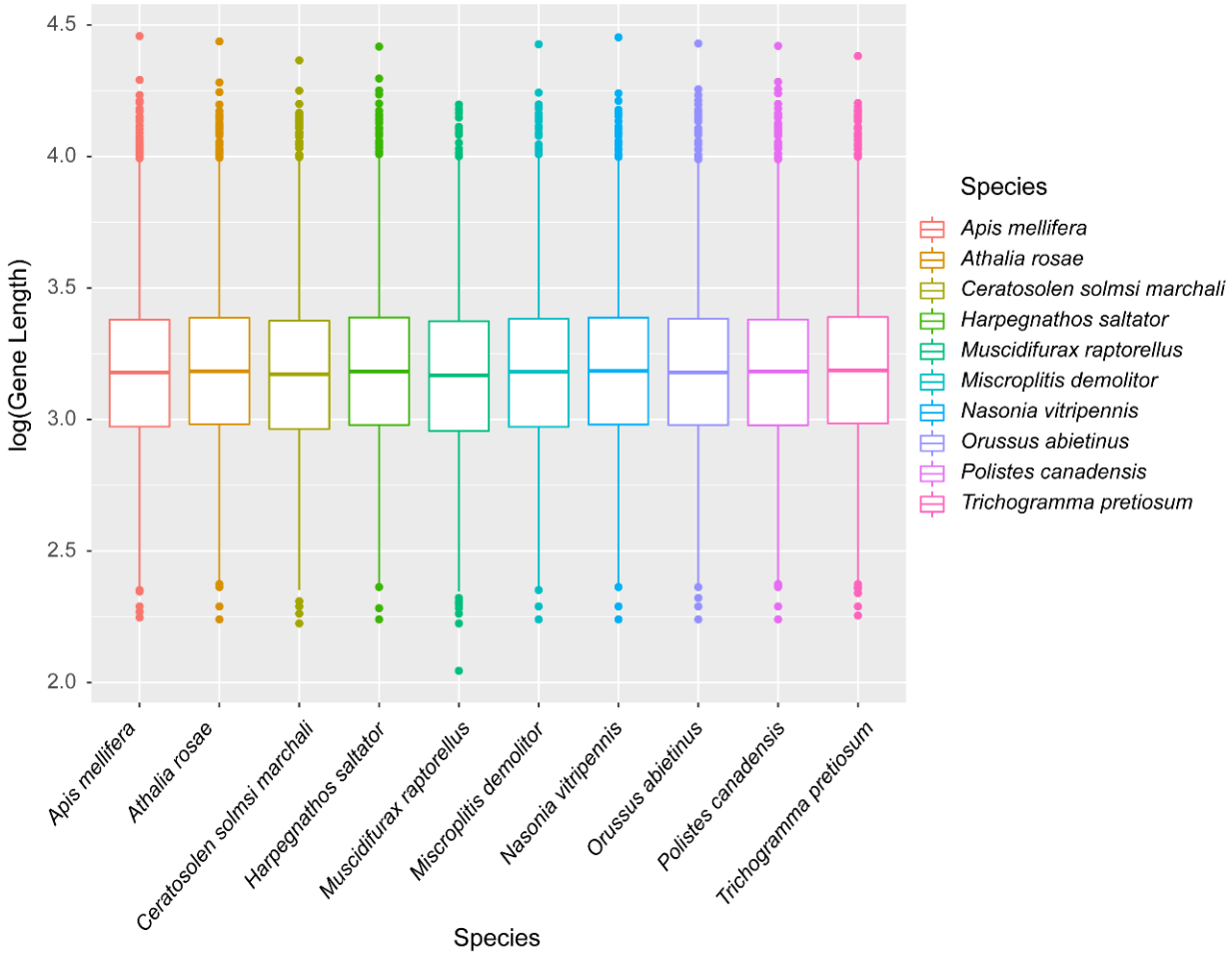


Figure 9. Comparison of the gene length distributions for 3,662 shared 1:1 single copy orthologs in *M. raptorellus* and nine representative Hymenopteran species.

2.4.4 Genome comparisons

2.4.4.1 Genome comparison between *N. giraulti* and *N. vitripennis*

Ng scaffolds were mapped to each chromosome of the *Nv* assembly (GCA_000002325.2) [98] with BWA-MEM aligner [119]. Overall, the alignments are consistent between *Ng* and *Nv* with a few insistencies (Figure 10). A total of 1,137 *Ng* scaffolds were aligned to the *Nv* chromosomes (Table 7 and Table 8), accounting for 89.3% of the total chromosome length in *Nv*. The average sequence identity in these aligned regions is 93.23%. As a useful tool for comparative analysis and interspecific mapping, I provide a set of 5,147,972 high-quality single nucleotide polymorphisms between the *Ng* and *Nv* genome assemblies. The SNPs fall 6.1% percent into exons (3.4% of these are synonymous and 2.7% are nonsynonymous), 16.3% percent in introns, and 77.6% percent are intragenic. These represent either species-specific or strain-specific differences, which will be resolved in the resequencing of multiple *Ng* strains in future work.

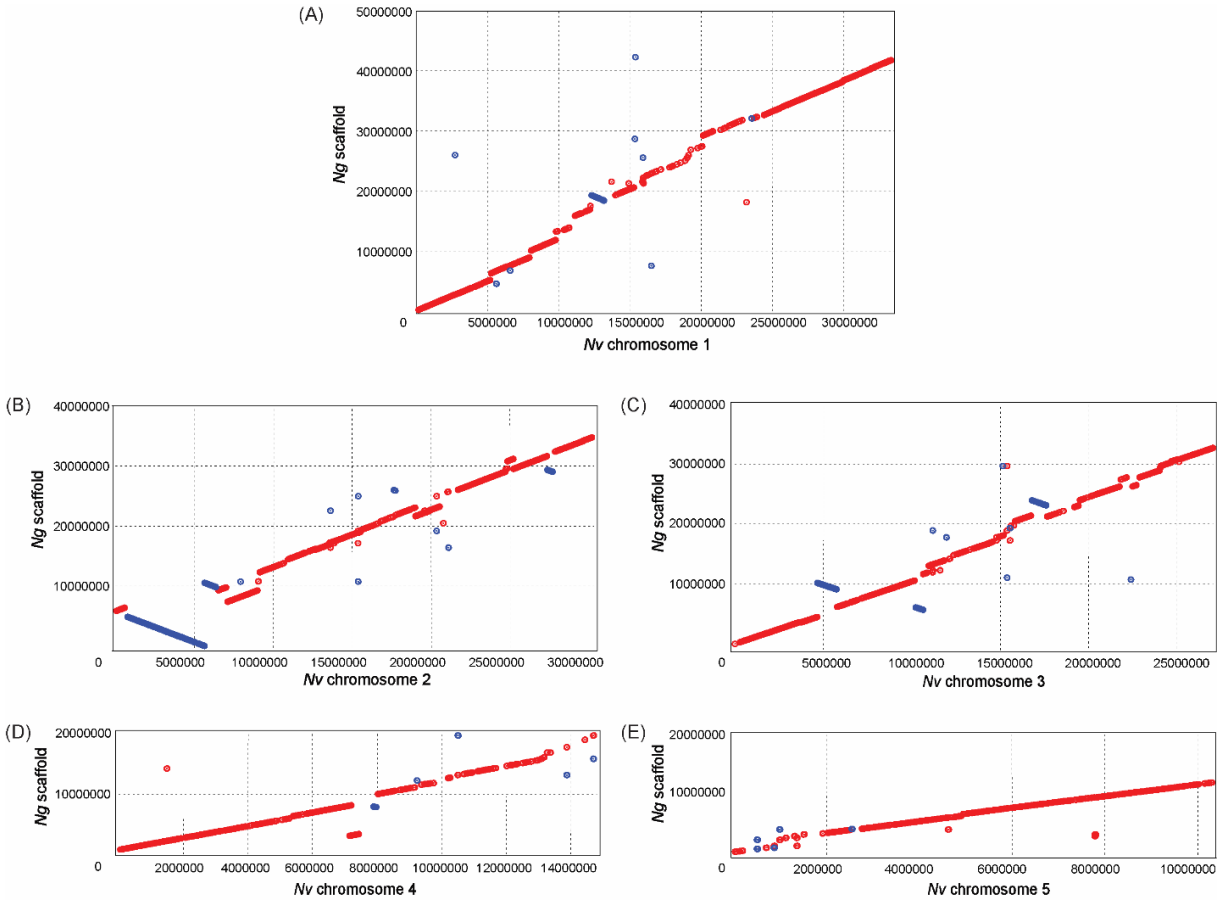


Figure 10. Chromosome level alignment between *N. giraulti* scaffolds and *N. vitripennis* chromosomes.

Dot plot showing comparison between *Ng* and *Nv* genomes. Red stands for a forward match and blue stands for a reverse match.

Table 7. Alignment length and percentage of *N. giraulti* scaffolds to *N. vitripennis* genome

Nv chromosome	Number of *Ng* scaffolds	Length (bp)	Sequence identity	Chromosome coverage (all)	Chromosome coverage (top 10)
Chr1	490	29,245,964	93.36%	87.11%	39.32%

Chr2	324	27,672,334	93.19%	91.34%	66.15%
Chr3	320	24,746,805	93.13%	91.53%	59.82%
Chr4	371	12,841,562	93.09%	86.72%	69.24%
Chr5	232	9,050,462	93.43%	87.74%	79.30%
Total	1,737	103,557,127	93.23%	89.30%	58.60%

Table 8. The top 10 *N. giraulti* scaffolds covering each of *N. vitripennis* chromosomes

***Nv* chromosome**	***Nv* chromosome length**	***Ng* scaffold name**	***Ng* scaffold length**	**Total match length**	**Average identity**
Chr1	33,571,687	SCAFFOLD14	2,118,795	1,855,993	91.43%
		SCAFFOLD20	1,784,763	1,613,333	93.16%
		SCAFFOLD17	1,992,319	1,495,766	93.65%
		SCAFFOLD16	2,032,568	1,396,719	93.12%
		SCAFFOLD11	2,387,156	1,379,907	92.37%
		SCAFFOLD33	1,338,229	1,259,619	93.34%
		SCAFFOLD34	1,334,651	1,203,410	93.49%
		SCAFFOLD30	1,374,614	1,202,006	91.63%
		SCAFFOLD41	1,098,727	998,038	93.93%
		SCAFFOLD51	930,897	796,142	93.80%
Chr2	30,297,376	SCAFFOLD1	6,445,087	5,124,810	92.76%
		SCAFFOLD4	4,081,301	3,812,105	92.98%
		SCAFFOLD7	3,257,095	2,989,469	91.66%
		SCAFFOLD12	2,386,129	2,225,524	93.03%
		SCAFFOLD29	1,416,399	1,249,871	93.76%
		SCAFFOLD31	1,357,603	1,133,339	92.07%
		SCAFFOLD28	1,438,783	936,736	92.30%
		SCAFFOLD47	988,996	912,318	94.59%
		SCAFFOLD48	974,174	829,937	93.48%
		SCAFFOLD55	901,160	828,761	93.80%

Chr3	27,037,145	SCAFFOLD6	3,380,798	3,081,875	93.49%
		SCAFFOLD9	3,152,063	2,851,521	93.52%
		SCAFFOLD21	1,778,028	1,709,509	93.56%
		SCAFFOLD23	1,747,365	1,644,700	92.61%
		SCAFFOLD24	1,726,916	1,568,524	91.65%
		SCAFFOLD15	2,064,749	1,480,470	92.03%
		SCAFFOLD38	1,106,630	1,040,392	91.98%
		SCAFFOLD36	1,137,620	979,265	90.51%
		SCAFFOLD35	1,161,403	954,878	94.10%
		SCAFFOLD1	6,445,087	862,861	90.86%
Chr4	14,808,294	SCAFFOLD3	5,274,031	2,846,201	92.24%
		SCAFFOLD26	1,575,316	1,505,432	93.49%
		SCAFFOLD32	1,347,799	1,258,721	91.59%
		SCAFFOLD27	1,476,458	1,183,034	93.97%
		SCAFFOLD46	1,000,992	926,513	93.08%
		SCAFFOLD56	894,343	840,807	91.17%
		SCAFFOLD70	697,421	597,759	92.42%
		SCAFFOLD80	606,933	545,622	92.43%
		SCAFFOLD67	716,332	292,228	93.97%
		SCAFFOLD124	412,340	256,496	92.39%
Chr5	10,315,142	SCAFFOLD5	3,899,944	2,156,984	94.06%
		SCAFFOLD3	5,274,031	2,149,273	94.13%
		SCAFFOLD13	2,227,120	2,076,974	92.33%
		SCAFFOLD22	1,756,827	1,411,978	94.50%
		SCAFFOLD101	507,482	234,307	92.14%
		SCAFFOLD78	614,452	45,731	95.88%
		SCAFFOLD903	38,670	37,454	94.52%
		SCAFFOLD187	259,829	27,322	91.96%
		SCAFFOLD598	76,236	24,715	91.81%
		SCAFFOLD625	72,788	15,116	91.50%

2.4.4.2 Genome comparison between *M. raptorellus* and *N. vitripennis*

N. vitripennis and the congeners of *M. raptorellus*, *M. uniraptor*, and *M. zaraptor*, have a haploid karyotype of n = 5 [172-174]. A total of 25 scaffolds from my *M. raptorellus* assembly with a total length of 187.4 Mb (59.7% of the whole assembly) were unambiguously aligned to the five assembled chromosomes in *N. vitripennis* genome (Figure 11). The *N. vitripennis* chromosome assembly was based on recombination data between two closely related species (*N. vitripennis* and *N. giraulti*) [175, 176], with all non-repetitive and non-centromeric regions correctly assembled and oriented (total chromosome size 159.4 Mb, 55% of the genome). The remaining 40% of repetitive regions (Table 6) were not assembled into *N. vitripennis* chromosomes. The majority of *N. vitripennis* chromosomal regions have a collinearity relationship to *M. raptorellus* scaffolds (Table 4), suggesting high evolutionary conservation. The synteny analysis results also identified regional inversion, translocation, and duplication events, which will shed light on the genome evolution in these two genera.

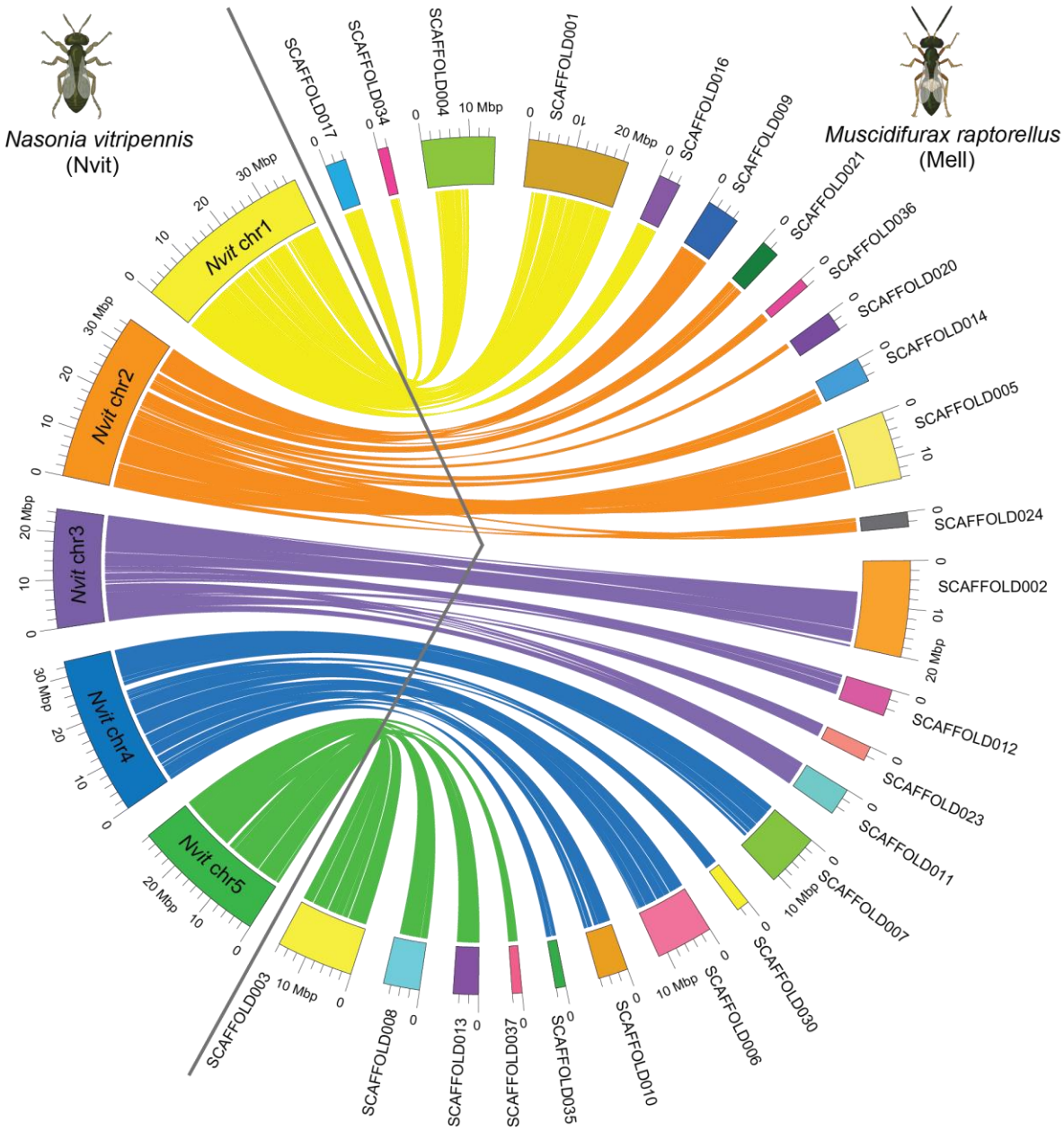


Figure 11. Genome comparisons between *Muscidifurax raptorellus* and *Nasonia vitripennis*.

A total of 25 largest scaffolds in the *M. raptorellus* assembly showed a one-to-one relationship with the five chromosomes in the *N. vitripennis* genome. Chrs 1-5 on the left of the circle represent

N. vitripennis chromosomes, and scaffolds on the right represent *M. raptorellus* assembled scaffolds. Parts of the figure were created with BioRender.com.

2.4.5 Phylogenetic relationships

2.4.5.1 Phylogenetic relationship between *N. giraulti* and eight selected arthropod genomes

I compared the *Ng* genome to 8 other sequenced arthropod genomes (fruit fly, pea aphid, honeybee, water flea, human lice, mosquito, silk moth and jewel wasp *Nv*), to identify a core gene set for phylogenomic analysis. A total of 348 single-copy 1:1 orthologs were identified. *Ng* is most closely related to *Nv*, and they cluster with honeybee, another Hymenoptera species (Figure 12). These 348 single-copy ortholog provide a useful gene set for evolutionary analysis.

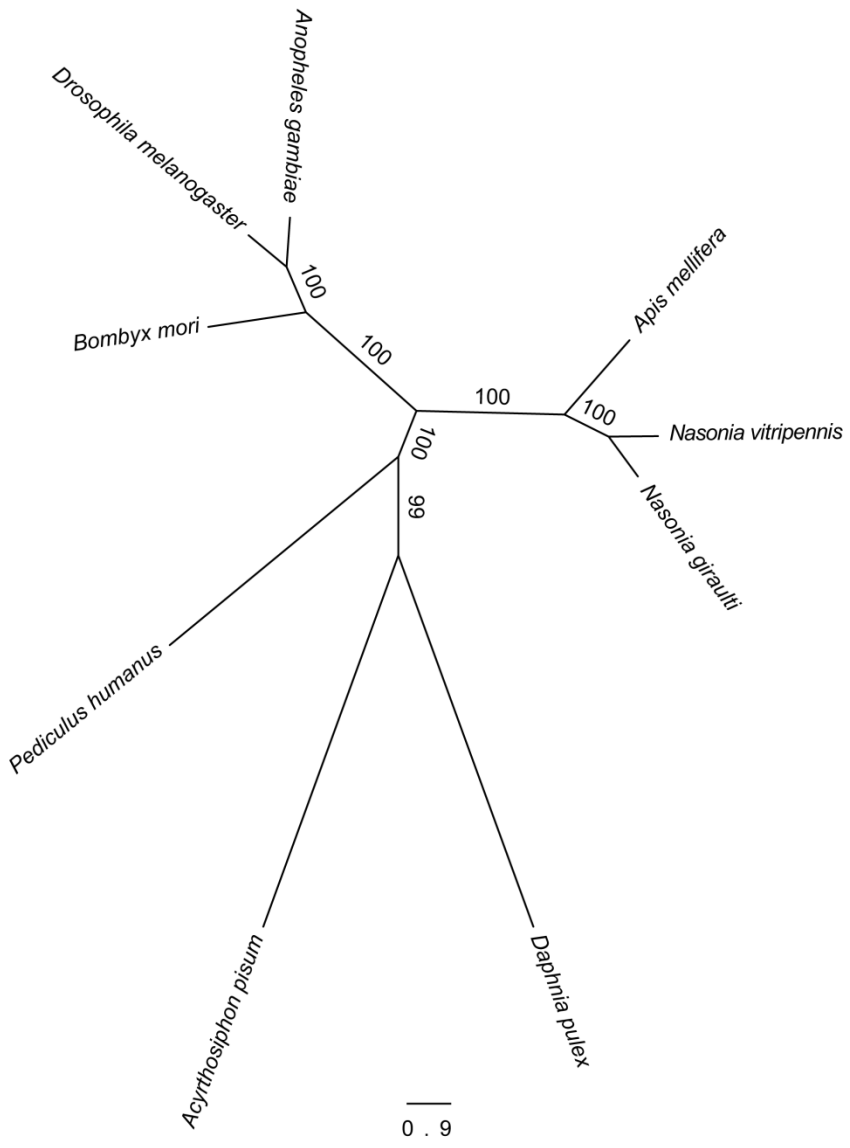


Figure 12. Phylogenetic relationships of *N. giraulti* with eight selected arthropod species.

A phylogenetic tree of *N. giraulti* with 8 other arthropod species was constructed based on a total of 348 single-copy 1:1 orthologs. The selected arthropod genomes are from fruit fly, pea aphid, honeybee, water flea, human lice, mosquito, silk moth and jewel wasp *Nasonia vitripennis*.

2.4.5.2 Phylogenetic relationship between *M. raptorellus* and nine representative hymenopteran species

To construct the phylogenetic tree of *M. raptorellus* and other hymenopteran species, I used 3,662 single-copy 1:1 orthologs in nine species (turnip sawfly, parasitic wood wasp, Braconid wasp, minute polyphagous wasp, jewel wasp, fig wasp, paper wasp, ant, and honeybee). *M. raptorellus* clustered with the chalcid wasp species within the superfamily Chalcidoidea (Figure 13). *M. raptorellus* is the closest outgroup species to the jewel wasp *Nasonia* genus that has a high-quality reference genome, which will facilitate the evolutionary studies in the *Nasonia* subgroup and parasitoid wasp comparative genomics.

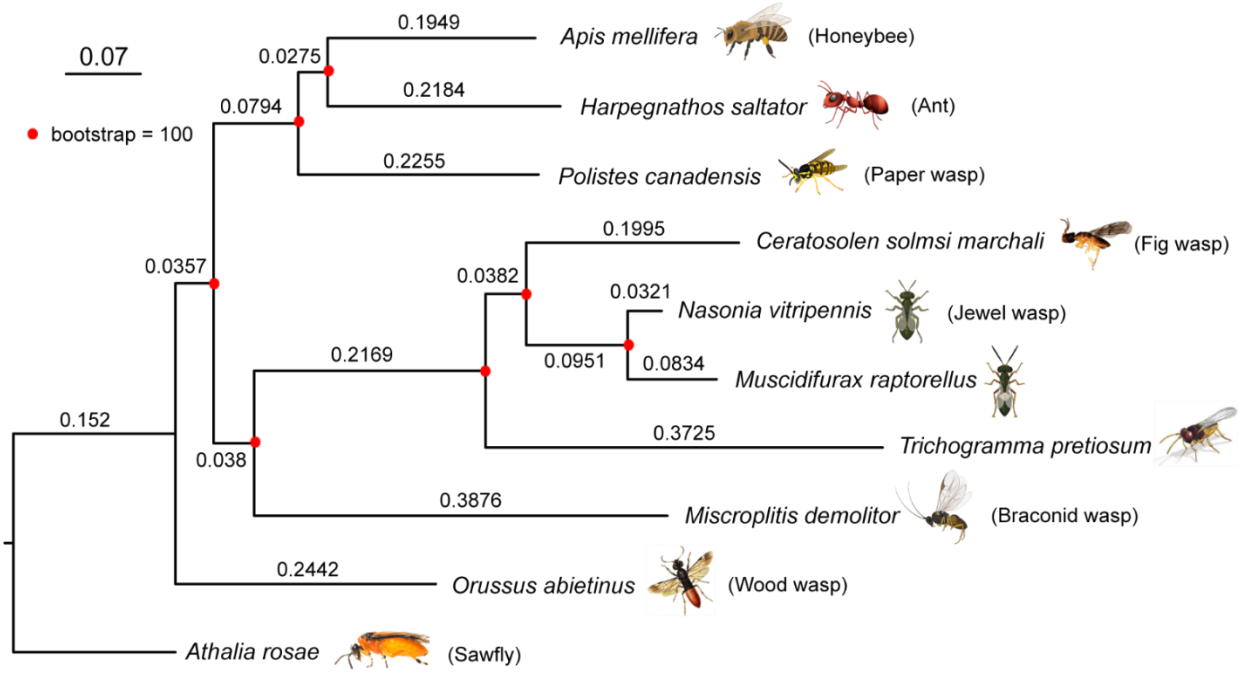


Figure 13. Phylogenetic relationship between *M. raptorellus* and nine representative hymenopteran species.

A maximum-likelihood phylogenetic tree of *M. raptorellus* with nine other hymenopteran species was constructed based on 3,662 shared 1:1 single-copy proteins using RAxML v8.2. The sawfly

Athalia rosae was used as the outgroup. The bootstrap values are supported at 100/100. The length of each branch is shown on the branches. Parts of the figure were created with BioRender.com.

2.5 Discussion

2.5.1 High-quality genome assembly of parasitoid wasp *Nasonia giraulti*

Supernova 2.0 assembler [113] was used for the *Ng* genomic assembly with a barcode subsampling strategy. The best Supernova assembly has a contig N50 of 36.14 Kb and a scaffold N50 of 400.25 Kb, which was obtained by using 20% barcode subsampling of 140 million input reads. Interestingly, using all available reads with no barcode subsampling provided the worst assembly result. This can be caused by the overkill of read coverage (~600×), which might lead to fragmented assembly due to the presence of sequencing errors.

The previous *Ng* assembly was based on 1× Sanger and 10× Illumina short-read alignments to an earlier *Nv* assembly [9]. Compared to reference-assisted *Ng* assembly, my *de novo* assembly was significantly improved in contig level with a much lower number of contigs and larger contig N50. The gap percentage is only 1.5% of the whole assembly, which surpasses most of the previous *Nasonia* genome assemblies. Although the scaffold N50 of the whole *Ng* genome is ~545 Kb, the scaffold N50 of the protein-coding gene-contained scaffolds (a total of 1,393 scaffolds) is 664.6 Kb, indicating the high quality of my current assembly in the genic regions.

2.5.2 Potential functions of *Ng*-specific genes

I further compared the *Ng* gene sets with the *Nv* annotated gene set OGS2 [177] to determine if there are any candidates for *Ng*-specific genes (methods see Figure 14). A total of 2,361 *Ng*-specific candidate genes were generated by Orthofinder [154]. The protein sequences of

these candidate genes were BLASTed to the *Nv* genome. A total of 112 *Ng* candidate genes showed no hits to the reference and *Nv* PSR1.1 genome assemblies [152]. To exclude potential pseudogenes in *Ng*, these 112 candidate genes were then aligned to the *Ng* transcripts annotated by Cufflinks [144] and 45 genes were retained. The protein sequences of these genes were aligned to *Nv* PSR1.1 again using tBLASTn and three more genes were excluded (E-value cutoff = 1E-5), resulting in the final list of 42 *Ng*-specific genes. Twenty-eight of these *Ng*-specific genes have a tBLASTn hit in *Trichomalopsis sarcophagae* (*Tsarc*), which is a sister species to the *Nasonia* genus, suggesting that they could be degenerated genes in *Nv*. Therefore, I divide this class further into 28 "*Nv* absent" genes, which are not present in the annotated *Nv* genome but are found in the closely related species *Trichomalopsis sarcophagae*, and 14 candidate "*Ng* novel" genes, which are not found in either *Nv* or *Tsarc*. Among these *Ng*-specific genes, eight genes are annotated with E-value < 1E-4 and identity >40% to the NCBI NR database. These include hypothetical protein TSAR_007225, NADH dehydrogenase (ubiquinone) flavoprotein 3, T-complex protein 1 subunit eta, gem associated protein 4, PREDICTED uncharacterized protein LOC107980813, collagen type II alpha, [histone H4]-N-methyl-L-lysine20 N-methyltransferase, and neuropeptides capa receptor-like gene (Table 9). The BLAST2GO functional analysis revealed that these 42 genes are enriched for genes involved in gluconate transmembrane transporter activity (Figure 15). The genes warrant further study to investigate their possible origins and functions.

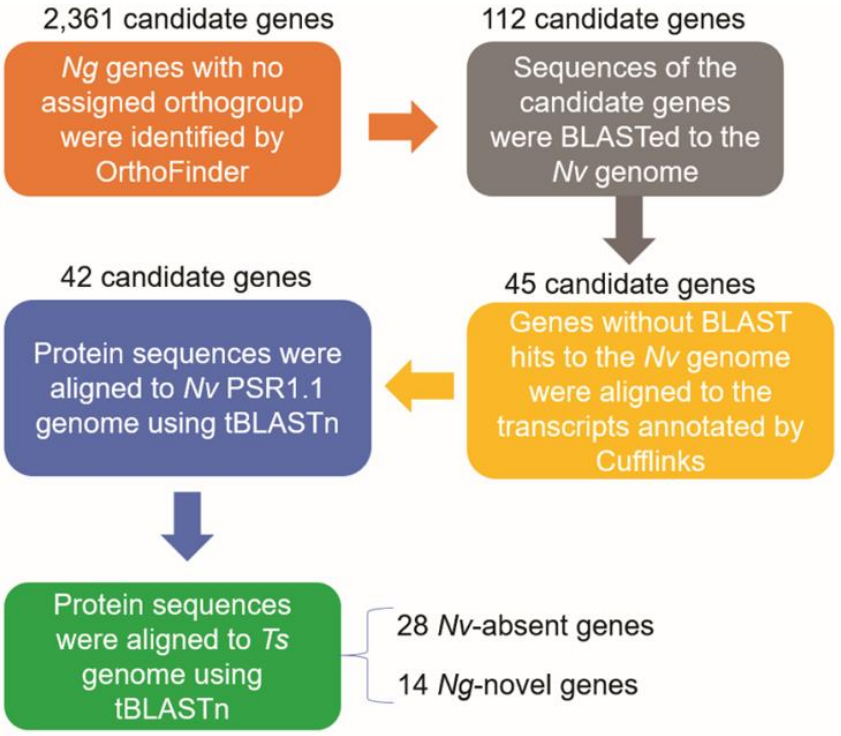


Figure 14. Workflow of bioinformatic identification *Ng*-specific genes compared to *Nv*.

Table 9. Summary of *Ng*-specific genes compared to *Nv*

***Ng*-specific gene ID**	***Ng* scaffold**	**Cufflinks count**	**Category**	**E-value**	**Annotation**	**Annotation software**
NgirV509666	SCAFFOLD76: 391662-392706	27	*Nv*-absent	2.30E-05	NADH dehydrogenase (ubiquinone) flavoprotein 3	KofamKOALA
NgirV511084	SCAFFOLD124: 18549-20323	46	*Nv*-absent	1.20E-14	hypothetical protein TSAR_007225	BLAST2GO
NgirV511319	SCAFFOLD135: 207692-208200	2	*Nv*-absent	2.60E-05	gem associated protein 4	KofamKOALA
NgirV511696	SCAFFOLD167: 222235-223057	14	*Ng*-novel	7.70E-05	collagen, type II, alpha	KofamKOALA
NgirV512673	SCAFFOLD270: 96662-97817	16	*Nv*-absent	1.10E-08	T-complex protein 1 subunit eta	KofamKOALA

NgirV512975	SCAFFOLD333: 51143-51978	8	*Ng*-novel	1.49E-09	PREDICTED: uncharacterized protein LOC107980813	BLAST2GO
NgirV514131	SCAFFOLD677: 6908-7458	5	*Nv*-absent	1.69E-58	neuropeptides capa receptor-like	BLAST2GO
NgirV514168	SCAFFOLD694: 51797-52427	4	*Nv*-absent	8.50E-05	[histone H4]-N-methyl-L-lysine20 N-methyltransferase	KofamKOALA

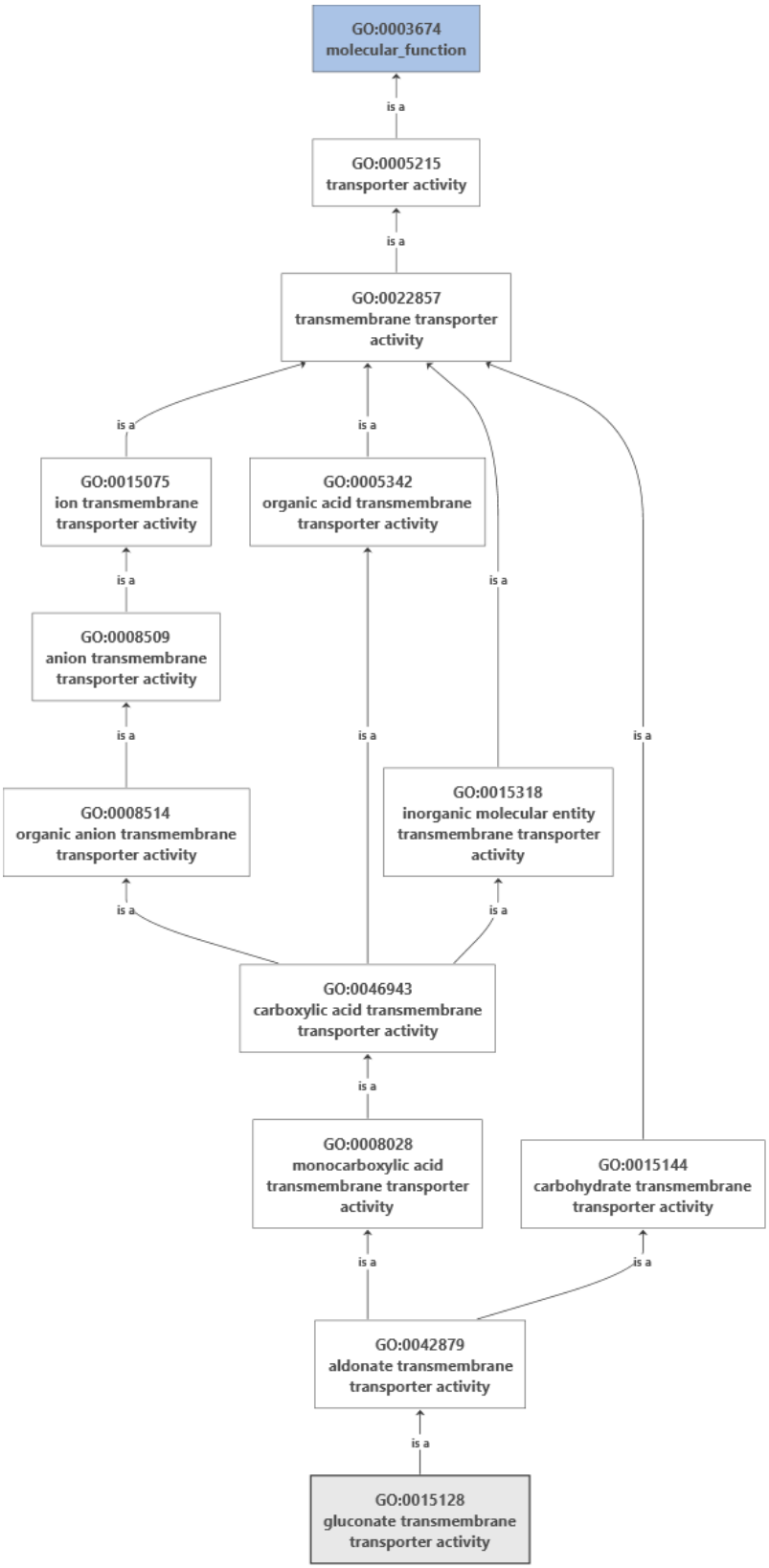

GO:0003674
molecular_function
is a
GO:0005215
transporter activity
is a
GO:0022857
transmembrane transporter activity
is a
is a
GO:0015075
ion transmembrane transporter activity
GO:0005342
organic acid transmembrane transporter activity
is a
is a
GO:0008509
anion transmembrane transporter activity
is a
is a
is a
GO:0008514
organic anion transmembrane transporter activity
GO:0015318
inorganic molecular entity transmembrane transporter activity
is a
is a
GO:0046943
carboxylic acid transmembrane transporter activity
is a
GO:0008028
monocarboxylic acid transmembrane transporter activity
GO:0015144
carbohydrate transmembrane transporter activity
is a
is a
GO:0042879
aldonate transmembrane transporter activity
is a
GO:0015128
gluconate transmembrane transporter activity

Figure 15. Functional clustering analysis of *Ng*-specific genes using Blast2GO.

2.5.3 The significance of parasitoid wasp *Nasonia* and *Muscidifurax* in genetic research

The genus *Nasonia* has been a frequently used model system for the subject of genetic, ecological, evolutionary, and developmental research in parasitoid wasps for over 70 years. The genome size of jewel wasp *Nasonia* is approximately 300 Mbp with 5 chromosomes, which is only ¼ of fruit fly *Drosophila*. In contrast to existing insect model systems, *Nasonia* have several advantages, including a haplodiploid sex determination mechanism [18], ease of handling and rearing, large family size, a complete set of DNA methyltransferases [9, 14], a short generation time, photoperiodic diapause response [2], and the availability of sequenced genomes [9]. These features collectively make it a superior organism for genetic research. *Nasonia* can be readily inbred to maintain the health of lines in the laboratory. The inbred strain is particularly suited for the study of species-specific trait inheritance. Moreover, four closely related species in the *Nasonia* genus can interbreed and produce interspecies hybrids after the removal of the endosymbiont *Wolbachia*. The interfertile *Nasonia* species allows us to investigate the chromosomal translocation and associated traits between different species. My high-quality genome assembly of *N. giraulti* will serve as a useful genomic resource for genetic and comparative genomic studies in the *Nasonia* genus.

The parasitoid wasp genus *Muscidifurax* is estimated to have diverged from the *Nasonia* genus approximately 15 million years ago [36]. Four closely related species were identified in the genus of *Muscidifurax*. The gregarious species *M. raptorellus* and the two solitary species *M. raptor* and *M. zaraptor* have received extensive studies and are commonly produced for commercial purposes. The feature that distinguishes *M. uniraptor* from its sibling species in the

Muscidifurax genus is the unique reproductive strategy known as thelytokous parthenogenesis, where each host produces only one female offspring, and no males are involved in the reproduction process. The parthenogenesis in *M. uniraptor* is triggered by an A-supergroup *Wolbachia w*Uni [32, 33]. The diverse reproductive strategies in the *Muscidifurax* genus make it an excellent model system for the study of sexual vs. asexual evolution. My assembled high-quality *M. raptorellus* reference genome holds particular significance as it can serve as the closest outgroup species to the jewel wasp *Nasonia* genus in evolutionary studies and parasitoid wasp comparative genomics. Therefore, the unique characteristics, genetic basis, and availability of genomic resources of both *Nasonia* and *Muscidifurax* genera provide a valuable model system for research in genome evolution, host preference, host-parasite coevolution, and interactions, which shed light on various aspects of parasitoid wasp biology and genetics.

2.5.4 The application of parasitoid wasp *Nasonia* and *Muscidifurax* in pest biological control

Pest flies can cause a huge economic loss on livestock operations. The impact of stable flies on feeder cattle dropped feed efficiency by 10% to 15%, and gains were reduced by 0.2 to 0.5 pounds per day [178]. When dairy cattle spend their energy getting the flies off their legs, bunching in a corner, or not eating, milk production can be reduced by almost 2.2 pounds per day. It is estimated that stable flies can cost the livestock industry $2.2 billion per year. Currently, the main pest fly control strategy is insecticide-based. However, insecticide is expensive, and resistance is a huge issue. For example, horn fly developed resistance to insecticide in a couple of years [179, 180]. The specific nucleotide substitutions present in resistance fly voltage-gated sodium channels have been identified to be associated with pyrethroid resistance [180, 181]. The potential pollution of the environment is another issue.

Parasitoid wasp species in the genus of *Nasonia*/*Muscidifurax* are excellent biological control agents for pest flies in agriculture and livestock [34, 182-184]. As natural biological control agents, the parasitoids are more environment-friendly and sustainable compared to chemical pesticides [179, 185]. *Nasonia* attacks the pupae of various flies. The fly species are primarily house flies [186], flesh flies [187-189], and blow flies [190, 191], belonging to the families *Muscidae*, *Sarcophagidae*, and *Calliphoridae*. The parasitoid wasp *M. raptorellus* is an effective biocontrol agent of dipteran filth flies, including house fly, stable fly, horn fly, black dump fly, and flesh fly [34, 182, 183]. Parasitoid wasps primarily control pest populations by disrupting the pest's life cycle. Female parasitoid wasps lay one or more eggs inside the host fly pupa. The immature parasitoid wasps feed on the developing fly pupa, disrupting its structure and preventing it from maturing into a fly, effectively breaking the host's life cycle. Mature parasitoid wasps emerge from the fly pupa, lay eggs, and continue their life cycle (Figure 1). The parasitoid wasp genus *Muscidifurax* is extremely easy to culture in large quantities with a short generation time (21 days). They do not suffer from resistance issues because the wasps co-evolve with the pest flies through an evolutionary arms race.

Chapter 3 Phylogenomic analysis of *Wolbachia* strains reveals patterns of genome evolution and recombination

3.1 Abstract

Wolbachia are widespread intracellular bacteria that mediate many important biological processes in arthropod species. In this study, I identified 210 conserved single-copy genes in 33 genome-sequenced *Wolbachia* strains in the A–F supergroups. Phylogenomic analyses with these core genes indicate that all 33 *Wolbachia* strains maintain the supergroup relationship, which was classified previously based on the multilocus sequence typing (MLST) genes. Using an interclade recombination screening method, 14 inter-supergroup recombination events were discovered in six genes (2.9%) among 210 single-copy orthologs. This finding suggests a relatively low frequency of intergroup recombination. Interestingly, they have occurred not only between A and B supergroups (nine events) but also between A and E supergroups (five events). Maintenance of such transfers suggests possible roles in *Wolbachia* infection-related functions. Comparisons of strain divergence using the five genes of the MLST system show a high correlation (Pearson correlation coefficient $r = 0.98$) between MLST and whole-genome divergences, indicating that MLST is a reliable method for identifying related strains when whole-genome data are not available. The phylogenomic analysis and the identified core gene set in this study will serve as a valuable foundation for strain identification and the investigation of recombination and genome evolution in *Wolbachia*.

3.2 Introduction

Wolbachia, alphaproteobacterial endosymbionts, are widespread and common in arthropods and filarial nematodes, either as reproductive parasites or mutualists [43-45]. More than

half of arthropods are infected with *Wolbachia* [46, 47] due to the frequent horizontal transfers of *Wolbachia* to new host species, although the typical transmission mode of these bacteria is vertical through the egg cytoplasm. Genomic studies of *Wolbachia* started with the first complete genome of the A-*Wolbachia* parasite of *Drosophila melanogaster* (*w*Mel) published in 2004 [192], and followed by the complete genome of D-*Wolbachia* (*w*Bm) in nematode *Brugia malayi* in 2005 [193]. Many more genomes have been published in the last decade, and a list of sequenced whole genomes of *Wolbachia* is summarized in Table 10.

Table 10. Summary of current sequenced *Wolbachia* genomes

Strain	Supergroup	Host species	Genome size (Mb)	Genome GenBank Accession	Reference
*w*Mel	A	*Drosophila melanogaster*	1.268	GCA_000008025.1	[192]
*w*Bm	D	*Brugia malayi*	1.080	GCA_000008385.1	[193]
*w*Pip	B	*Culex pipiens*	1.482	GCA_000073005.1	[194]
*w*Ri	A	*Drosophila simulans*	1.446	GCA_000022285.1	[195]
*w*VitB	B	*Nasonia vitripennis*	1.108	GCA_000204545.1	[196]
*w*AlbB	B	*Aedes albopictus*	1.240	GCA_000242415.3	[197]
*w*Di	B	*Diaphorina citri*	1.241	GCA_000331595.1	[198]
*w*Oo	C	*Onchocerca ochengi*	0.960	GCA_000306885.1	[199]
*w*Ha	A	*Drosophila simulans*	1.296	GCA_000376605.1	[200]
*w*No	B	*Drosophila simulans*	1.302	GCA_000376585.1	[200]
*w*Suzi	A	*Drosophila suzukii*	1.415	GCA_000333795.2	[201]
*w*Bol1	B	*Hypolimnas bolina*	1.378	GCA_000333775.1	[202]
*w*Ov	C	*Onchocerca volvulus*	0.961	GCA_000530755.1	[203]
*w*Pip_Mol	B	*Culex molestus*	1.436	GCA_000723225.2	[204]
*w*Gmm	A	*Glossina morsitans*	1.020	GCA_000689175.1	[205]

*w*Cle	F	*Cimex lectularius*	1.250	GCA_000829315.1	[206]
*w*Au	A	*Drosophila simulans*	1.268	GCA_000953315.1	[207]
Ob_Wba	B	*Operophtera brumata*	1.121	GCA_001266585.1	[208]
*w*VitA	A	*Nasonia vitripennis*	1.212	GCA_001983615.1	[209]
*w*Uni	A	*Muscidifurax uniraptor*	1.049	GCA_001983635.1	[209]
*w*Tpre	B	*Trichogramma pretiosum*	1.134	GCA_001439985.1	[210]
*w*Wb	D	*Wuchereria bancrofti*	1.061	GCA_002204235.2	[211]
*w*Fol	E	*Folsomia candida*	1.802	GCA_001931755.2	[212]
*w*Con	B	*Cylisticus convexus*	2.110	GCA_003344345.1	[213]
*w*Spc	A	*Drosophila subpulchrella*	1.420	GCA_002300525.1	[214]
*w*Stri	B	*Laodelphax striatella*	1.232	GCA_001637495.1	Unpublished
*w*DacA	A	*Dactylopius coccus*	1.171	GCA_001648025.1	[215]
*w*DacB	B	*Dactylopius coccus*	1.498	GCA_001648015.1	[215]
*w*MelPop	A	*Drosophila melanogaster*	1.239	GCA_000475015.1	[216]
*w*Npa	A	*Nomada panzeri*	1.344	GCA_001675775.1	[217]
*w*Nfe	A	*Nomada ferruginata*	1.338	GCA_001675785.1	[217]
*w*Nleu	A	*Nomada leucophthalma*	1.367	GCA_001675715.1	[217]
*w*Nfla	A	*Nomada flava*	1.333	GCA_001675695.1	[217]
*w*OneA1	A	*Nasonia oneida*	1.293	GCA_009012935.1	[218]

Because of its endosymbiotic nature, multiple different *Wolbachia* strains can be present in the same host cells, allowing the potential for homologous recombination between strains [219, 220]. Studies have observed recombination across strains and supergroups [221-223], which may be mediated by bacteriophage and lead to mosaic genomes in *Wolbachia* [195, 196, 202]. Although co-infection of different strains exists in the same arthropod host, with recombination, particularly

in associated phage [224], the supergroups may still remain genetically distinct clades [200]. Recombination events in *Wolbachia* have been discovered in *Wsp* [223] and other genes in Crustaceans [225], mites [226], and various arthropod species [221, 223, 227-229]. No inter-strain recombination has been reported in the filarial nematode *Wolbachia* strains [230].

Most of the previous research on recombination has focused on five MLST genes, *Wolbachia* surface protein (*wsp*), and 16S rRNA, or on a subset of genomes from the A-D and F supergroups [210]. Therefore, whole-genome analyses in a large number of *Wolbachia* strains of all supergroups are needed to identify additional homologous recombination events among *Wolbachia* across the different supergroups. In this study, I performed phylogenomic analyses on 33 annotated *Wolbachia* genomes, and analyzed the individual gene trees to identify potential recombination events across the supergroups. Relatively low frequencies of inter-supergroup recombination events were found, indicating a general genetic cohesiveness of supergroups. However, between supergroup recombination is still evident, and could play a role in *Wolbachia* adaptation.

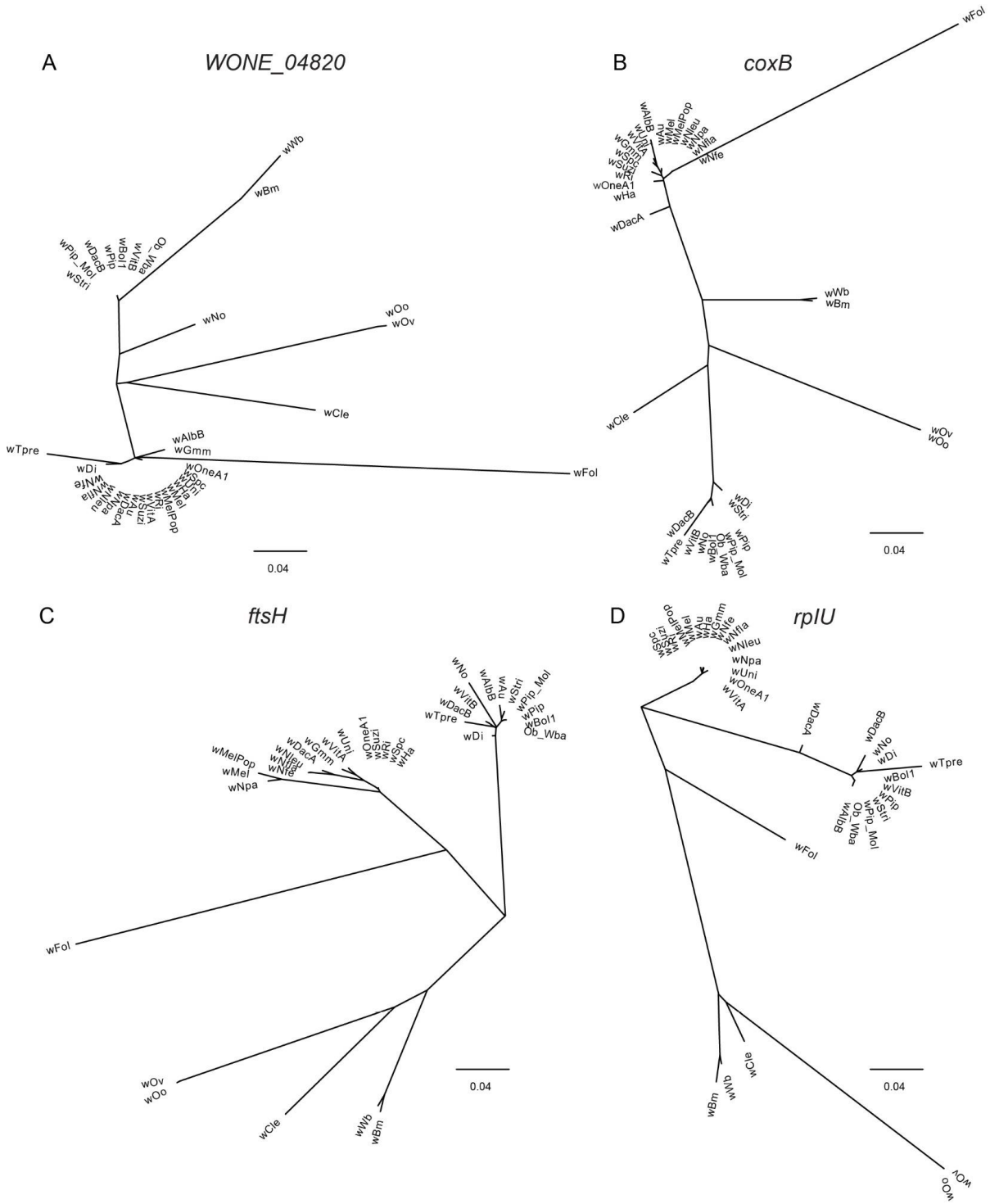
3.3 Materials and Methods

3.3.1 Phylogenomic analysis of annotated *Wolbachia* genomes

To examine the phylogeny of *Wolbachia* at the genome level, I conducted a phylogenomic analysis using 34 annotated *Wolbachia* genomes (GenBank accession numbers and reference papers listed in Table 10). Homologous genes and ortholog clusters among all 34 *Wolbachia* genomes were determined by using OrthoFinder v1.1.8 [165] with default settings. A total of 210 single-copy ortholog groups were identified, and gene IDs in each of the ortholog groups were

used to extract the corresponding nucleotide and protein sequences from 34 *Wolbachia* genomes. PAL2NAL [231] was used to check the consistency between the nucleotide and protein sequences, and all inconsistent nucleotide sequences downloaded from GenBank were manually corrected. These 210 core single-copy genes were used for the subsequent analysis. The 210 core single-copy genes in all 34 *Wolbachia* genomes were aligned with MAFFT [166] at both nucleotide and protein sequence levels. These single-gene alignments were concatenated into one alignment to use in the subsequent phylogenetic analysis. A Maximum Likelihood (ML) tree was constructed with the GTRGAMMA model and 1,000 bootstrap replicates by RAxMLv8.2 [168] using the concatenated nucleotide sequence alignment of the core gene set. For phylogenetic analysis of protein sequences from the core gene set, the best-fit model of protein evolution was searched by ProtTest 3 [167]. The final ML phylogenetic tree was inferred by using RAxML v8.2 [168] with the FLU protein model (best-fit model identified by ProtTest 3) and 1,000 rapid bootstrap replicates.

The single gene ML trees for all 210 core genes were constructed with their corresponding nucleotide sequence alignments using the GTRGAMMA model and 1,000 bootstrap replicates by RAxML v8.2 [168]. I also constructed protein trees for these identified genes with their corresponding protein sequence alignments using the best-fit protein model detected by ProtTest 3 [167] and 1,000 rapid bootstrap replicates by RAxML v8.2 [168]. The gene trees and protein trees were visualized using FigTree v1.4.4 [232]. For better viewing of short branches, transformation and rerooting were performed in FigTree to generate the main figures. The original gene trees are shown in Figure 16.


A
WONE_04820
wWb
wBm
Ob_Wba
wVitB
wBol1
wPip
wDacB
wPip_Mol
wStri
wNo
wOo
wOv
wCle
wTpre
wAlbB
wGmm
wFol
wDi
wOneA1
wMelPop
wMel
wNpa
wNfla
wNleu
wDacA
wUni
wVitA
wSuzi
wRi
wSpc
wHa
0.04
B
coxB
wFol
wAlbB
wMel
wMelPop
wNpa
wNleu
wNfla
wNfe
wOneA1
wHa
wDacA
wWb
wBm
wCle
wOv
wOo
wDacB
wTpre
wVitB
wNo
wPip
wPip_Mol
wStri
wDi
0.04
C
ftsH
wNo
wAlbB
wAu
wStri
wPip_Mol
wVitB
wDacB
wTpre
wPip
wBol1
Ob_Wba
wDi
wMelPop
wMel
wNpa
wNfla
wNleu
wDacA
wGmm
wVitA
wUni
wOneA1
wSuzi
wRi
wSpc
wHa
wFol
wOv
wOo
wCle
wWb
wBm
0.04
D
rplU
wMelPop
wMel
wNpa
wNfla
wNleu
wGmm
wHa
wNfe
wOneA1
wUni
wNpa
wNleu
wVitA
wDacA
wDacB
wNo
wDi
wTpre
wBol1
wVitB
wPip
wStri
wPip_Mol
Ob_Wba
wAlbB
wFol
wBm
wWb
wCle
wOv
wOo
0.04

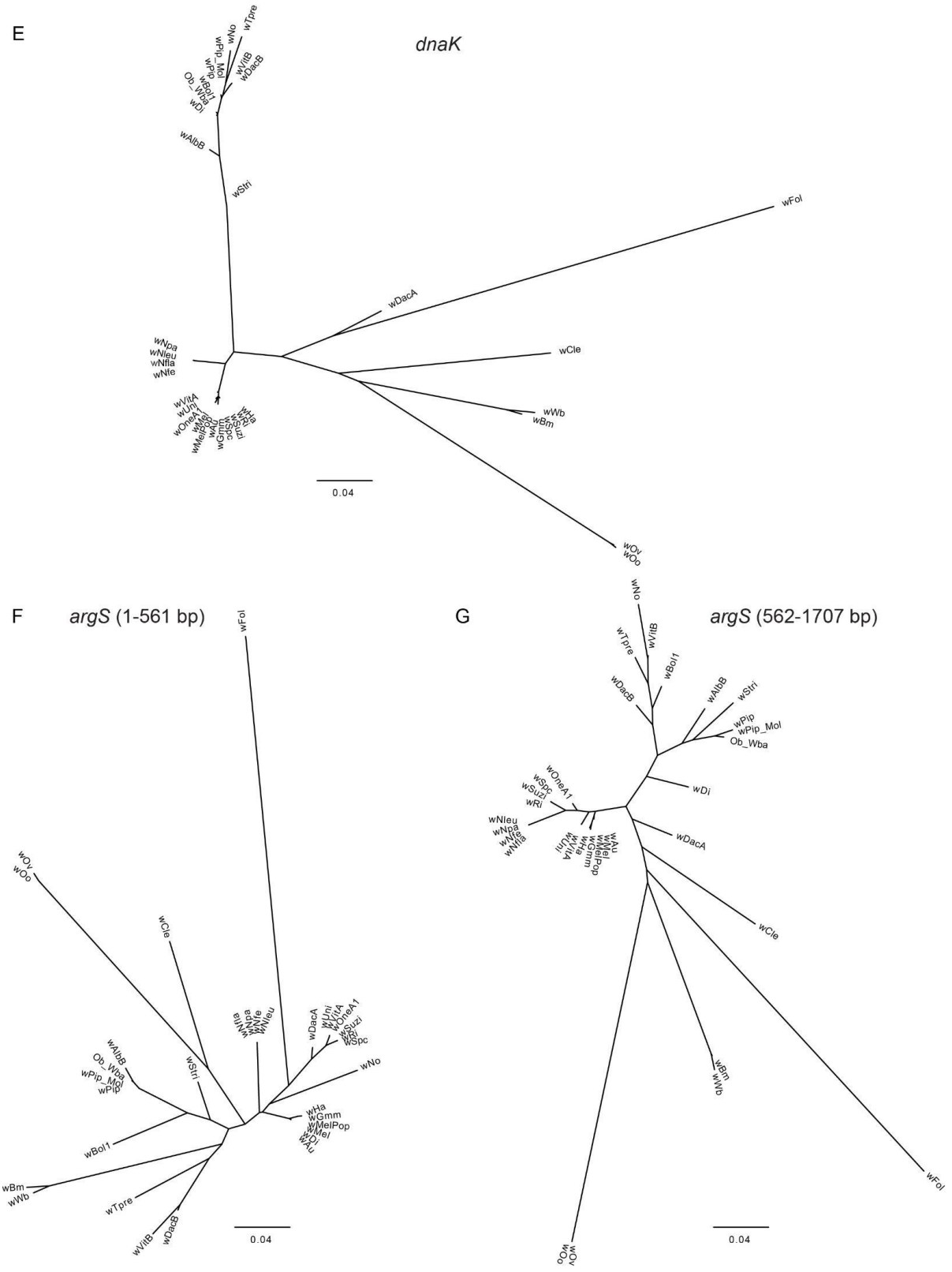

E
dnaK
wTpre
wNo
wPip
wMol
wBol1
wVitB
wDacB
wOb_Wba
wDi
wAlbB
wStri
wFol
wDacA
wNpa
wNleu
wNfla
wNfe
wCle
wVitA
wUni
wOneA1
wMelPop
wAu
wGmm
wSpc
wSuzi
wRi
wWb
wBm
0.04
wOv
wOo
F
argS (1-561 bp)
wFol
wOv
wOo
wCle
wAlbB
wOb_Wba
wPip_Mol
wPip
wStri
wNpa
wNfe
wNfla
wNleu
wDacA
wUni
wVitA
wOneA1
wSuzi
wRi
wSpc
wNo
wHa
wGmm
wMelPop
wMel
wDi
wAu
wBol1
wBm
wWb
wTpre
wVitB
wDacB
0.04
G
argS (562-1707 bp)
wNo
wVitB
wTpre
wBol1
wDacB
wAlbB
wStri
wPip
wPip_Mol
Ob_Wba
wDi
wOneA1
wSpc
wSuzi
wRi
wNleu
wNpa
wNfe
wNfla
wUni
wVitA
wAu
wMel
wMelPop
wGmm
wDacA
wCle
wBm
wWb
wFol
wOv
wOo
0.04

Figure 16. Nucleotide ML trees for six *Wolbachia* genes with interclade recombination events.

(A) *WONE_04820*. (B) *coxB*. (C) *ftsH*. (D) *rplU*. (E) *dnaK*. (F) *argS* (1-561 bp). (G) *argS* (562-1707 bp).

3.3.2 Identification of individual gene trees with intergroup recombination events

To search for interclade recombination events, I developed a prescreening tool for the identification of specific gene/protein recombinants that move a particular gene/protein outside its respective supergroup. Based on the concatenated strain phylogeny, I assign a supergroup identity for each strain. For every gene, I calculate the branch length between all strain combinations, and then determine the nearest neighbor based on the shortest branch length. Candidate recombination vents are then identified as those for which the nearest neighbor is in a different supergroup. Because some supergroups only have a single representative, the method is most effective at finding candidate recombination to or from A, B, C, and D supergroups. Next, an Interclade Recombination (IR) score was used to quantify the degree of divergence of the gene from its strains' supergroup. The distance from the candidate gene to its nearest neighbor (Nn) is compared to the average interclade distance (IC) distance of the recombination candidate gene to other members of its strain's supergroup (based on the concatenated phylogeny) using the IR metric below.

$$\textbf{IR score} = \left(1 - \frac{\boldsymbol{Nn}}{\boldsymbol{IC}}\right) \times \mathbf{100}$$

An IR score can range from 0 to 100, with a larger score indicating a recombination between supergroups.

I further manually compared gene trees in both nucleotide and protein level as follows: 1) the nucleotide ML trees were compared to the concatenated ML tree to manually confirm the recombination events; 2) nucleotide sequence alignments were further inspected for informative SNPs that separate different supergroups; 3) the single-gene protein trees were also compared to the concatenated protein tree to check for consistency of supergroup classification. Inference of intragenic recombination events and the breakpoints was conducted on nucleotide sequence alignments using the GARD algorithm [233] with default parameters using the datamonkey web server (http://www.datamonkey.org/).

3.3.3 Phylogenetic analysis of *Wolbachia* in *Nasonia* using MLST genes.

The five MLST (Multi Locus Sequence Typing) genes [59, 234] were examined to further characterize the phylogenetic relationships of *Wolbachia* strains in *Nasonia*. These genes include *gatB* (aspartyl/glutamyl-tRNA (Gln) amidotransferase, subunit B), *coxA* (cytochrome c oxidase subunit I), *hcpA* (conserved hypothetical protein), *ftsZ* (cell division protein) and *fbpA* (fructose-bisphosphate aldolase). The pairwise evolutionary divergence distances between 33 *Wolbachia* species were estimated with both the core gene set identified in this study, five MLST genes and the concatenated sequence of these five MLST genes in 33 *Wolbachia* species by using the Maximum Composite Likelihood model [235] in MEGA7 [236]. Estimates of evolutionary divergence using the *ftsZ* gene were only conducted among 31 *Wolbachia* species, excluding *w*Bm and *w*Wb, because of the inability to correctly annotate *ftsZ* in these species. The Pearson correlation coefficient of estimated evolutionary divergences with the core gene set and the MLST gene set (each MLST gene and the concatenated sequence of five MLST genes) was calculated with the Hmisc package [237] in R.

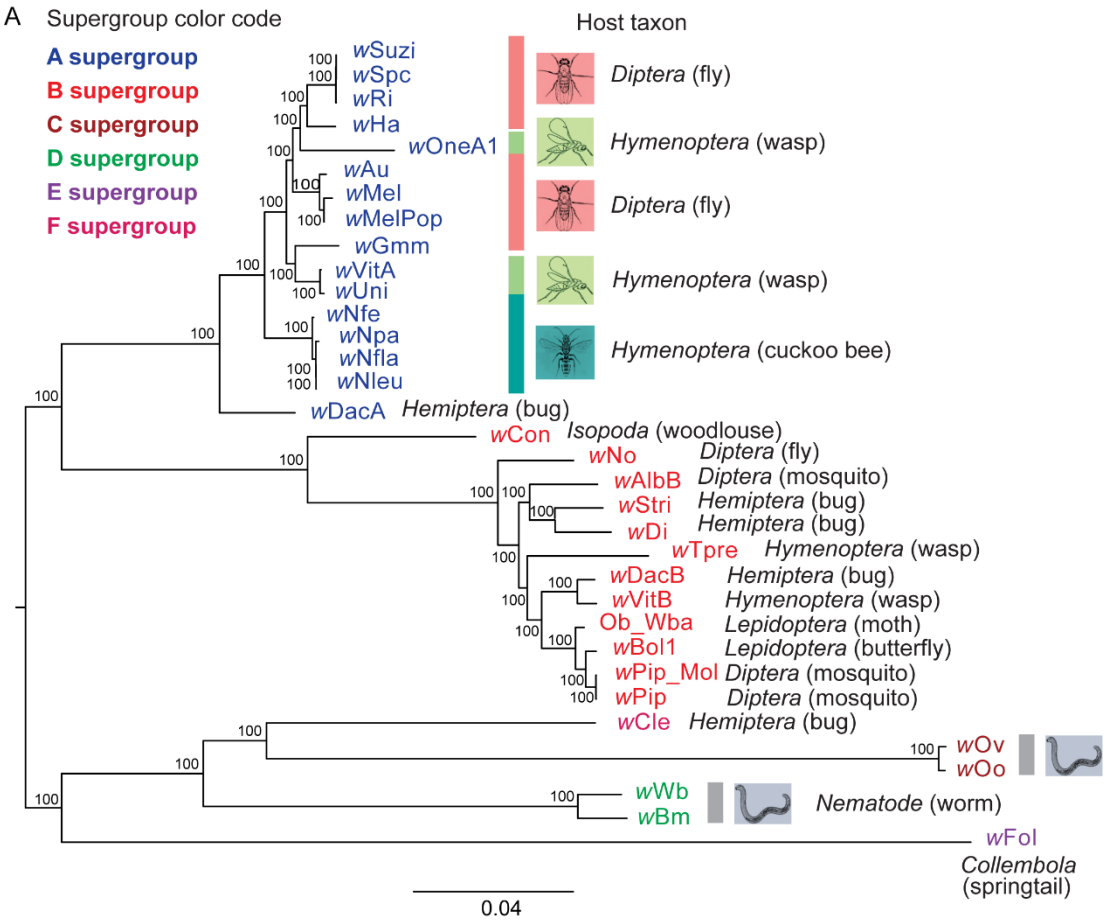
3.4 Results

3.4.1 Phylogenomic analysis of annotated *Wolbachia* genomes

To identify a core gene set for phylogenomic analysis of *Wolbachia* strains, I initially compared 34 publicly available and annotated *Wolbachia* genomes as of November 2019, which include sixteen A-group, twelve B-group, two C-group, two D-group, and one for E and F-group strains from diverse host species (Table 10). Single-gene ortholog clusters were generated using the procedure described in the Methods. A total of 210 single-gene ortholog clusters were identified that are shared among the 34 *Wolbachia* genomes. This is a smaller set than the 496 *Wolbachia* gene orthologs detected in [210] for 16 *Wolbachia* strains, but this study included a larger strain set (34 *Wolbachia* strains), and I restricted the analysis to single-copy orthologs across all of the genomes.

Based on the concatenated coding nucleotide and protein sequences of this core gene set, Maximum Likelihood (ML) phylogenetic trees of 34 *Wolbachia* genomes confirmed the separation of different supergroups A (*w*Suzi, *w*Spc, *w*Ri, *w*Ha, *w*Au, *w*Mel, *w*MelPop, *w*Gmm, *w*Uni, *w*DacA, *w*Nfe, *w*Npa, *w*Nfla, *w*Nleu, *w*VitA, *w*OneA1), B (*w*AlbB, *w*Stri, *w*Di, *w*No, *w*Tpre, *w*DacB, *w*VitB, Ob_Wba, *w*Bol1, *w*Pip_Mol, *w*Pip), C (*w*Oo, *w*Ov), D (*w*Bm, *w*Wb), E (*w*Fol) and F (*w*Cle) with 100% bootstrap support (Figure 17). One of the B-*Wolbachia*, *w*Con, appeared to be phylogenetically distant from other B strains (Figure 17). However, its genome size is 2.11Mb, almost double the B-*Wolbachia* average (1.288 Mb). Further examination of the genome assembly suggested *w*Con is potentially a mixed assembly of one A and one B *Wolbachia* genomes. Therefore, I excluded *w*Con and reconstructed the ML tree using the rest 33 *Wolbachia* nucleotide sequences (Figure 18). For comparisons of nucleotide and protein phylogenies, I also

constructed an ML phylogenetic tree of concatenated protein sequences from these core genes using RAxML [168]. The protein ML phylogenetic tree (Figure 19) matched well with the nucleotide coding sequence ML tree (Figure 18).



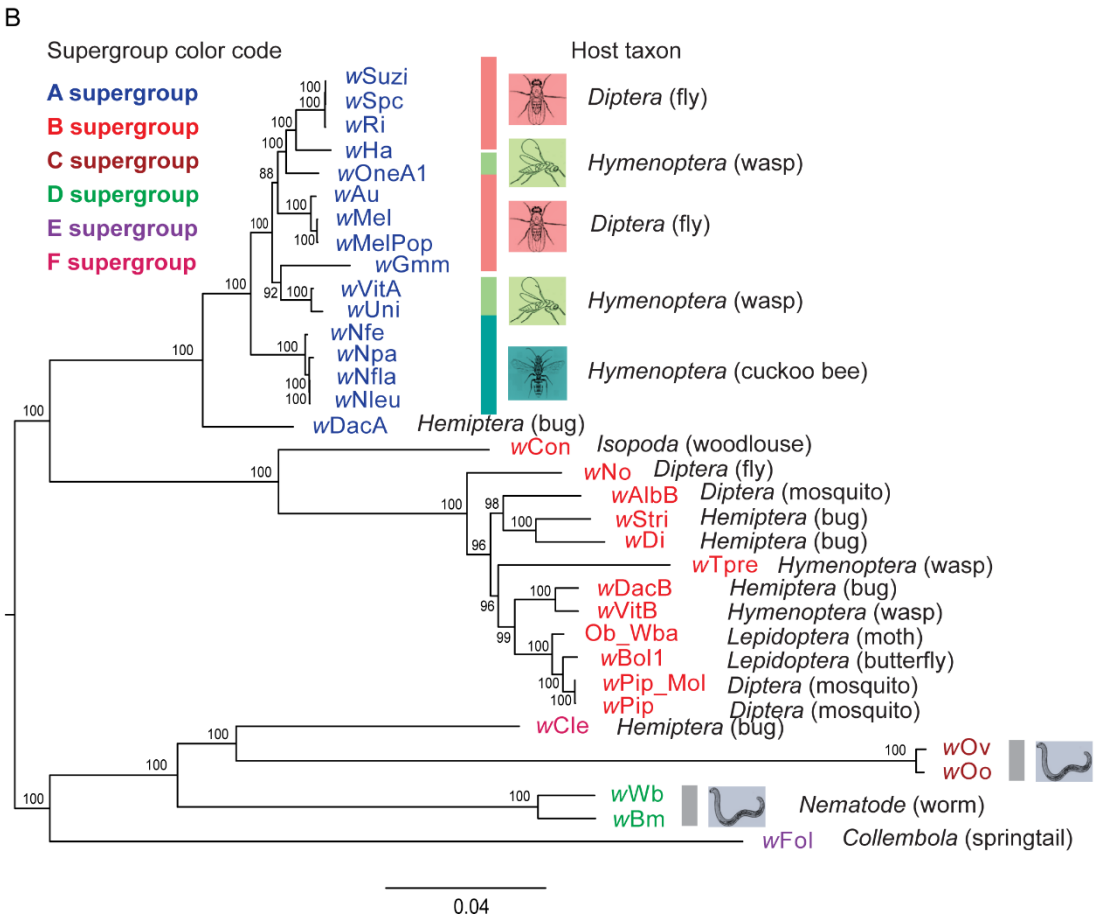


Figure 17. Phylogenetic analysis of 34 *Wolbachia* genomes.

(A) Phylogenetic ML tree was constructed using concatenated nucleotide sequence alignment of 210 single-copy orthologous genes; (B) Phylogenetic FLU tree was constructed using concatenated protein sequence alignment of 210 single-copy orthologous genes.

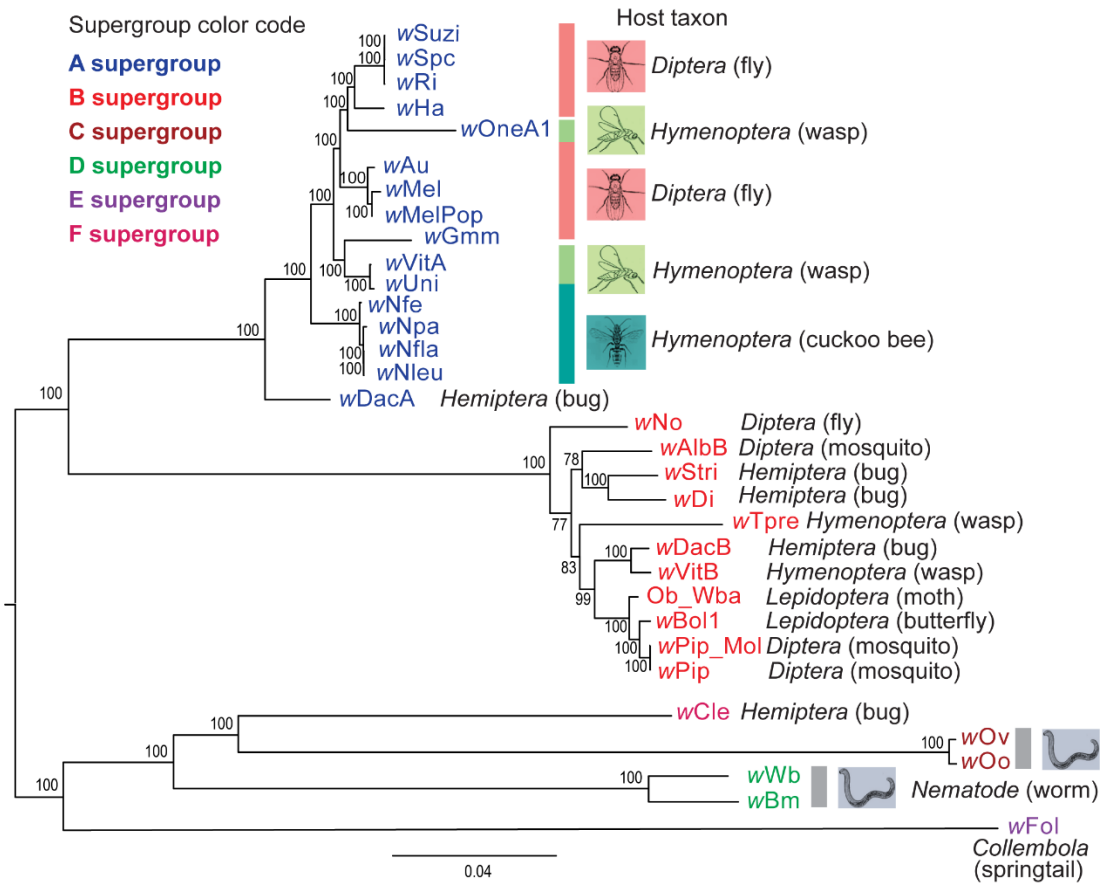


Figure 18. Phylogenomic relationships of 33 *Wolbachia* strains.

The phylogenetic tree was constructed using Maximum Likelihood method from a concatenated nucleotide sequence alignment of 210 single-copy orthologous genes among 33 genome-sequenced *Wolbachia* strains. Numbers on the branches represent the support from 1,000 bootstrap replicates. Branch transformation and rerooting were performed in FigTree 1.4.4. The assembly names were color-coded based on supergroup identity (A-F). Host taxonomic classifications and species common names were labeled.

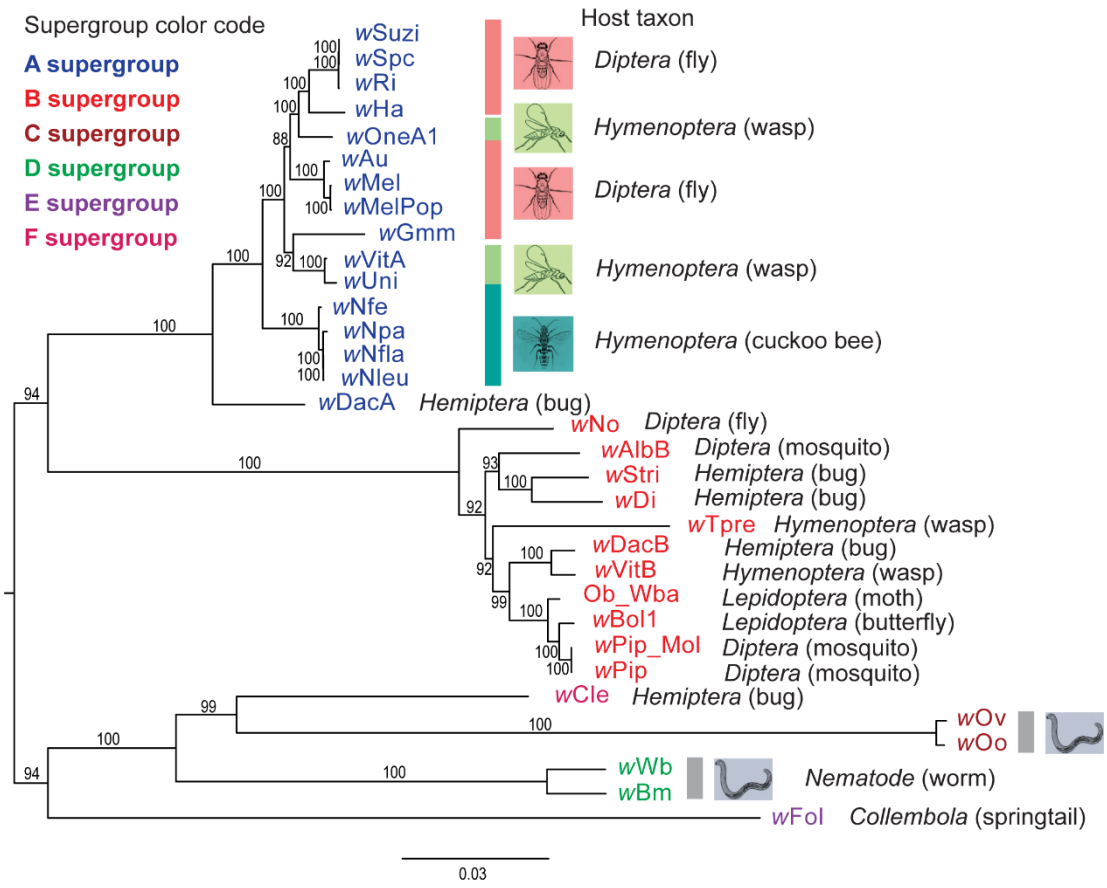


Figure 19. Phylogenetic analysis of 33 *Wolbachia* genomes (*w*Con excluded) from concatenated protein sequence alignment of 210 single-copy orthologous genes.

As expected, my genomic analyses support extensive horizontal movement of *Wolbachia* strains between divergent host species. For example, *w*OneA1, which is an A-supergroup bacterium in the parasitoid *Nasonia oneida* [238] is more closely related to a subset of A-*Wolbachia* found in *Drosophila* (*w*Ha, *w*Ri, *w*Spc, and *w*Suzi) than to *w*VitA and *w*Uni in closely related parasitoid wasps. This pattern was previously observed using MLST genes in *Wolbachia* [57], but is now supported by a much larger data set. The B-supergroup mosquito *Wolbachia w*Alb in *Aedes albopictus* gives another example of an obvious major host shift (Figure 18).

3.4.2 Identification of inter-supergroup recombination events

Our focus in this study is to evaluate recombination between supergroups in *Wolbachia*. I therefore developed a prescreening method to detect candidates between supergroup recombination events, and applied it to the 210 single-copy ortholog gene set. For each *Wolbachia* strain on an individual gene nucleotide tree, I computed the branch length distance to all other strains. I defined recombination candidates if their nearest neighboring strain belongs to a different supergroup based on the concatenated gene tree (see Methods). The interclade recombination score (IR score) can range from 0 to 100. This method works for detecting recombinants within supergroups containing more than one genome-sequenced strain (see Methods). Five genes with IR >65 were chosen as the cutoff for further investigation. I identified recombination events between A and B, and A and E supergroups. I also examined all 210 RAxML gene trees with both the corresponding protein and nucleotide sequence alignments. Both tree topologies and bootstrap values support the recombination events detected by the screening method. One additional recombination event was found for an A-group strain that contains an E- group version of the *dnaK* gene (Table 11).

Table 11. List of six *Wolbachia* genes with interclade recombination events

Gene Name	Gene Description	Intragenic Recombination Breakpoint	Species with Interclade Recombination	Interclade Recombination Score	Gene Name
ftsH	ATP-dependent metalloprotease FtsH	816 bp (P = 0.0002)	*w*Au (B-in-A)	99.9	Figure 20A
rplU	50S ribosomal protein L21	None	*w*DacA (B-in-A)	71.0	Figure 20B

coxB	Cytochrome c oxidase subunit II	None	*w*AlbB (A-in-B)	92.3	Figure 21A
			*w*Fol (A-in-E)	NA	
WONE_04820	Hypothetical protein	None	*w*Di (A-in-B)	91.3	Figure 21B
			*w*AlbB (A-in-B)	82.6	
			*w*Tpre (A-in-B)	67.1	
			*w*Fol (A-in-E)	NA	
argS	Arginine-tRNA ligase	1-561 bp	*w*Di (A-in-B)	99.9	Figure 22A
		(P = 0.0006)	*w*No (A-in-B)	43.9	
			*w*Fol (A-in-E)	NA	
		562-1,707 bp	*w*Di (A-in-B)	23.3	Figure 22B
			*w*DacA (E-in-A)	NA	
dnaK	Chaperone protein DnaK	None	*w*DacA (E-in-A)	NA	Figure 23

A total of 5 genes (2.4%) with 9 recombination events were identified between A and B supergroups, including B-supergroup genes *FtsH* (ATP-dependent metalloprotease) and *rplU* (50S ribosomal protein L21) in A-supergroup strains *w*Au (Figure 20A) and *w*DacA (Figure 20B) respectively (B-in-A events in Table 11), and 7 A-in-B recombination events in *coxB* (cytochrome c oxidase subunit II), *WONE_04820* (hypothetical protein) and *argS* (arginine-tRNA ligase) (Table 11 and Table 12). GARD algorithm [233] was used to detect intragenic recombination in these

events and identify recombination breakpoints if intragenic recombination is involved. Two breakpoint positions among the identified genes were detected by GARD, including one breakpoint position at 816 bp in *ftsH* gene with a *P*-value of 0.0002, another breakpoint position at 561 bp in *argS* gene with a *P*-value of 0.0006 (Table 11).

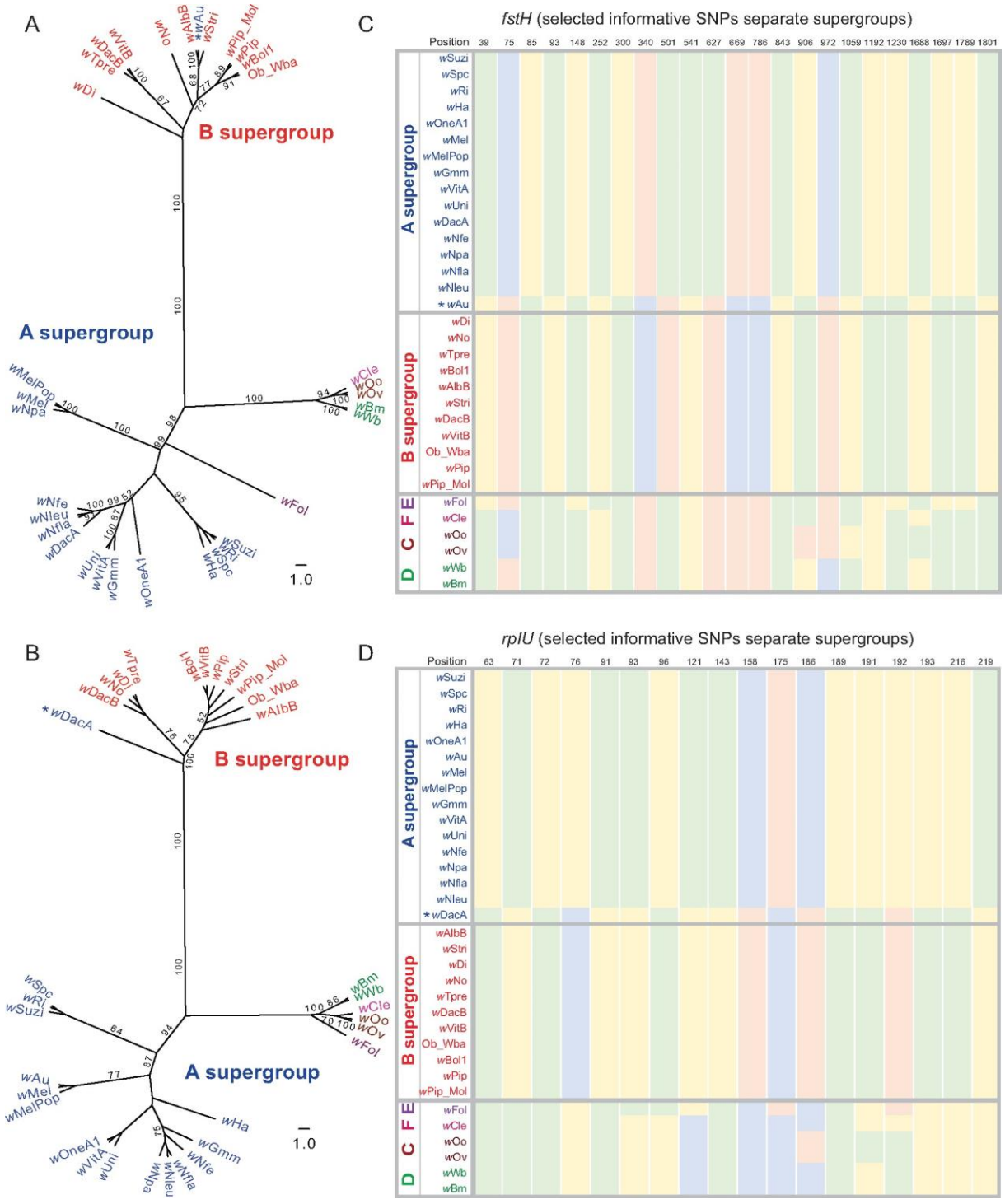

A
B supergroup
A supergroup
1.0
C
fstH (selected informative SNPs separate supergroups)
Position 39 75 85 93 148 252 300 340 501 541 627 669 786 843 906 972 1059 1192 1230 1688 1697 1789 1801
A supergroup
wSuzi
wSpc
wRi
wHa
wOneA1
wMel
wMelPop
wGmm
wVitA
wUni
wDacA
wNfe
wNpa
wNfla
wNleu
* wAu
B supergroup
wDi
wNo
wTpre
wBol1
wAlbB
wStri
wDacB
wVitB
Ob_Wba
wPip
wPip_Mol
D C F E
wFol
wCle
wOo
wOv
wWb
wBm
B
D
rplU (selected informative SNPs separate supergroups)
Position 63 71 72 76 91 93 96 121 143 158 175 186 189 191 192 193 216 219

Figure 20. Inter-supergroup recombination events of B supergroup genes *fstH* and *rplU* in A-*Wolbachia* strains.

(A, B) Nucleotide ML trees reveal IR events, in which genes from an A-*Wolbachia* clusters with B supergroup. The supergroup identities are labeled using the same color code as in Figure 18. Bootstrap values above 50 are shown in the figure. (C, D) Supergroup informative SNP positions are plotted for all strains (green: A; blue: C; yellow: G; and pink: T). These SNPs showed the general pattern of recombination, whether entire genes between clades or between clade recombination within genes.

Table 12. Interclade recombination events detected in 33 *Wolbachia* genomes between A and B supergroups

Ortho group	Strain 1	Super-group 1	Strain 2	Super-group 2	Recombination event	Distance	Average distance	IR score
OG0000228	*w*Au	A	*w*AlbB	B	B->A	2E-06	0.214	99.9
OG0000384	*w*DacA	A	Ob_Wba	B	B->A	0.040	0.138	71.0
OG0000160	*w*AlbB	B	*w*VitA	A	A->B	0.021	0.274	92.3
OG0000129	*w*Di	B	*w*Npa	A	A->B	0.009	0.107	91.3
OG0000129	*w*AlbB	B	*w*Spc	A	A->B	0.019	0.109	82.6
OG0000129	*w*Tpre	B	*w*Npa	A	A->B	0.046	0.140	67.1

OG0000301 (1-561bp)	*w*Di	B	*w*Mel	A	A->B	<0.001	0.154	99.9
OG0000301 (1-561bp)	*w*No	B	*w*Au	A	A->B	0.121	0.215	43.9
OG0000301 (562-1707bp)	*w*Di	B	*w*MelPop	A	A->B	0.091	0.119	23.3

Here I describe the recombination events in more detail. There are two cases of A-*Wolbachia* strains that contain a B-*Wolbachia* gene transfer. For *ftsH*, the A-*Wolbachia w*Au strain gene clusters with B-*Wolbachia* strains with an IR score of 99.9, and this recombination event is supported in the nucleotide tree with a bootstrap value of 100 (Figure 20A). As a universally conserved gene in bacteria, *ftsH* is known to be crucial for the proteolytic degradation of specific integral membrane proteins and cytoplasmic proteins, and it also targets soluble signaling factors like heat-shock sigma factor σ32 and transcriptional activator λ-CII [239]. A second B into A recombination event involves a B-group *rplU* gene that has inserted into the A-*Wolbachia w*DacA (IR = 71), which is also supported with a bootstrap value of 100 in the corresponding nucleotide tree (Figure 20B). Less is known about the function of *rplU*, except for its interaction with 23S rRNA [240].

Three additional genes reveal recombination events of individual A-*Wolbachia* genes into B-*Wolbachia* strains. The *coxB* gene from an A-*Wolbachia* was transferred to B-*Wolbachia w*AlbB (IR = 92), supported by the corresponding nucleotide and trees with a bootstrap value of 99 (Figure 21A). The *coxB* protein is a component of the electron transport chain which drives oxidative phosphorylation. The second case of an A-to-B transfer involves the hypothetical protein

WONE_04820 gene. An A-*Wolbachia* gene is present in three B-*Wolbachia* strains *w*Di, *w*AlbB and *w*Tpre (IR = 91, 83 and 67, respectively). The corresponding nucleotide tree supports the general pattern with a bootstrap value of 74 (Figure 21B). Based on the concatenated tree topology, it is difficult to resolve whether these indicate a single or independent transfer event, given that the three strains are not monophyletic within the B supergroup (Figure 18). The function of this gene is currently unknown.

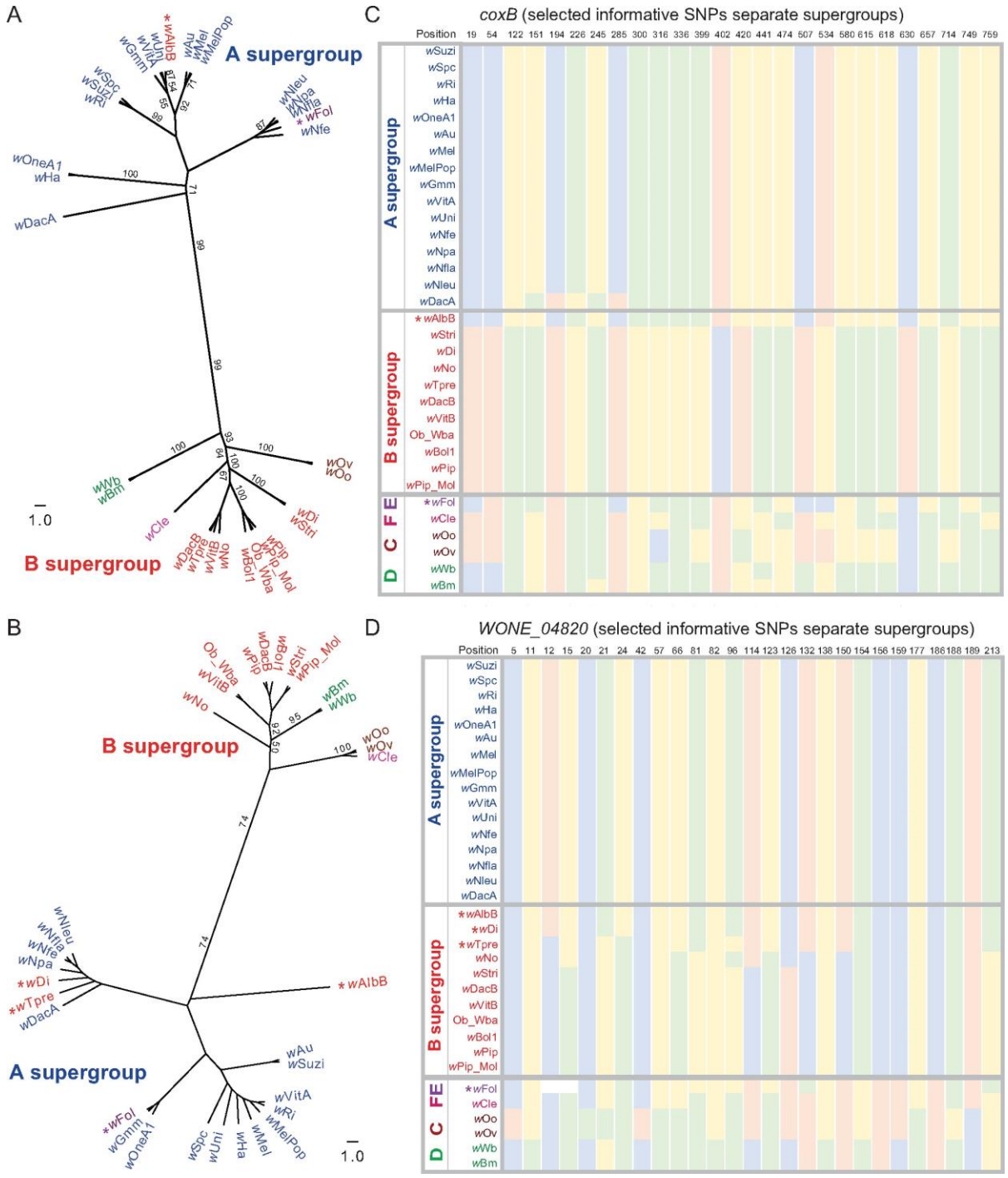


Figure 21. Inter-supergroup recombination events of A supergroup genes *coxB* and *WONE_04820* in B-*Wolbachia* and E-*Wolbachia* strains.

(A) Nucleotide ML trees reveal interclade recombination events, in which *coxB* genes from *w*AlbB (B-*Wolbachia*) and *w*Fol (E-*Wolbachia*) cluster with A supergroup, with a bootstrap support of 99. (B) Nucleotide ML trees reveal interclade recombination events, in which hypothetical protein *WONE_04820* from *w*AlbB, *w*Di, *w*Tpre, (B-*Wolbachia*) and *w*Fol (E-*Wolbachia*) clusters with A supergroup, with a bootstrap support of 74. (C, D) Supergroup informative SNP positions are plotted for all strains (green: A; blue: C; yellow: G; pink: T). Bootstrap values above 50 are shown in the figure.

In each case for the above examples, the complete gene was recombined into a different supergroup. However, recombination events can also occur within genes, as has been documented for the highly recombinogenic *wsp* gene [221]. For *argS*, I found evidence for intragenic recombination (Figure 22A and Figure 22B), with significantly different topologies between the 5' region (positions 1-561 bp) compared to the rest of the gene (positions 562-1707 bp). Intragenic recombination is supported by GARD, which identified the breakpoint at 561 (*P*-value = 0.0006). As a member of the class I aminoacyl-tRNA synthetase family, the expression of *argS* is reported to increase the aminoacyl-tRNA synthetase activity in bacteria [241]. In addition, there is also an apparent A-B recombinant event in the *coxB* gene of *w*DacA based on a stretch of 5 A-B diagnostic SNPs (position 151, 194, 226, 245 and 285 in Figure 21C).

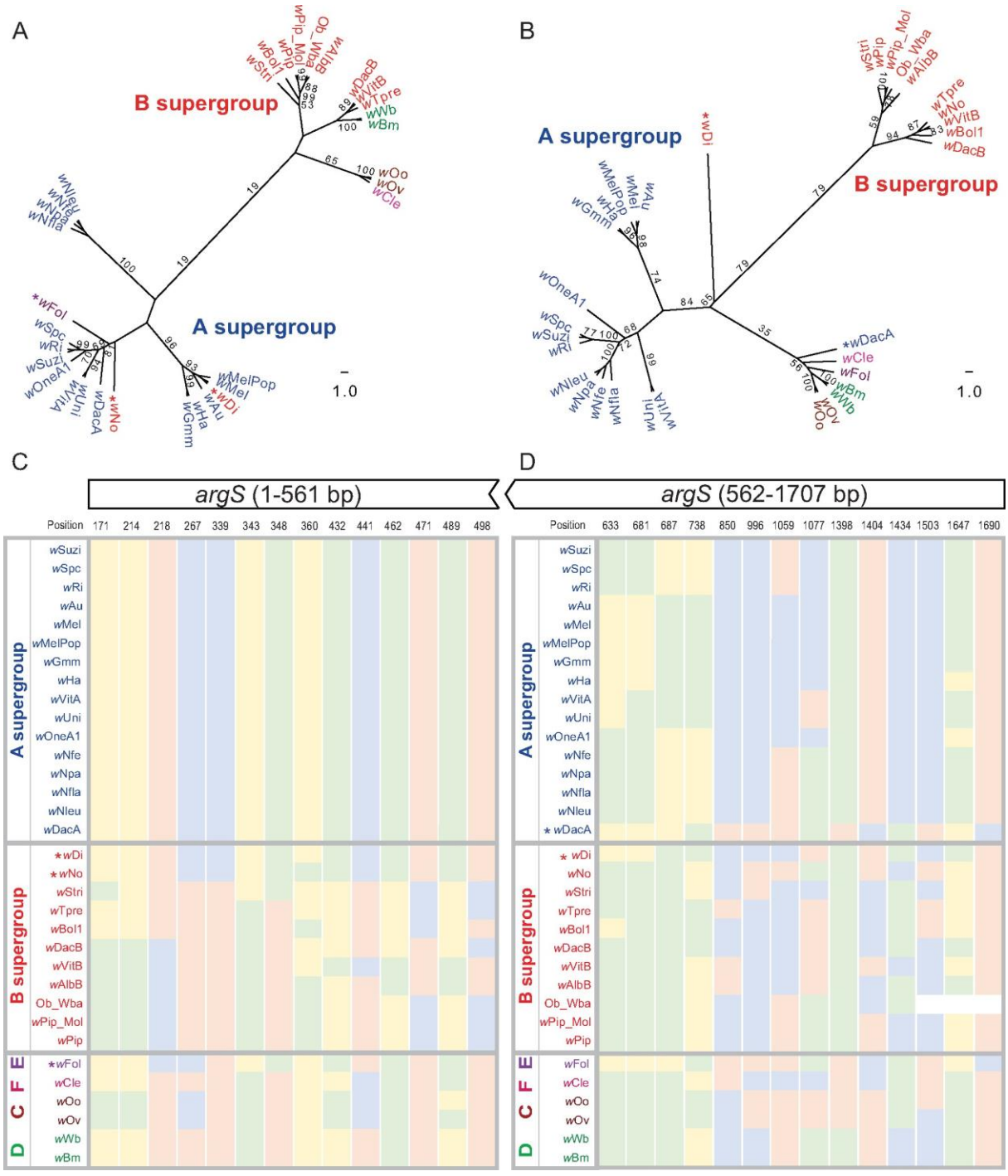


Figure 22. Intragenic recombination event between supergroups in *argS* gene.

Intragenic recombination event was detected by the GARD method in *argS* (*P*-value = 0.0006), and the inferred breakpoint is at 561 bp position in this gene. (A, B) Nucleotide ML trees for the 5' region (1-561 bp) and 3' region (positions 562-1707 bp), respectively. *argS* genes from *w*Di, *w*No, (B-*Wolbachia*), and *w*Fol (E-*Wolbachia*) cluster with A supergroup (A). The supergroup classifications follow the color code in the previous figures. Bootstrap values above 50 are shown in the figure. (A) *argS* from starting site to 561 bp, B-*Wolbachia w*Di, *w*No and E-*Wolbachia w*Fol cluster with A supergroup with 19 bootstrap support; whereas (B) *argS* from 562 bp to stop site, *w*Di (B-*Wolbachia*) clusters with A supergroup with 79 bootstrap support, and *w*DacA (A-*Wolbachia*) clusters with E-*Wolbachia* with 35 bootstrap support, indicating intragenic recombination events; (C) Nucleotides at selected positions (1-561 bp in *argS*) support the tree topology in (A); (D) Nucleotides at selected positions (562-1707 bp in *argS*) supported the tree topology in (B).

For the E supergroup there is only one released genome (*w*Fol). Nevertheless, I also found some evidence for recombination events between A and this single representative of the E supergroup. For instance, *w*Fol genes cluster with A-*Wolbachia* in *coxB, WONE_04820* and *argS* (Figure 21A, Figure 21B and Figure 22A). Given the high similarity among most sequenced A-*Wolbachia*, it is not possible to confidently identify which is the likely source. In addition, there appear to be two E-group genes that have transferred into the A-*Wolbachia* strain *w*DacA, *argS* and *DnaK* (Figure 22B and Figure 23). A better understanding of the evolutionary history of these transfers will be gained with additional E supergroup genome sequences.

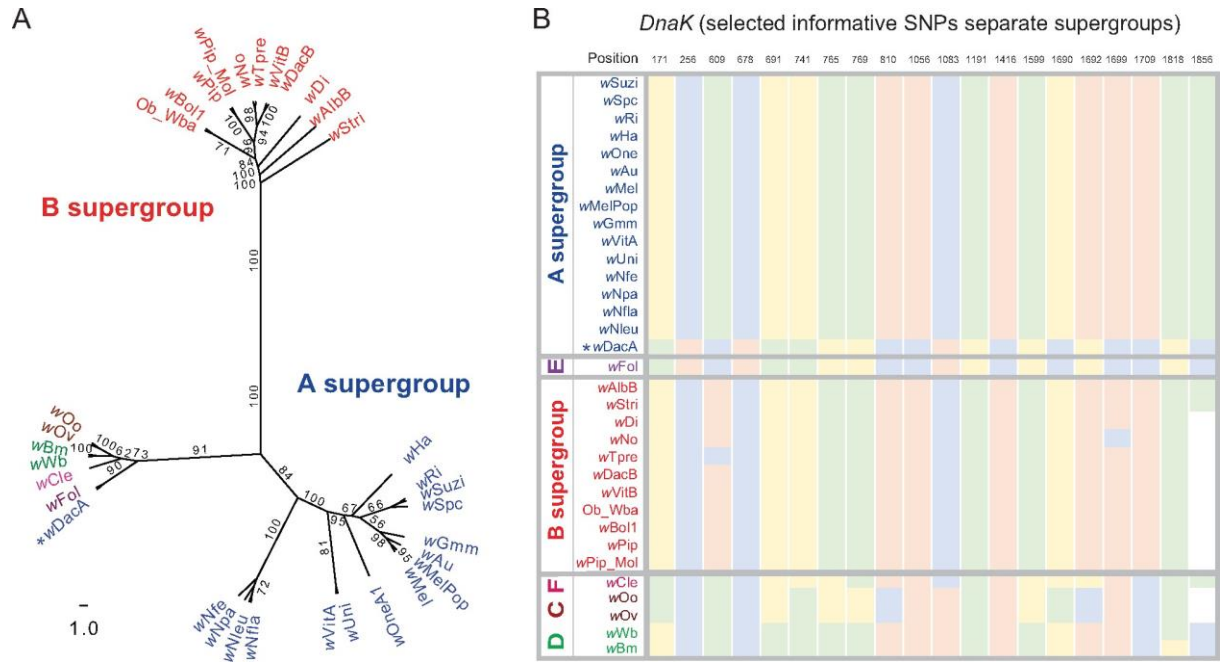


Figure 23. The nucleotide ML tree reveals recombination event where A-*Wolbachia* cluster with E-*Wolbachia* in *dnaK* gene.

The supergroup classifications follow the color code in earlier figures. Bootstrap values above 50 are shown in the figure. Nucleotides at selected positions are shown in the right panels. *w*DacA (A-*Wolbachia*) clusters with *w*Fol (E supergroup) with 90 bootstrap support. (A) The supergroup classifications follow the color code in the previous figures. Bootstrap values above 50 are shown in the figure. (B) Supergroup informative SNP positions in *dnaK* are plotted for all strains (green: A; blue: C; yellow: G; and pink: T).

Taken together, 97% of the single-copy orthologs agree with the supergroup classification in *Wolbachia***,** with a few cases of likely recombination events between *Wolbachia* strains of different supergroups. The recombination between A and B supergroups in gene *coxB* was reported by a previous study of 6 *Wolbachia* strains [200], and the remaining identified inter-supergroup

recombination events are novel findings in my study. The finding also indicates that these recombination events involve relatively small regions, rather than large recombination events involving many genes. The frequent gene order rearrangements observed in *Wolbachia* may make larger recombination tracks between supergroups less successful, as they are more likely to involve vital gene losses due to lack of synteny.

3.4.3 Concordance of MLST genes and whole genome divergence

The MLST system [59] has been variously used for strain typing of *Wolbachia*, identification of related strains, recombination within genes (e.g., the *wsp* locus) and phylogenetic inferences among strains. Recently, the reliability of the MLST system has been criticized [242], with whole genome sequencing stated to be preferred. Although whole genome data sets would always be desirable, the number of *Wolbachia* whole genome sequences is small compared to the many hundreds of MLST sequences currently available for comparative analyses. Therefore, I undertook to compare genetic divergence based on the MLST to the set of 210 genes in 33 different *Wolbachia* strains.

The MLST performed very well in both identifying closely related strains and in genetic divergence among strains compared to the genome-wide data set. The Pearson correlation coefficient of estimated evolutionary divergence with core gene set and *gatB*, *fbpA*, *hcpA*, *coxA*, *ftsZ* is 0.96, 0.9, 0.97, 0.92, and 0.97, respectively, with *P*-value $< 2.2 \times 10^{-16}$ (Table 13). The Pearson correlation coefficient of estimated evolutionary divergence with the core gene set and the concatenated MLST set is 0.98 with *P*-value $< 2.2 \times 10^{-16}$ (Figure 24). Eventually, whole genome data sets will supplant the MLST system. However, with over 1900 isolates in the *Wolbachia* MLST database, this will likely take some time, and until then, MLST remains a reliable method

for identifying closely related *Wolbachia* strains and their host associations. Furthermore, closely related *Wolbachia* strains identified by MLST that differ in host type of phenotypic effects on hosts (e.g., cytoplasmic incompatibility, feminization, male-killing, parthenogenesis, viral suppression), can be used for targeted whole genome sequencing to reveal possible mechanisms involved in host and phenotypic shifts.

Table 13. Correlation of evolutionary divergence estimates between *Wolbachia* species using 210 core gene set and five MLST genes

Correlation Coefficient (ρ)	**core gene set**	***gatB***	***fbpA***	***hcpA***	***coxA***	***ftsZ****
core gene set	1	0.96	0.90	0.97	0.92	0.97
gatB		1	0.86	0.91	0.90	0.94
fbpA			1	0.87	0.84	0.92
hcpA				1	0.89	0.96
coxA					1	0.92
ftsZ*						1

* Estimates of evolutionary divergence using *ftsZ* gene were only conducted among 31 *Wolbachia* species excluding *w*Bm, *w*Wb, and *w*Con, because of the inability to correctly annotate *ftsZ* in these three species.

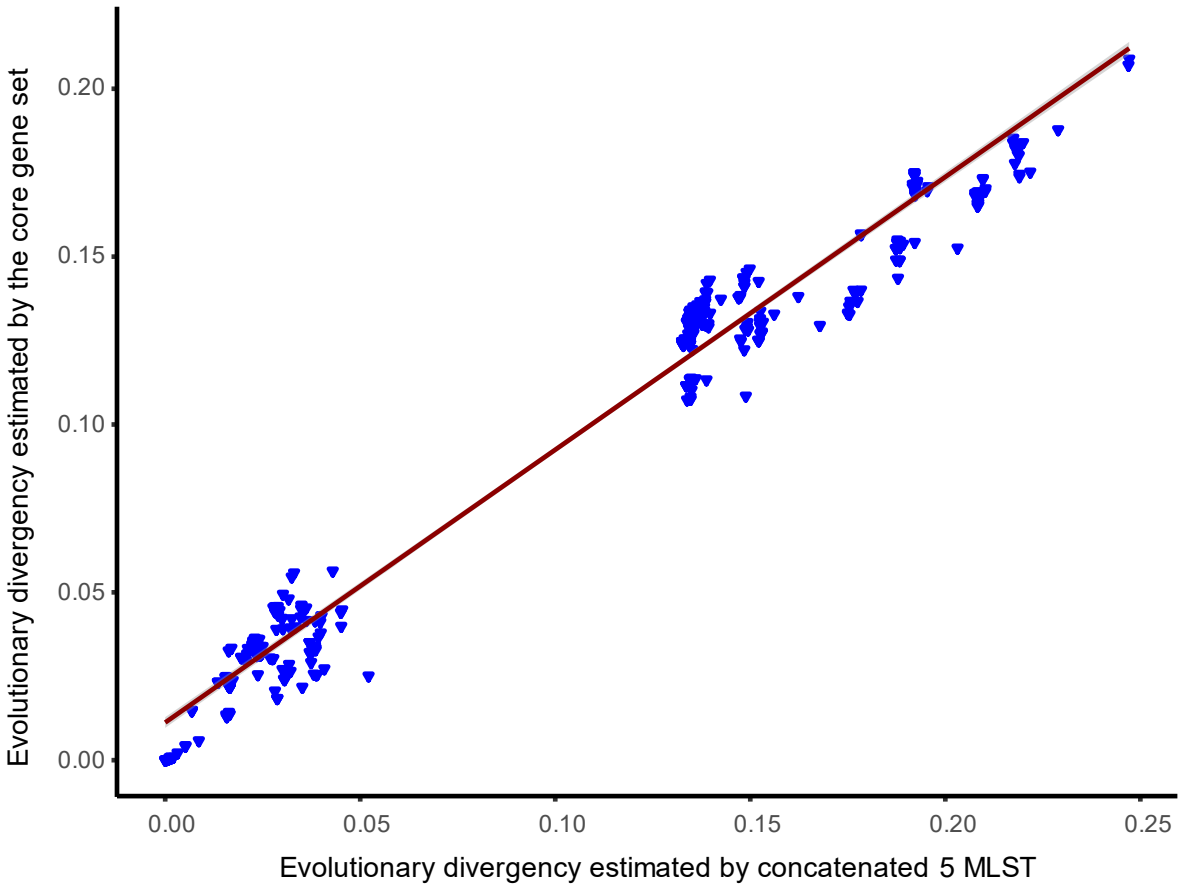


Figure 24. Correlation of evolutionary divergence estimated by core gene set and the five concatenated MLST genes.

Pearson correlation coefficient $r = 0.98$, P-value $< 2.2 \times 10^{-16}$

3.5 Discussion

The phylogenomic analysis of 33 annotated *Wolbachia* genomes in my study is the most comprehensive phylogenomic and evolutionary analysis conducted in *Wolbachia* strains to date. By including almost all available *Wolbachia* genomes in NCBI, I confirmed at the genome level that these *Wolbachia* strains group into distinct clusters (A, B, C, D, E, F supergroups) and different *Wolbachia* co-infected in the same host kept strain boundaries [200]. 205 of the 210 single-gene trees are consistent with the strain tree. Six gene trees have major rearrangements among *Wolbachia* groups (Figures 20-23), indicating potential recombination events between strains. I estimated that recombination events between supergroups occurred in at least 2.9% of the core genes in the *Wolbachia* genomes, and recombination may be one of the evolutionary forces shaping the *Wolbachia* genomes.

In total, there are a total of 14 recombination events detected in six genes. Nine of these involve A-B recombination in five genes. The five genes with distinct tree structure differences from the consensus *Wolbachia* tree include *ftsH*, *rpIU*, *coxB*, hypothetical protein *WONE_04820* and *argS*. In addition, five events were detected between the A and E supergroups. Most recombination events involved the entire gene, whereas a single intragenic event was found in *argS*. A second intragenic event may also be present in *coxB* (Figure 21C), although it was not detected by the Interclade Recombination or GARD methods. I conclude that inter-supergroup recombination is uncommon among the set of 210 core single ortholog genes used in this study. Recombination may be more frequent in other genes, and clearly is so in phage-associated genes [243, 244] and the surface protein *wsp* [227]. Furthermore, within supergroup recombination is

also likely to be more common, although also more difficult to quantify due to the greater similarity within these groups.

Among the 14 recombination events observed, *argS* (5 events) and *WONE_04820* (4 events) appear to be particularly prone to inter-supergroup recombination (Table 11). *argS* is a class I aminoacyl-tRNA synthetase, which catalyzes the ligation of arginine to its transfer RNA, while the function of *WONE_04820* is not clear. *WONE_04820* is conserved in *Wolbachia*, and no known functional domains could be identified. In addition, two bacterial strains appear to be more prone to inter-supergroup recombination (*w*DacA and *w*Di). Notably, both are found in hemipterans. More sequencing of *Wolbachia* from different insect orders is needed, as the current set are predominantly from Diptera and Hymenoptera.

Recombination events among A and B *Wolbachia* supergroups have been documented in previous studies, and I identified additional cases through the phylogenomic analysis among 33 sequenced genomes. Interestingly, I also discovered recombination events between A and E supergroups, which were not known previously. The E group *Wolbachia* is found in springtails [245-247]. A recent study characterized the *Wolbachia* in 11 collembolan species by MLST, and found that nearly all are E-group *Wolbachia* that are monophyletic, based on phylogenetic reconstruction using MLST genes [248]. My genome analysis of the single collembolan *Wolbachia* genome reveals a number of candidate recombination events, including intergroup recombination between A and E in *coxB*, *dnaK*, *WONE_04820*, and *argS*. Targeted sequencing of these genes in the additional collembolan species or additional genome sequencing will help reveal the origins and directions of these events. I further speculate that selective maintenance of such transfers could suggest a possible role in E *Wolbachia* function, such as parthenogenesis induction found in this

springtail [248]. The focus of this study has been on recombination between supergroups, where the phylogenetic signal-to-noise ratio is much stronger. However, although more difficult to document, intra-supergroup recombination is likely to be more extensive than between supergroups, and is a topic worthy of future study.

It has been recently argued that MLST genotyping has little utility in phylogenetic analyses, and should be supplanted by genomic studies [242]. When the MLST system was developed, it was pointed out by the authors that the system would be most useful for identifying relatively closely related *Wolbachia*, due to potential recombination among more divergent strains [58]. However, the comparison on genome sequence indicates that MLST typing is largely valid, both for supergroup identification and detection of closely related strains. Related *Wolbachia* based on MLST results are also closely related in the genome-wide analysis. This suggests that, until *Wolbachia* genome sequencing becomes much less expensive and can be readily performed on single arthropods, that MLST will remain a useful tool for the identification of strains, their relationships, and host affinities. Nevertheless, caution should be exercised due to some documented recombination events within MLST genes and among them [57]. Therefore, topologies should be compared among genes for evidence of discordance, rather than simply relying on phylogenetic reconstructions of concatenated sequences.

Chapter 4 Genome evolution and transmission mechanism of the microsporidian pathogen *Nosema muscidifuracis*

4.1 Abstract

Nosema is a diverse genus of microsporidian parasites, which are all unicellular, obligate fungal symbionts and pathogens of insects and other arthropods. *Nosema muscidifuracis* infects parasitoid wasp species of *Muscidifurax zaraptor* and *M. raptor* (Hymenoptera: Pteromalidae). In this study, I report high-quality assemblies (14,397,169 bp in 28 contigs) of *N. muscidifuracis* genomes and comparisons to other *Nosema* genomes. Using PacBio long-read sequencing technology, a novel composite 4-bp (TAGG)n and 5-bp (TTAGG)n telomeric repeat motif was discovered at the ends of chromosomes, which represent the first identified telomeres for *Nosema* (and other microsporidia). A total of 2,782 protein-coding genes were annotated, with 66.2% of the genes having two copies and 24.0% of genes having three copies. These duplicated genes are highly similar, with a sequence identity of 99.3%. The complex pattern suggests extensive gene duplications and rearrangements across the genome. I annotated 57 rDNA loci, which are extremely GC-rich (37%) in a GC-poor genome (25% genome average). *Nosema*-specific qPCR primer sets were designed based on 18S rDNA annotation as a diagnostic tool to determine its titer in host samples. I discovered high *Nosema* titers in *Nosema*-cured *M. raptor* and *M. zaraptor* using heat treatment in 2017 and 2019, suggesting that the remedy did not completely eliminate the *Nosema* infection. *N. muscidifuracis* shared 449 orthologous genes with six other genome-sequenced *Nosema* species. Comparative phylogenomic analyses revealed incongruency in the *Nosema* and host species trees, indicating a host switch event between parasitoid wasps and bees. In *N. muscidifuracis*, a highly significant ACCC motif was found within 20 bp upstream of the

translation start codon ATG. This motif is present in 90% of highly expressed genes, in sharp contrast to ~20% in lowly expressed genes, and therefore serves as a candidate *cis*-element for positive regulation of gene expression. Strikingly, similar (C)3 and (C)4 motifs were also discovered in other distantly related *Nosema* species, suggesting a conserved *cis*-regulatory mechanism. Cytogenetic analyses revealed a substantial *Nosema* load within the ovaries of *M. raptor* and *M. zaraptor*, consistent with a heritable component of infection and per ovum vertical transmission. *Nosema* are widespread pathogens, including inducing epizootics in honeybees. The parasitoid-*Nosema* system is laboratory tractable, and, therefore, can serve as a model to inform future genome manipulations of *Nosema*-host system for investigations of Nosemosis. This study also provides novel insights into the genetic architecture, gene regulation, and genome evolution of *Nosema* species and will enhance the understanding of host-parasite interactions.

4.2 Introduction

As a genus of microsporidia, *Nosema* can infect a wide range of hosts. Most species parasitize insects and other arthropods [63, 64], such as *Apis mellifera* [69], *Apis cerana* [70], *Pieris rapae* [71], *Bombyx mori* [72], *Antheraea pernyi* [73], and *Gammarus duebeni* [74]. *Nosema* also infects other beneficial insects, including the parasitoid wasp genus *Muscidifurax* (Hymenoptera: Pteromalidae). The genus *Muscidifurax* is a natural biocontrol agent of the dipteran filth flies with nine identified species, all of which are pupal parasitoid wasps. *Nosema muscidifuracis* (*N. muscidifuracis*) infects parasitoid wasp species of *Muscidifurax zaraptor* and *M. raptor*, causing ~50% reduction in longevity and ~90% reduction in fecundity [38].

Nosema disease reduction methods have been explored to cure the *N. muscidifuracis* infection. Exposure of parasitoid eggs within host pupae at several temperatures (45°C, 47°C, and

50°C) was shown to be effective in managing *Nosema* disease and increasing parasitoid fecundity [249]. A 100% cure rate was achieved at 50°C for 45 minutes with a relative survival of 18%. To completely eliminate the pathogen, heat treatment in combination with the Pasteur method was applied: heat treatment minimizes disease prevalence and ensures an adequate genetic base for healthy parasitoids, and the Pasteur method (based on visual examination for patent infections) isolates uninfected wasps as parents for rearing their progeny [38, 250]. However, the efficacy of this approach depends on the efficiency of visual detection, which may not detect low-level infections.

Understanding the *Nosema* disease transmission mechanism is critical for developing new control methods, but the transmission of *Nosema* in parasitoid wasps is not fully understood yet. The following is known about transmission patterns: 1) maternal transmission is highly efficient; 2) adults can acquire infection by feeding on spore suspensions or infected parasitoid immatures within hosts; 3) infected male adults do not transmit infections to healthy females; 4) house fly hosts do not become infected; and, 5) horizontal transmission occurs when healthy immatures feed on infected larvae in superparasitized hosts [38]. The intracellular nature of *Nosema* suggested that transmission can occur vertically, either within the egg (cytoplasmic), and/or on the egg surface (per ovum). If per ovum or through ingestion by feeding larvae, parasitoid wasp lines can be cured of the infection by surface sterilization or egg transfer experiments. To establish a parasitoid-*Nosema* model for Nosemosis research, a high-quality reference genome is essential for studying gene expression changes and gene manipulations in *Nosema*.

To adapt the intracellular parasitism, microsporidia exhibit dramatic reduction in common eukaryotic features at the molecular, cellular, and biochemical levels [68, 75, 251]. In

Encephalitozoon cuniculi, the mitochondria are highly reduced into streamlined "mitosomes" [252]. A gene encoding the heat shock protein HSP70, derived from the mitochondrial endosymbiont, is involved in folding other proteins in the process of import into the mitochondrion. No mitochondrion physically identified in microsporidian cells, and the presence of the mitochondrial-type HSP70 in *Encephalitozoon cuniculi*, *Vairimorpha necatrix*, and *Nosema locustae* provide evidence for loss of mitochondria in Microsporidia [253-255]. Phylogenetic analysis clearly placed the HSP70 gene in *N. locustae* in the mitochondrial group, suggesting the mitochondrial evolutionary origin of Microsporidia [256]. Collectively, Microsporidia were derived from lineages containing mitochondria and retained some mitochondrial-type genes, but it is unknown whether the mitochondria were completely lost in these highly reduced, intracellular parasites.

In the previous study, a sequence motif characterized by a thymine homopolymer upstream of a highly over-represented cytosine triplet was identified in *N. ceranae*, which predominantly located within 15 bp of the start codon [257]. A similar motif with a shorter length containing a cytosine triplet was identified in *E. cuniculi*, but it is not statistically significant [257]. The motif was identified only in *N. ceranae*, which infects honey bees, and the comparison was not conducted across different *Nosema* species in a wide range of hosts. In host-parasite associations, events such as host shift, duplication, or extinctions disrupt cophylogenetic patterns, result in incongruences between host and parasite phylogenies [258]. The *Dictyocoela muelleri* microsporidia, which infects *Gammarus roeselii* only in the recently colonized region, revealed recent host shifts from local host species after the spread of *G. roeselii* [259]. The phylogenetic congruence of *Dictyocoela duebenum* and *Gammarus balcanicus* gives another example of host shifts between different *G.*

balcanicus cryptic lineages [260]. I performed comparative genomic analysis to investigate the phylogenetic relationship between *N. muscidifuracis* and other *Nosema* species in different hosts.

In this study, I sequenced and assembled the *N. muscidifuracis* reference genome in the parasitoid wasp species *M. zaraptor* using PacBio long-read sequencing, characterized the genetic architecture of the *N. muscidifuracis* genome, developed qPCR assays for accurate quantification of *Nosema* titer, and explored the vertical transmission mechanisms using cytogenetic analyses. The pattern of motif was identified in seven *Nosema* genomes, and I also determined the potential effect of the sequence motif on gene expression. Furthermore, I extended my efforts in comparative genomic analysis of *Nosema* species in wide range of hosts to identify the evolutionary relationship between *Nosema* species and their hosts.

4.3 Materials and methods

4.3.1 Sample source and insect rearing

The genus *Muscidifurax* is a natural biocontrol agent of the dipteran filth flies with nine identified species, all of which are pupal parasitoid wasps. *Muscidifurax raptor* Girault and Sanders was the first species characterized in 1910 [23]. In 1970, four sibling species in this genus were described: *M. zaraptor* Kogan and Legner, from the southwestern United States; *M. raptoroides* Kogan and Legner from Central America and Mexico; *M. raptorellus* Kogan and Legner from Uruguay and Chile; and a thelytokous species *M. uniraptor* Kogan and Legner from Puerto Rico [261]. *M. uniraptor* only produces a female offspring, and the parthenogenesis is caused by an intracellular bacterium, which is an A-supergroup *Wolbachia w*Uni [32, 33].

In this study, three *Muscidifurax* species were used. The source of *M. zaraptor* was from

two independent colonies, and both of them were derived from the same USDA colony maintained by Dr. Chris Geden's laboratory, which was originally collected in 2015 from dairy farms in Minnesota, Nebraska, and California. The USDA *M. zaraptor* colony is maintained in the Geden laboratory at the Center for Medical, Agricultural and Veterinary Entomology, USDA Agricultural Research Service (USDA-ARS, Gainesville, FL, USA). An attempt was made to cure the colony for *Nosema* infection in 2019 using a heat shock treatment approach [262]. The *M. zaraptor* colony maintained in the Wang laboratory (AUB colony) at Auburn University College of Veterinary Medicine (Auburn, AL, USA) was obtained from the University of Rochester in 2019, which was derived from the USDA colony infected with *N. muscidifuracis*. The *M. uniraptor* colony at AUB was also obtained from the University of Rochester in 2019, and it is *Nosema*-free. The *M. raptor* colony maintained in the Wang laboratory at AUB was derived from a *Nosema*-cured colony treated in 2017 in the Geden laboratory (see Materials and Methods). All AUB colonies were maintained on commercial flesh fly (*Sarcophaga bullata*) pupae (Ward's Science, Rochester, NY, USA) at a constant temperature of 25°C and 24h constant light in the Wang laboratory. The USDA colonies were maintained on housefly (*Musca domestica*) pupae in the Geden laboratory (Gainesville, FL).

4.3.2 High molecular weight DNA extraction, PacBio CCS library preparation and sequencing

High molecular weight (HMW) genomic DNA (gDNA) was extracted from *M. zaraptor* whole-body samples collected 24 hours after eclosion from the AUB colony using Genomic-tip 20/G kit (Qiagen, MD, USA). The DNA concentration was measured on a Qubit 3.0 Fluorometer instrument (Thermo Fisher Scientific, MD, USA). The gDNA quality and the size distribution

were assessed on an Agilent TapeStation 4200 machine (Agilent Technologies, CA, USA) with genomics screen tapes. A total of 10 μg high-quality *M. zaraptor* HMW gDNA was sheared into 20 Kb fragments. After end-repair and ligating the specific adapter oligos, the DNA fragments were annealed with sequencing primer v2 and Sequel II DNA Polymerase, bound to the SMRTbell templates, and the library was prepared using the SMRTbell Template Prep kit v2 with the CCS HiFi Library construction protocol (Pacific Biosciences, CA, USA) at the HudsonAlpha Genome Sequencing Center (HGSC, Huntsville, AL, USA). The concentration and the size distribution for the prepared library were determined on LabChip GX Touch HT (PerkinElmer, MA, USA), and sequenced on a PacBio Sequel II System at HGSC.

4.3.3 10× Genomics linked-read library construction and Illumina sequencing.

HMW gDNA from AUB *M. zaraptor* was diluted to ~0.8 ng/ul with EB buffer through a series of dilutions, with concentrations determined by Qubit 3.0 Fluorometer (Thermo Fisher Scientific, USA). The Chromium Genome Reagent Kit v2 (10× Genomics Inc., CA, USA) was used for linked-read library preparation, according to the manufacturer's instructions. The genome chip was loaded with diluted denatured gDNA, sample master mix, and gel beads following the protocol. Gel Bead-In-EMulsions (GEMs) were generated using a 10× Chromium Controller. After the incubation and cleanup of the obtained GEMs, Chromium i7 Sample Index was ligated and served as the library barcode to provide linked information. The size distribution of the prepared library was assessed using Agilent TapeStation 4200 (Agilent Technologies, CA, USA), final library quantity was checked with Qubit 3.0 Fluorometer (Thermo Fisher Scientific, USA). After quality control, the 10× Genomic library was sequenced on an Illumina NovaSeq 6000 machine.

4.3.4 Assembly of the *N. muscidifuracis* Genome

De novo genome assembly for the *M. zaraptor* genome was performed with 23.7 Gb PacBio HiFi reads (377.2 Gb raw reads) using dedicated long-read assemblers hifiasm v0.13 [124]. The 10× Genomics reads were aligned to the assembled contigs using the LongRanger pipeline v2.1.6 [128]. The identity of *M. zaraptor* contigs was determined by coverage depth (54.5X Illumina read depth) and homology to closely-related *M. raptorellus* with a high-quality reference genome available [263]. Contigs from six microbial species were also detected in the hifiasm assembly, including five bacterial species and *N. muscidifuracis*. A total of 30 *N. muscidifuracis* PacBio HiFi contigs were identified based on coverage depth (226.7X Illumina read depth) and GC content (25.2%) [238]. *De novo* assembly of 10× Genomics data was performed using Supernova v2.1.1 with default parameters [123]. Overlap of HiFi contigs and 10× scaffolds were detected by quickmerge v0.3.0 [131]. The 30 *N. muscidifuracis* contigs were merged into 28 contigs based on manual inspection of the contig overlap.

4.3.5 Assessment of *Nosema* genomes

The final genome completeness of *N. muscidifuracis* was assessed by BUSCO (Benchmarking Universal Single-Copy Orthologs) v5.3.2 [264]. The BUSCO scores were computed using microsporidia_odb10 with a total of 600 orthologs. The BUSCO scores were also computed for an outgroup microsporidian species, *Encephalitozoon cuniculi* [75, 76], as well as eight other *Nosema* species/strains, including *N. ceranae* BRL 01 [257], *N. ceranae* PA08 [79], *N. ceranae* BRL [70], *N. apis* BRL 01 [69], *N. bombycis* CQ1 [72], *N. granulosis* Ou3-Ou53 [74], *N. antheraeae* YY [73], and *Nosema sp.* YNPr [71] (Table 14).

Table 14. Genome assembly statistics for *Nosema muscidifuracis* and comparison with other microsporidian genomes

Species (strain)	Accession/citation	Size (bp)	scaffold/contig N50 (Kb)	# of scaffolds/ contigs	BUSCO complete (fragmented/ missing)
N. muscidifuracis	this assembly	14,397,169	544.3/544.3	28/28	97.0% (1.2%/1.8%)
N. apis (BRL01)	GCA_000447185.1	8,569,501	24.3/14.0	554/1133	75.0% (6.0%/19.0%)
N. ceranae (BRL)	GCA_004919615.1	8,816,425	177.3/177.3	110/110	97.0% (1.2%/1.8%)
N. ceranae (PA08)	GCA_000988165.1	5,690,748	42.6/42.6	536/536	97.0% (1.3%/1.7%)
N. ceranae (BRL01)	GCA_000182985.1	7,860,219	2.9/2.9	5465/5465	93.9% (3.5%/2.6%)
Nosema sp. (YNPr)	NA	3,637,996	12.2/3.8	462/2272	84.5% (2.7%/12.8%)
N. antheraeae (YY)	NA	7,100,626	172.2/25.6	202/719	95.3% (1.5%/3.2%)
N. bombycis (CQ1)	GCA_000383075.1	15,689,776	57.4/6.1	1607/3558	83.0% (5.0%/12.0%)
N. granulosis (Ou3-Ou53)	GCA_015832245.1	8,859,703	12.7/9.4	1754/2007	95.9% (1.8%/2.3%)
E. cuniculi (GB-M1)	GCA_000091225.2	2,497,519	220.3/218.3	11/12	100.0% (0.0%/0.0%)

4.3.6 Genome size estimation

Illumina short-reads aligned to *N. muscidifuracis* assembly were utilized to estimate the genome size. Low-quality bases and adapter sequences were trimmed using Trimmomatic version 0.39 [115], with the parameters "ILLUMINACLIP:adapter:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:60." High-quality trimmed reads were used for multiple k-mer counting using Jellyfish version 2.3.0 [265] with parameters count "-m 25 -s 20G -t 48." The genome size and heterozygosity were estimated using GenomeScope [266].

4.3.7 Telomeric repeat identification

TRIP pipeline (Telomeric Repeats Identification Pipeline) [267] was used to *de novo* predict the candidate telomeric repeat motifs (TRMs) from publicly available short-read sequencing data of *Nosema* genomes (Table 15). With PacBio long-read assembly, the telomeric repeat motifs of *N. muscidifuracis* were determined from the repetitive regions at the termini of multiple contigs. To confirm the telomeric repeats, I extracted and aligned the sequences of assembled telomeric and sub-telomeric regions in the *N. muscidifuracis* genome. The phylogenetic tree based on 27 nucleotide sequences in ~20 Kb subtelomeric region was constructed to determine the evolutionary relationship of the conserved sub-telomeric sequences.

Table 15. Short-read genome sequencing data in *Nosema* species for telomeric repeat motif identification

Species (strains)	Data Accession numbers	Total number of reads	Candidate telomeric repeat motif	Sequencing platform
N. apis (BRL01)	SRX245851	493,431	N/A	454 GS FLX
N. ceranae (BRL)	SRX5338655	2,186,202	N/A	MinION
N. ceranae (PA08)	SRX318182	6,106,172	TTAGG	Illumina HiSeq 2000
N. ceranae (BRL01)	SRX003255	1,063,647	N/A	454 GS FLX
Nosema sp. (YNPr)	NA	Raw data not available		
N. antheraeae (YY)	NA	Raw data not available		
N. bombycis (CQ1)	SRX7209795	26,849,840	N/A*	Illumina HiSeq 2000
N. granulosis (Ou3-Ou53)	SRX5286701	7,327,280	N/A	Illumina MiSeq
E. cuniculi (GB-M1)	ERS610230	Raw data not available		

* cDNA was sequenced for this species.

4.3.8 Repeat annotation

Before gene prediction, repeat annotation was performed to identify the repetitive elements in *N. muscidifuracis* genome. I first constructed a *de novo N. muscidifuracis* repeat database using RepeatModeler v2.0.1 [137], which provides a list of repeat family sequences. The repeat-identifying was implemented by three complementary computational programs, RECON v1.0.8 [138], RepeatScout v1.0.5 [139], and Tandem Repeats Finder (TRF) [140]. Based on the transposon element library, the homologous repeats and low-complexity DNA sequences were masked using RepeatMasker v4.1.1 [141] with RMBlast v2.10.0 sequence search engine (http://www.repeatmasker.org).

4.3.9 Noncoding RNA annotation

Noncoding RNAs (ncRNAs) were predicted using the Infernal software version 1.1.2 [268] based on the multiple sequence alignments using the covariance models in the Rfam database [269]. Before prediction, the esl-seqstat program was used to determine the total database size for the *N. muscidifuracis* genome. The ncRNAs in the *N. muscidifuracis* genome were predicted and annotated using the cmscan program in Infernal software [268] with RNA families in the Rfam [269] database.

4.3.10 *Nosema* inspection and treatment procedures

Colonies of *M. raptor* were established in 2015 from parasitoids collected from dairy farms in Florida, Minnesota, Nebraska, and California, and samples of parasitoids were examined visually for *Nosema* infection by crushing wasps in a drop of sterile water on a microscope slide and examining at 400× for the presence of spores [95]. None of the specimens examined at the

time showed patent infections. After one year in colony, 100% of examined parasitoids had patent infections, indicating that low-level infections had escaped detection at the time of colony founding. In 2017, colonies were heat-treated by exposing multiple cohorts of newly parasitized house fly pupae to 50^{o}C for 45 min and holding them for adult emergence [249]. Emerged adults were examined visually for infection as before, and progeny from cohorts that were *Nosema*-free based on visual inspections were used to start new colonies. Examination of these colonies after three additional generations indicated a resurgence of infection in three out the four strains. A final attempt was made to eliminate *Nosema* disease from the colonies by first subjecting three successive generations of parasitoids from the four strains to heat treatment as before and holding them for adult emergence. This procedure was conducted with three separate lineages of each of the four strains. Parasitized pupae were then held in individual gelatin capsules for emergence. Pairs of parasitoids (n=100 pairs for each strain) were then held with 70 house fly pupae/pair for three days, then given another set of 70 for three more days. Each female was examined individually for the presence of spores; it was expected that six days would be sufficient time for infections to be evident by visual inspection. Progeny of any females with patent infections were discarded, and the parasitized pupae from females without patent infections were pooled to start a new colony.

4.3.11 Genomic DNA extraction and *Nosema* titer determination using quantitative PCR

To determine the levels of *Nosema* infection in the *Nosema*-cured *M. raptor* AUB colony, *Nosema*-infected *M. zaraptor* AUB colony and *Nosema*-cured *M. zaraptor* USDA colony, I extracted genomic DNA from 24-hour adult male and female samples with three replicates per sex using AllPrep DNA/RNA Mini Kit (Qiagen, MD). Three primer sets were designed to target the

different regions of the 18S rDNA gene in *N. Muscidifuracis* using Oligo 7 primer analysis software (Molecular Biology Insights Inc., Cascade, CO, USA) (Table 16). Primers were synthesized at Eurofins Genomics LLC (Louisville, KY, USA). All primer sets were evaluated by PCR using the *M. zaraptor* DNA template, with *Nosema*-free DNA as a negative control, followed by electrophoresis on 2% agarose gel (Figure 25). The PCR experiments were performed with Q5 High-Fidelity 2x Master Mix on an Eppendorf Mastercycler Pro PCR machine (Eppendorf North America, Enfield, CT, USA). The thermal cycling protocol for the primers was as follows: initial denaturation step at 98°C for 30 s, followed by 30 cycles of initial denaturation at 98°C for 10 s, annealing at 55°C for 30 s, and extension at 72°C for 15 s. The qPCR reaction was conducted in a 20 µL system using Luna® Universal qPCR Master Mix (New England BioLabs, Ipswich, MA, USA). Each reaction contained 10 µL of Luna Universal qPCR Master Mix, 8 µL of nuclease-free water, 0.5 µL of forward primer and 0.5 µL reverse primer (10 µmol/L), and 1 µL of DNA template. The Bio-Rad C1000 Touch Thermal Cycler with CFX96 Real-Time PCR Detection Systems (Bio-Rad Laboratories, Hercules, CA, USA) was used to conduct qPCR experiments with the SYBR scan mode using the *Nosema* 18S NP2 primer set (Table 16). The relative quantification method was applied to determine the *Nosema* abundance in *M. raptor* AUB colony and *M. zaraptor* from both AUB and USDA colonies. The thermocycling conditions for the qPCR assays were 95 °C for 60 s, followed by 40 cycles at 95 °C for 15 s and 50 °C for 30 s.

Table 16. The 18S primers used for the quantification of *Nosema* titer in the *Muscidifurax zaraptor* genome

Species name	Target gene	Primer name	Primer sequence	Product size	PCR Ta
N. muscidifuracis	18S	NP1_F	GAAGAAGTATCTGAAAAATGGAC	151 bp	49.3°C
N. muscidifuracis	18S	NP1_R	CGTTACTGCCTTGTTAAGCC		
N. muscidifuracis	18S	NP2_F	AAGAAGTATCTGAAAAATGG	213 bp	49.8°C
N. muscidifuracis	18S	NP2_R	CTTAGACTTAGTAGCCGTCTC		
N. muscidifuracis	18S	NP3_F	TTATAGACAGACACAATCAG	225 bp	49.7°C
N. muscidifuracis	18S	NP3_R	ATATCATCTTAGATAGCGACGG		

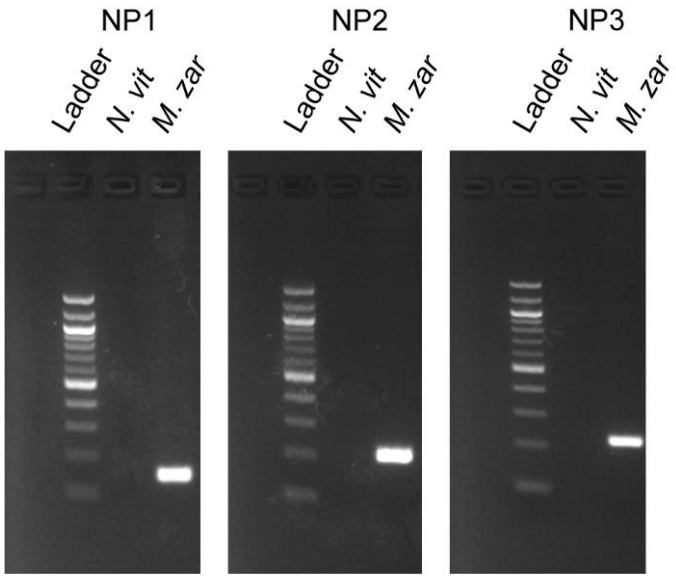


Figure 25. PCR results for three primer sets targeting the 18S rDNA gene in *Nosema muscidifuracis*.

The PCR amplification products for three 18S rDNA gene primer sets in *M. zaraptor* DNA sample were evaluated by a 2% agarose gel alongside a Quick-Load® 100 bp DNA Ladder, with *N. vit*

(*Nasonia vitripennis*, *Nosema*-free) as a negative control. The gel was run at 120 V for 20 min. The sizes of PCR products for NP1, NP2, and NP3 are 151 bp, 213 bp, and 225 bp, respectively. NP1: 5'-GAAGAAGTATCTGAAAAATGGAC-3' and 5'-CGTTACTGCCTTGTTAAGCC-3'; NP2: 5'-AAGAAGTATCTGAAAAATGG-3' and 5'-CTTAGACTTAGTAGCCGTCTC-3'; NP3: 5'-TTATAGACAGACACAATCAG-3' and 5'-ATATCATCTTAGATAGCGACGG-3'.

4.3.12 RNA sample quality control, RNA-seq library preparation and sequencing

Adult male and female *M. zaraptor* were collected 24 hours after eclosion from both the USDA colony and AUB colony in the Wang laboratory. Total RNA was extracted from the adult whole-body samples in three biological replicates for each sex. RNA extractions were performed with AllPrep DNA/RNA Mini Kit (Qiagen, MD, USA) following the manufacturer's protocol. The RNA yield was quantified using a Qubit 3.0 Fluorometer instrument (Thermo Fisher Scientific, MD, USA), followed by a quality check on an Agilent TapeStation 4200 Bioanalyzer (Agilent Technologies, CA, USA). For RNA-seq library preparation, 1 μg of total RNA was used as input for all samples. To remove the abundant rRNA (ribosomal RNA) in the sample, an rRNA removal protocol was performed using the NEBNext rRNA Depletion Kit (New England Biolabs, MA, USA). The remaining mRNA was used for RNA-seq library construction using NEBNext Ultra II Directional RNA Library Prep Kit (New England Biolabs, MA, USA) with the manufacturer-provided protocol. After quality control, the RNA-seq library was sequenced on an Illumina NovaSeq6000 platform.

4.3.13 RNA-seq data processing and gene annotation

On average, 128 million 150-bp reads were generated for each of the 12 *M. zaraptor* RNA-

seq libraries. The raw RNA-seq data was checked for sequencing quality using FastQC [122]. Adapter sequences and low-quality bases in the paired-end RNA-seq reads were trimmed with Trimmomatic v0.39 [115] (Table 17). A total of 6.2 million non-rDNA reads mapped uniquely to the *N. muscidifuracis* were used for *de novo* transcriptome assembly by Trinity v2.4.0 [135]. Protein-coding genes were predicted and annotated in the *N. muscidifuracis* genome assembly using Fungal Genome Annotation Pipeline (FunGAP) [270], and filtered RNA-seq reads as input. FunGAP masked the repeats in the genome and assembled the RNA-seq reads. Augustus [271] was used for gene model prediction with species parameter "--augustus_species encephalitozoon_cuniculi_GB", which is the closest to *Nosema*. The microsporidian protein database used by FunGAP was downloaded by "download_sister_orgs.py" script in FunGAP.

Table 17. RNA sequencing sample information, data yield, quality control summary statistics

Library ID	Sex	Host	Replication information	# of paired-end reads	# of reads after QC	% of reads after QC
Mzar_adultF_FL_rep1	Female	fleshfly	replicate 1	85,853,492	84,584,114	98.52%
Mzar_adultF_FL_rep2	Female	fleshfly	replicate 2	57,015,645	56,227,715	98.62%
Mzar_adultF_FL_rep3	Female	fleshfly	replicate 3	69,336,408	68,326,961	98.54%
Mzar_adultF_HL_rep1	Female	house fly	replicate 1	76,735,261	75,692,065	98.64%
Mzar_adultF_HL_rep2	Female	house fly	replicate 2	77,203,402	75,967,748	98.40%
Mzar_adultF_HL_rep3	Female	house fly	replicate 3	56,990,139	56,109,147	98.45%
Mzar_adultM_FL_rep1	Male	fleshfly	replicate 1	49,640,375	48,903,854	98.52%
Mzar_adultM_FL_rep2	Male	fleshfly	replicate 2	60,168,883	59,363,068	98.66%

Mzar_adultM_FL_rep3	Male	fleshfly	replicate 3	61,031,813	60.048,859	98.39%
Mzar_adultM_HL_rep1	Male	house fly	replicate 1	57,421,992	56,504,800	98.40%
Mzar_adultM_HL_rep2	Male	house fly	replicate 2	51,761,573	50,944,361	98.42%
Mzar_adultM_HL_rep3	Male	house fly	replicate 3	64,986,450	64,014,005	98.50%

4.3.14 Comparative genome analysis

To compare the genomes of *N. muscidifuracis* and its closely-related species *N. ceranae*, MCscanX [156] was used to perform synteny analysis and identify syntenic blocks between these genomes based on core orthologous gene sets identified using BlastP with default settings (E-value ≤ 1e−5; minimum number of genes in a syntenic block ≥ 5). The genome and annotation files of *N. ceranae* (BRL strain) were downloaded at NCBI Assembly with the accession number GCA_004919615.1. The genomic circle of collinearity was visualized in Circos [157]. To check the location of duplicated genes in *N. muscidifuracis*, I detected and plotted the collinear blocks within the *N. muscidifuracis* genome using MCscanX [156]. The duplicated genes were aligned using mafft software (version 7.475) [272]. The identity of the two sequences for each gene pair was computed by the Needle program with the default parameters (Gap penalty=10, Extend penalty=0.5) using the Needleman-Wunsch algorithm [273]. The numbers of synonymous (dS) and non-synonymous (dN) substitutions between two sequences were calculated using the alignments by KaKs_Calculator software (version 2.0) with γ-YN method [274-277].

4.3.15 Confirmation of gene copy number differences using qPCR

I selected one single-copy gene *bim1*, *mfs1* with two homologous, and *tef1* with three

homologous copies in *N. muscidifuracis* genome for qPCR validation. The Oligo 7 primer analysis software (Molecular Biology Insights Inc., Cascade, CO, USA) was used to design the primer sets targeting the three selected protein-coding genes. The primers used for *bim1* are 5'-GTAGAAGAGAATTGCTTGAATG-3' and 5'-ACTCATACTCTGATGAAGGATTT-3', the primers used for *mfs1* are 5'-TTTAGCCACAAAATTATGTCC-3' and 5'-ATGTTAAATACTTGTGCTCT-3', and the primers used for *tef1* are 5'-GCTGCTGAAAATAACAAGTCT-3' and 5'-GCTGGTACAATAACTACACCT-3'. All primers were synthesized by Eurofins Genomics LLC (Louisville, KY, USA). Before the qPCR experiments, I checked the size of PCR products and the amplification efficiency using 2% agarose gel electrophoresis. The qPCR experiments were performed using Luna® Universal qPCR Master Mix (New England BioLabs, Ipswich, MA, USA) on a Bio-Rad C1000 Touch Thermal Cycler with CFX96 Real-Time PCR Detection Systems (Bio-Rad Laboratories, Hercules, CA, USA). The 20 µL reaction system consisted of 10 µL of Luna Universal qPCR Master Mix, 0.5 µL of forward primer (10 µmol/L), 0.5 µL of reverse primer (10 µmol/L), 8 µL of nuclease-free water, and 1 µL DNA template (*M. zaraptor* DNA samples extracted from AUB colony). The qPCR reaction conditions were 95 °C for 60 s, followed by 40 cycles at 95 °C for 15 s and 50 °C for 30 s.

4.3.16 Functional and pathway annotation of *N. muscidifuracis* proteins

The pathway annotation was performed using 2,783 annotated genes in *N. muscidifuracis* and 5,886 genes in *Saccharomyces cerevisiae* (*S. cerevisiae*, accession number: GCA_002571405.2) [278]. Assignments to genes in metabolic and regulatory pathways were performed by the KEGG's internal annotation tools (https://www.kegg.jp/) [279, 280]. GhostKOALA was used to assign the most appropriate K numbers to the query genes by GHOSTX

program [281] and KOALA (KEGG Orthology And Links Annotation) algorithm [282], which is based on the sequence similarity search against the structured KEGG GENES database [283]. Subsequently, a set of K numbers was linked to KEGG pathway maps with the KEGG Mapper Reconstructed tool. The number of genes in selected KEGG pathways in *N. muscidifuracis* and *S. cerevisiae* genomes was manually counted according to the KEGG pathway maps. Statistical significance was evaluated using the Chi-squared test.

4.3.17 Motifs prediction in regulatory regions

In microsporidia, the patterns of transcript initiation are quite different. The regulatory motifs of translation start sites appear to be concealed in common eukaryotes due to the compact, gene-dense genomes [284, 285]. To discover the potential regulatory motifs of the 5′ context of *N. muscidifuracis* coding sequences, 200 bp sequences upstream of the start codon for all genes were extracted from *N. muscidifuracis* genome. MEME [286] was applied to search for novel 5′ motifs using all gene set (n = 2,718) with the maximum motif width of 12 positions. To characterize the pattern of motif in *Nosema* species, the same analysis was performed in other *Nosema* genomes and *E. cuniculi*. I used MEME to identify the motifs upstream of the start codon in a small gene set (n=449, shared orthologous genes) and other predicted genes in seven *Nosema* species and *E. cuniculi*. The RNA-seq data was analyzed to provide additional evidence for the discovery of regulatory motifs. I plotted the average coverage across the 200 bp upstream (-200 bp) and 500 bp downstream (+500 bp) region of 2,155 protein-coding genes. The presence of motifs in different expressed genes was examined to determine the relationship between predicted motifs and gene expression.

4.3.18 Phylogenetic analysis between *N. muscidifuracis* and other microsporidia

To infer the evolutionary history of my assembled genome, I estimated the phylogenetic relationship between *N. muscidifuracis* and other microsporidia. Homologous orthologs were determined for 10 strains of 7 *Nosema* species (*N. ceranae* BRL 01 [257], *N. ceranae* PA08 1199 [287], *N. ceranae* BRL [70], *N. apis* BRL 01 [288], *N. bombycis* CQ1 [72], *N. granulosis* Ou3-Ou53 [74], *N. antheraeae* YY [73], *Nosema sp.* YNPr [71], *N. muscidifuracis* Mzar and *N. muscidifuracis* Mrap), as well as the outgroup species of *Encephalitozoon cuniculi* (*E. cuniculi*) [75, 76]. The orthologs of *Nosema* genus and the orthologous proteins of *N. ceranae* PA08 1199 strain were extracted from OrthoDB v10.1 (https://www.orthodb.org/) [52] using TaxonKit [289]. The genomic data of five *Nosema* strains and *E. cuniculi* was downloaded from NCBI, the accession number was presented in Table 14. The genomic data of *N. antheraeae* YY was downloaded from SilkPathDB database [73], and protein sequence of *Nosema sp.* YNPr was provided by Dr. Xu Jinshan [71]. Then, I performed a BLASTp search with the set of *N. ceranae* (PA08 1199 strain) proteins to identify the protein sequences of orthologs in other selected microsporidia genomes and my assembled *N. muscidifuracis* genomes with a minimum of 20% sequence identity (E-value < 10e-5). A total of 449 orthologs among these eleven genomes were identified. Subsequently, MAFFT v7.407 [290] was used to align the protein sequences of orthologs among the above genomes independently. The protein alignments were concatenated into one super-sequence to construct the phylogeny with Jones-Taylor-Thornton (JTT) protein model using RAxML v8.2 [291]. 1,000 rapid bootstrap replicates were applied to evaluate the branch supports. The phylogenetic tree was finally visualized in FigTree v1.4.4 software (http://tree.bio.ed.ac.uk/software/figtree/).

4.3.19 Cytogenetic analysis of *Nosema* distribution in the ovaries of three *Muscidifurax*

species

The ovaries of infected *M. zaraptor* and *M. raptor* were compared with the closely-related *Nosema*-free species *M. uniraptor*. Ovaries were fixed for 15 minutes in 10% Formaldehyde and 0.2% Tween in Phosphate Buffered Saline (PBS-T). Actin staining (Rhodamine Phalloidin, Invitrogen) was performed overnight or longer at 4℃ in PBS-T. Ovaries were then washed 3X in PBS-T and moved to Vectashield with DAPI (Vector Labs) for at least 16 hours before mounting. Microscopy was performed using a Leica SP5 point scanning confocal (63x/1.4 HCX PL Apo CS oil lens). Images were collected with LAS AF. Minor processing (Gaussian blur) was performed using FIJI.

4.4 Results

4.4.1 *Nosema muscidifuracis* genome assembly and statistics

I sequenced the *M. zaraptor* genome using a combination of PacBio long-read and Illumina 10× Genomics linked-read sequencing technologies in *Nosema*-infected *M. zaraptor* samples (see Materials and methods). A total of 23.7 Gb PacBio Sequel II HiFi reads (61.2-fold coverage of *M. zaraptor* genome), and 63.6 Gb of Illumina linked-reads (54.5-fold coverage) were generated (Table 18), resulting in a high-quality assembly of *M. zaraptor* genome. Bioinformatic analyses revealed that symbiotic microbes are among the assembled contigs, including *Nosema muscidifuracis*, a known microsporidian species infecting *M. zaraptor*. *Nosema muscidifuracis* contigs were separated from the host contigs based on a much higher sequencing depth (258.2×; Table 19) compared to *M. zaraptor* (61.2×) and a much lower GC content (22.6% compared to 42% in *M. zaraptor*). The drastic differences in sequence coverage and GC content allowed the

complete separation of *N. muscidifuracis* contigs from host sequences (Figure 26). In addition, *N. muscidifuracis* contigs were aligned to a closely related, *Nosema*-free *M. raptorellus* genome [263] to confirm the absence of host contaminations. The final assembly of the *N. muscidifuracis* genome contains 14,397,169 bp in 28 contigs, with contig length ranging from 299,473 to 982,164 bp (Figure 27A).

Table 18. Summary PacBio long-read and Illumina (10× Genomics) linked-read sequencing data generated for *Muscidifurax zaraptor* and *Nosema muscidifuracis* genome assembly

Genome Statistics	***M. zaraptor***	***N. muscidifuracis***
Genome size (bp)	386,836,632	14,397,169
PacBio data: # of HiFi reads	1,842,231	-
PacBio: total HiFi sequences	23.7 Gbp	-
PacBio: average depth	61.2×	258.2×
Pacbio: % scaffold mapped	99.98%	99.58%
Illumina data: # of linked reads	424,317,074	-
Illumina: total sequences	63.6 Gb	-
Illumina: average depth	54.5×	226.7×
Illumina: % scaffold mapped	99.95%	99.42%

Table 19. Average PacBio coverage depth against the *Nosema muscidifuracis* genome

contig	**length (bp)**	**covered %**	**depth**
contig01	982,164	99.99%	207.709

contig02	933,025	100.00%	303.615
contig03	833,206	99.76%	281.799
contig04	682,066	100.00%	282.248
contig05	669,200	98.96%	272.678
contig06	635,296	99.83%	315.956
contig07	610,585	98.52%	206.116
contig08	588,184	99.98%	280.003
contig09	577,561	99.78%	204.839
contig10	554,543	100.00%	281.667
contig11	544,348	99.75%	328.366
contig12	522,755	99.70%	167.12
contig13	491,432	98.60%	229.028
contig14	490,971	100.00%	194.356
contig15	459,695	99.79%	331.034
contig16	442,036	99.98%	311.178
contig17	421,492	99.76%	231.984
contig18	407,860	99.99%	224.255
contig19	396,865	99.67%	260.539
contig20	396,051	98.77%	213.116
contig21	388,771	100.00%	354.814
contig22	379,414	100.00%	346.008
contig23	368,489	100.00%	176.693
contig24	350,964	99.41%	191.966
contig25	328,750	99.80%	170.164
contig26	325,008	98.41%	244.626

contig27	316,965	100.00%	292.97
contig28	299,473	100.00%	280.375

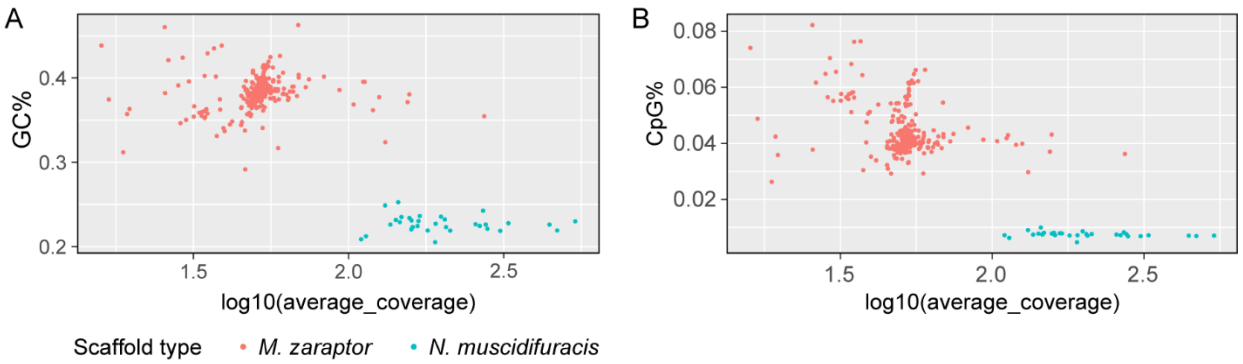


Figure 26. The plot of GC content and CpG percentage versus average coverage for all scaffolds from the initial assembly suggested that the scaffolds of *N. muscidifuracis* are separated from other scaffolds in *M. zaraptor* genome.

(A) The plot showing the GC content versus average coverage for all scaffolds from the initial assembly. The scaffolds of *N. muscidifuracis* with higher coverage and extremely low GC content are plotted using blue dots, and the scaffolds in *M. zaraptor* genome are labeled in red. (B) The plot showing the CpG percentage and average coverage for all scaffolds from the initial assembly.

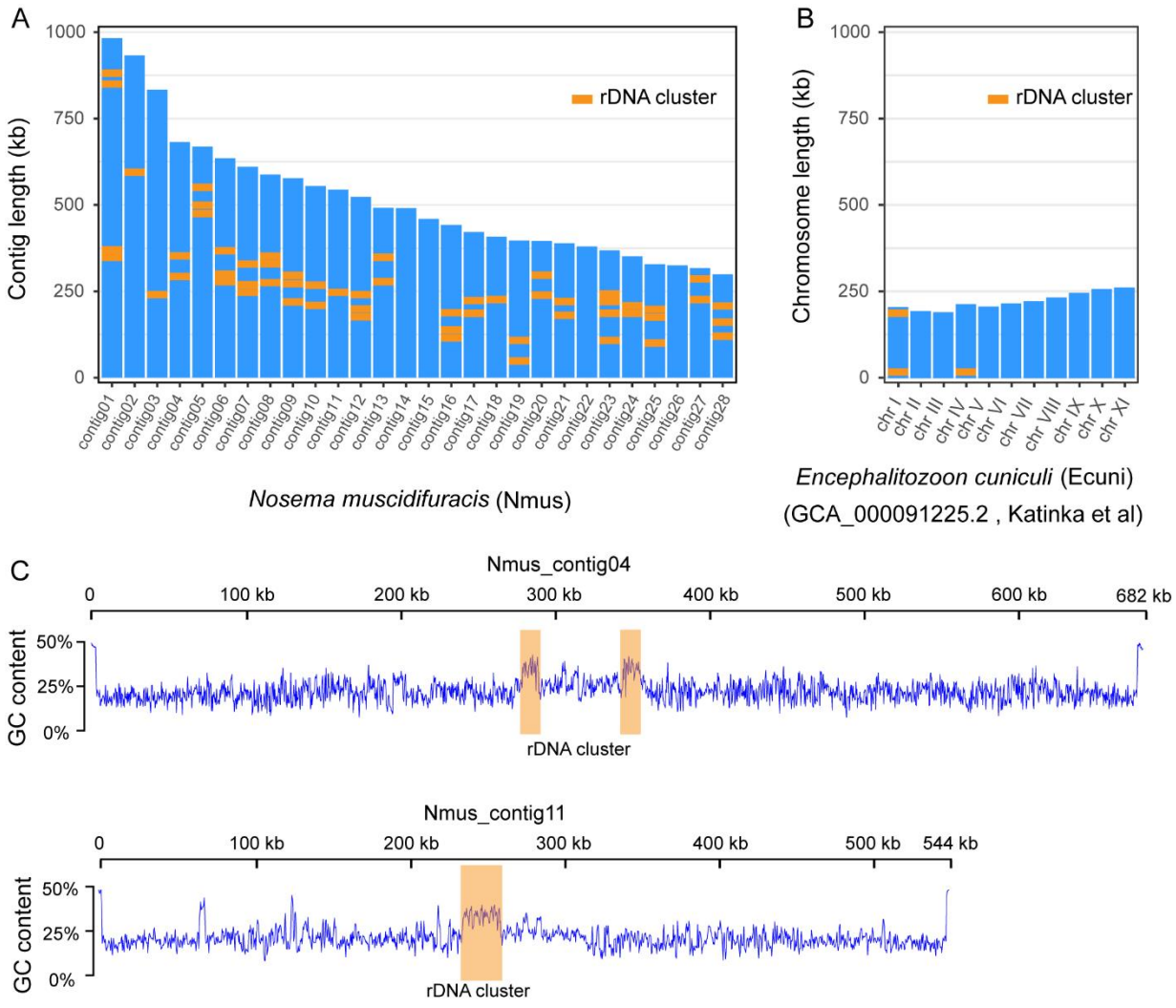


Figure 27. Chromosome level genome assembly of *Nosema muscidifuracis* showing GC content and rDNA clusters.

(A) The length of all 28 contigs in *N. muscidifuracis* genome. A total of 57 rDNA clusters are plotted using orange boxes. (B) The length of 11 chromosomes in *Encephalitozoon cuniculi* genome with three rDNA clusters labeled in orange. Data is from Katinka *et al* (2001) [75]. (C) Plots of GC content along *N. muscidifuracis* contig04 and contig11 show low average GC-content of the *N. muscidifuracis* genome and elevated GC content at rDNA clusters.

4.4.2 Assessment of the continuity and completeness of the *N. muscidifuracis* genome

The assembled *N. muscidifuracis* contigs had an N50 of 544,348 bp, which is much larger than all published *Nosema* genomes, indicating excellent continuity (Table 14). Nine contigs have regions at termini with elevated GC content (>50%), which are putative telomeric regions, suggesting chromosome-level assembly. To evaluate the completeness of the genome assembly, I aligned 10× Genomics short-read reads using BWA, and 99.42% of the assembled contigs were covered (Table 18). The BUSCO completeness score is 97.0%, which is the same as *N. ceranae* assemblies and much higher than other *Nosema* assemblies (Table 14). The assembly statistics indicate that the quality of my assembly is high in both genome continuity and completeness.

4.4.3 A novel composite 4-bp and 5-bp telomeric repeat motif in *N. muscidifuracis*

De novo telomeric repeat motifs (TRMs) prediction was performed in *Nosema* species using the TRIP pipeline [267]. With the short-read sequencing data publicly available, I identified the potential TRMs in *Nosema* genomes. The 4-bp (TAGG)n type of motif is the most abundant short tandem repeat in *Nosema*. The (TTAGG)n TRM is also found at the chromosome ends with great assembled length. Another possible telomeric repeat is (TAGGG)n. I characterized the assembled length of three types of telomeric repeat motifs (TRMs), and found a novel composite 4-bp and 5-bp telomeric repeat motif at the ends of chromosome. The combination of PacBio long-read assembly and Illumina short-read correction ensured the accuracy and completeness of genome assembly. With the high-quality *N. muscidifuracis* assembly, I examined the telomeric repeat motifs at the ends of chromosomes, and the result is consistent with the TRIP prediction. The (TAGG)n and (TTAGG)n types of TRMs were detected at 19 chromosome ends among 28 contigs (Figure 28A, Figure 28B, and Figure 28C). I then extracted the sequences of assembled

telomeric and sub-telomeric regions, the sequence alignments provided additional evidence to confirm that the telomere in *Nosema* species is a novel composite form of 4-bp and 5-bp telomeric repeat motif, and revealed a conserved ~20 Kb sub-telomeric region of the chromosomes (Figure 28C and Figure 28D). The phylogenetic tree based on 27 nucleotide sequences in ~20 Kb subtelomeric region suggested two types of conserved subtelomeric sequences in *Nosema muscidifuracis* (Figure 28E).

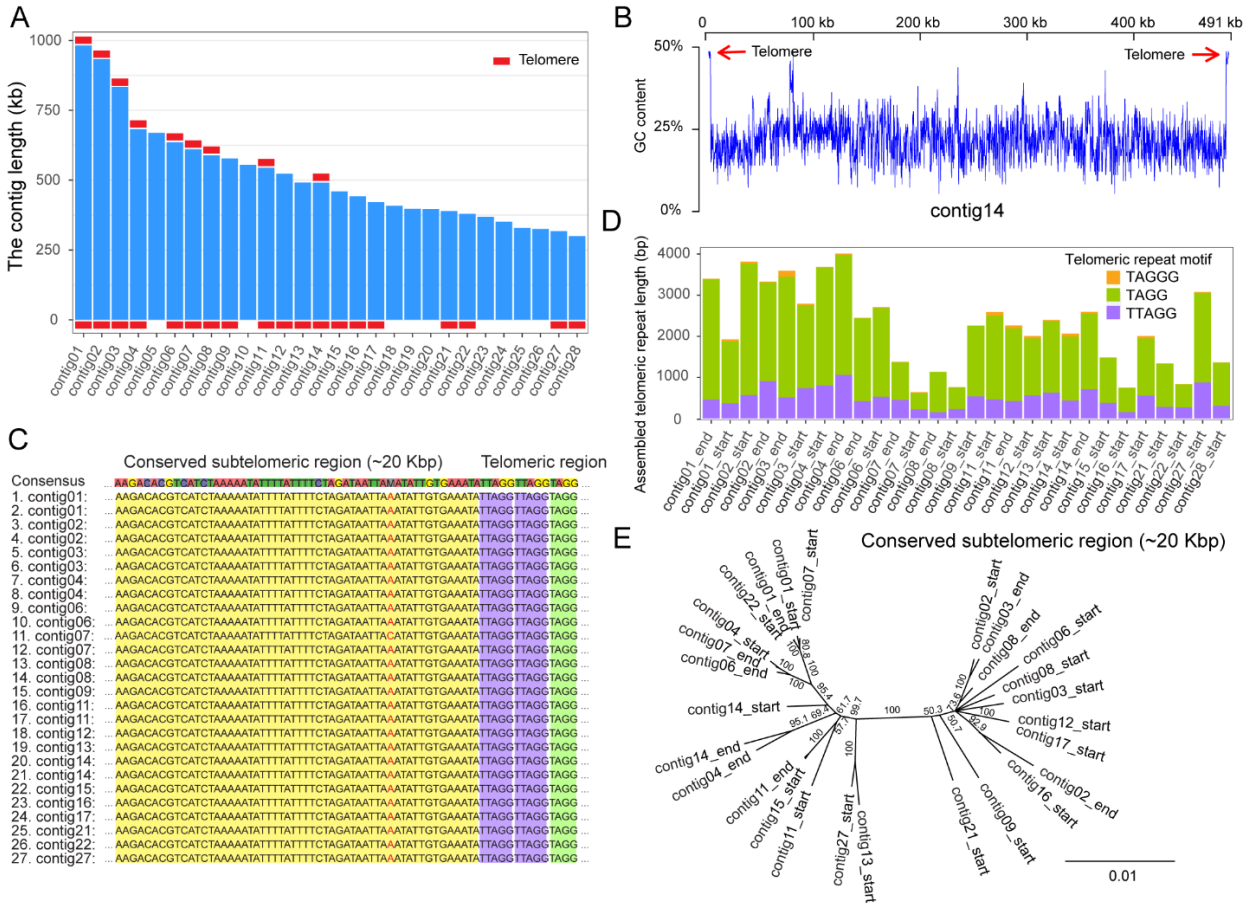


Figure 28. A novel type of telomere in the *Nosema muscidifuracis* genome.

(A) Presence of telomeric sequences at the termini of 28 *N. muscidifuracis* genome contigs. (B) Plot of GC content along contig14 showing the high GC content at telomeric regions. (C) Sequence

alignment at the telomere-subtelomere boundaries, showing the novel composite 4-bp and 5-bp telomeric repeat motif. (D) Total length and relative abundance of telomeric repeat motifs (TAGG, TTAGG, and TAGGG) in telomeric regions. (E) Phylogenetic tree of 27 subtelomeric sequences from different genomic contigs in *N. muscidifuracis*.

4.4.4 Repeat annotation

In the *N. muscidifuracis* genome, a total of 4,078,013 bp repetitive regions (28.3% of the genome) were identified. Most repeats belong to the unclassified category (58.8% of total repeats). Among classified repetitive elements, the Gypsy/DIRS1 LTR element is the most abundant, accounting for 37% of all known repeats, and 4.32% of the whole genome. The following classes of repetitive elements account for more than 1% of the *N. muscidifuracis* genome: DNA transposons (3.49%), LINEs (1.68%), and simple repeats (1.71%) (Table 20).

Table 20. Summary of annotated repeat elements in *Nosema muscidifuracis* genome

Categories	# of elements	Length (%)
Retroelements	1,008	863,391 (6%)
LINEs (RTE/Bov-B)	374	241,929 (1.68%)
LTR elements (Gypsy)	634	621,462 (4.32%)
DNA transposons	219	502,982 (3.49%)
Unclassified	1,462	2,396,435 (16.65%)
Simple repeats	3,313	246,854 (1.71%)
Low complexity	1,215	68,351 (0.47%)
Total	7,217	4,078,013

4.4.5 Noncoding RNA annotation identified 57 rDNA clusters located in the middle of *N. muscidifuracis* chromosomes.

A total of 327 noncoding RNA (ncRNA) genes were predicted and annotated in *N. muscidifuracis* genome based on the Rfam [269] database of RNA families by using the Infernal software package [268]. The 170 tRNA genes account for 0.09% of the entire genome. I also found seven snRNA associated with U2/U4/U6 small nuclear ribonucleoproteins, and two CD-box, small nucleolar RNA U3, suggesting that the splicing function may be present in *N. muscidifuracis*. The most abundant ncRNA genes in *N. muscidifuracis* are the rDNA genes. I identified 57 complete rDNA clusters encoding 18S/28S ribosomal RNAs (Table 21). *Encephalitozoon cuniculi*, a microsporidian species infecting various mammals, has 11 chromosomes, and the rDNA clusters are located in the telomeric regions [75, 76]. However, only three rDNA regions were present in the *E. cuniculi* reference genome due to the difficulties in assembling these regions with highly similarity (Figure 27B). In my *N. muscidifuracis* assembly, the rDNA clusters are clearly in the middle of the contigs/chromosomes (Figure 1A), suggesting that the chromosome termini location of rDNA clusters may be a special case only in *Encephalitozoon.* A hallmark of *N. muscidifuracis* rDNA region is increased GC context (~37%) compared to the genome average (~25%; see Figure 27C). Collectively, the rDNA clustered are more than 206 Kb in length, accounting for 1.4% of the *N. muscidifuracis* genome (Table 21).

Table 21. The annotation of noncoding RNAs in the *Nosema muscidifuracis* genome

Type	Class	# of copy	Average length (bp)	Total length (bp)	% of genome
tRNA		170	75.1	12,765	0.089%

rRNA	5S	34	119.2	4,052	0.028%
	LSU	57	2315.1	131,958	0.917%
	SSU	57	1237.1	70,516	0.490%
	Total rRNA	**148**	**1395.4**	**206,526**	**1.434%**
snRNA	CD-box	2	179.0	358	0.002%
	splicing	7	142.6	998	0.007%
	Total snRNA	**9**	**150.7**	**1,356**	**0.009%**

4.4.6 Reoccurrence of *Nosema* infection in *M. zaraptor* after cured by a combination of heat treatment and Pasteur method.

The genome assembly used *M. zaraptor* in the AUB colony, which has never been treated for *Nosema* (Figure 29A). In 2019, treatment to eliminate *Nosema* was performed at the USDA colony, using a heat incubation approach [249]. The *M. raptor* USDA colony was established in 2015 from *Nosema*-free founders based on microscopic examination [95]. The recurrence of *Nosema* infection was cured in 2017 using an extensive treatment procedure described in the Materials and Methods (summarized in Figure 29A). To determine whether the treatment has effectively eliminated or reduced the *Nosema* infection, I designed PCR primer sets (Table 16) to target the 18S rDNA genes with sequence information in my genome assembly. The *Nosema* titers in the AUB (Ct values 10.2~11.7) and USDA (Ct value 12.1~13.8) *M. zaraptor* samples were extremely high compared to the control uninfected *Muscidifurax uniraptor* samples (Ct value > 38; $P < 0.001$, t test), indicating heavy infection in both colonies (Figure 29B). The relative

abundance of *Nosema* in AUB samples is significantly higher than the USDA colony for both females (2.8-fold; $P < 0.01$, t test) and males (3.3-fold; $P < 0.001$, t test), suggesting a significant reduction of the *Nosema* load in treated wasps (Figure 29B). The results indicated that heat treatment reduced the *Nosema* titer in a short time but failed to eradicate the *Nosema*, and the infection came back and reached high titer rapidly. For the *M. raptor* samples, the relative abundance of *Nosema* (Ct values 10.8~11.7) is also high compared to the *Nosema*-free *M. uniraptor* samples ($P < 0.001$, t test), which is comparable to AUB *M. zaraptor* samples (Figure 29B). It is clear the *Nosema* has rebounded in the heat-treated line, either due to incomplete elimination or cross-infection from infected wasps.

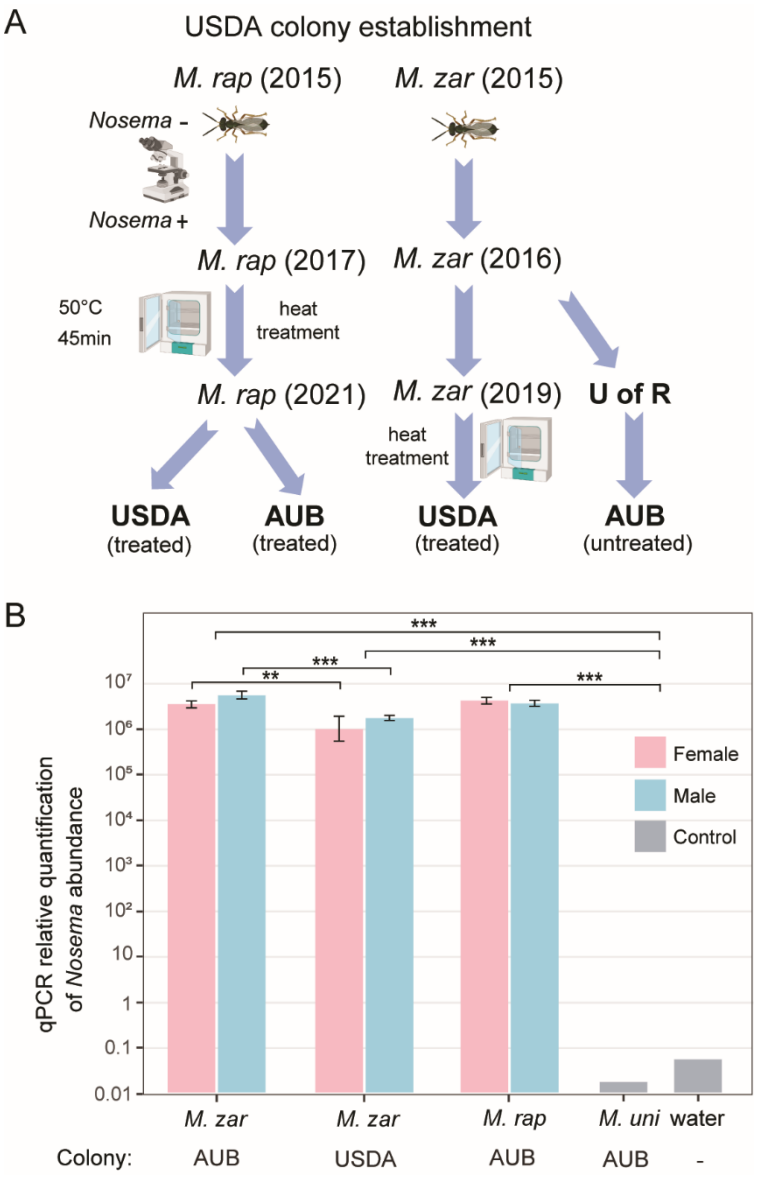


Figure 29. Quantification of the *Nosema* titer in AUB and USDA *Muscidifurax zaraptor* colony.

(A) The USDA *M. zaraptor* and *M. raptor* colonies were founded in 2015 in Dr. Geden's laboratory. The USDA *M. raptor* colony was treated using a combination of heat treatment and Pasteur method in 2017, resulting in a *Nosema*-free colony confirmed by microscopic examination. The AUB *M. zaraptor* was derived from a colony from the University of Rochester (U of R), which

was never treated. The USDA *M. zaraptor* colony was cured for *Nosema* infection using a heat treatment approach in 2019 [262]. (B) Results of qPCR quantifications of *Nosema muscidifuracis* in *M. zaraptor* and *M. raptor*. Female samples were plotted using pink bars and male samples were plotted in blue. Negative control (water) and *Nosema*-free wasp *M. uniraptor* were included as controls. Statistical significance was assessed by t test (*, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$).

4.4.7 Comparative genomic analysis of *N. muscidifuracis* with *N. ceranae* and an outgroup microsporidian *E. cuniculi*

I identified 2,782 protein-coding gene models using the Fungal Genome Annotation Pipeline (FunGAP) [270], with RNA-seq reads support. The number of protein-coding genes I annotated for *N. muscidifuracis* is close to that of *N. ceranae* (N = 2,905) [257] and *N. apis* (N = 2,764). Comparative genomic analysis revealed that 26/28 *N. muscidifuracis* contigs have syntenic regions in the *N. ceranae* genome based on protein-coding genes (Figure 30). There is a moderate level of conservation in gene order, with many genome rearrangement events (Figure 3). When an outgroup microsporidian species was compared, all 11 chromosomes in *E. cuniculi* [75] have syntenic regions in *N. muscidifuracis*, which mapped to 24 *N. muscidifuracis* contigs (Figure 31).

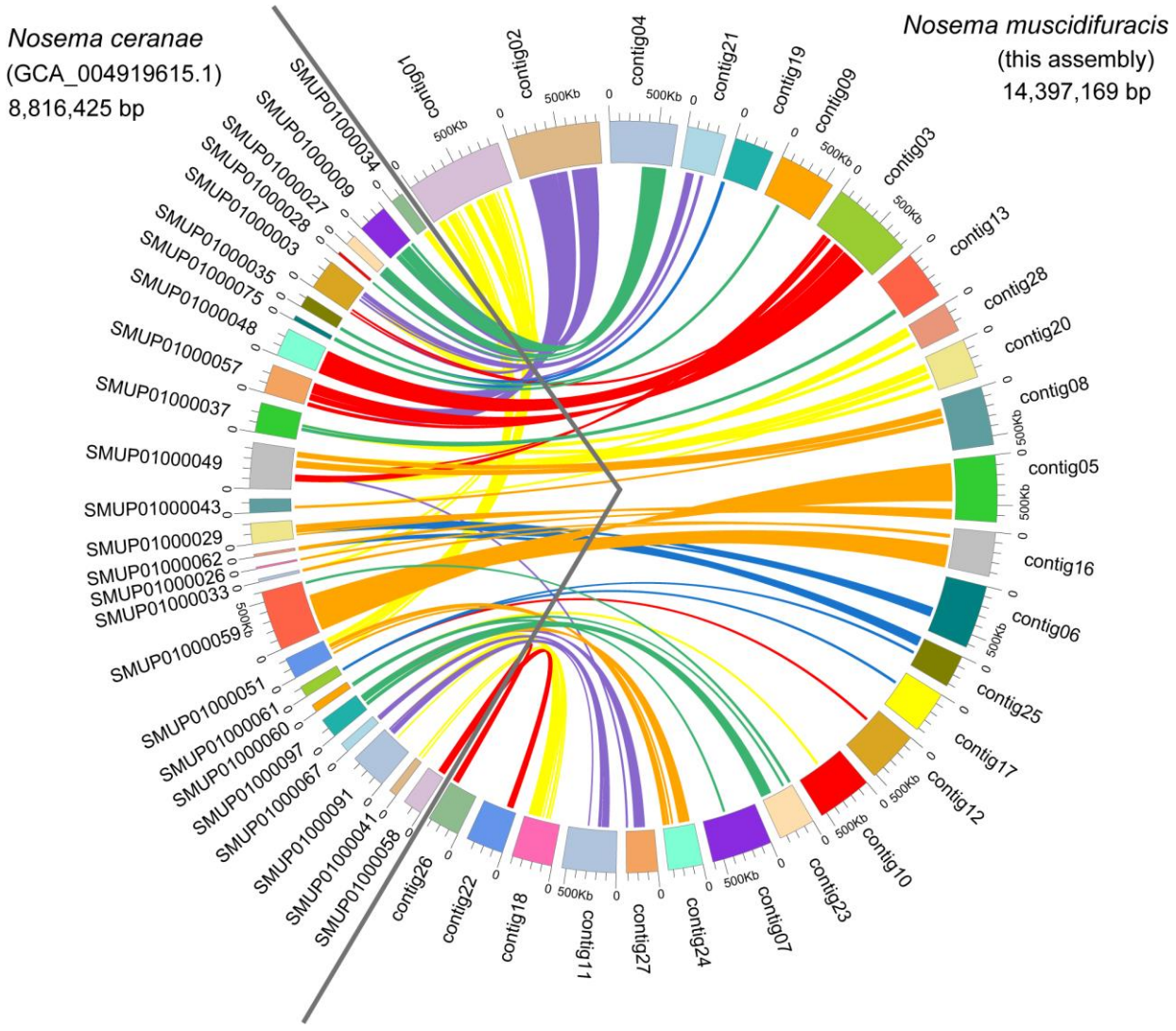


Figure 30. Genome comparisons between *Nosema muscidifuracis* and *N. ceranae*.

A total of 26 contigs of *N. muscidifuracis* (93.4% of assembly in this research) show a one-to-one relationship with 25 scaffolds in the *N. ceranae* genome (54.1% of assembly GCA_004919615). The scaffolds on the left of the circle represent *N. ceranae* scaffolds, and the contigs on the right represent *N. muscidifuracis* contigs.

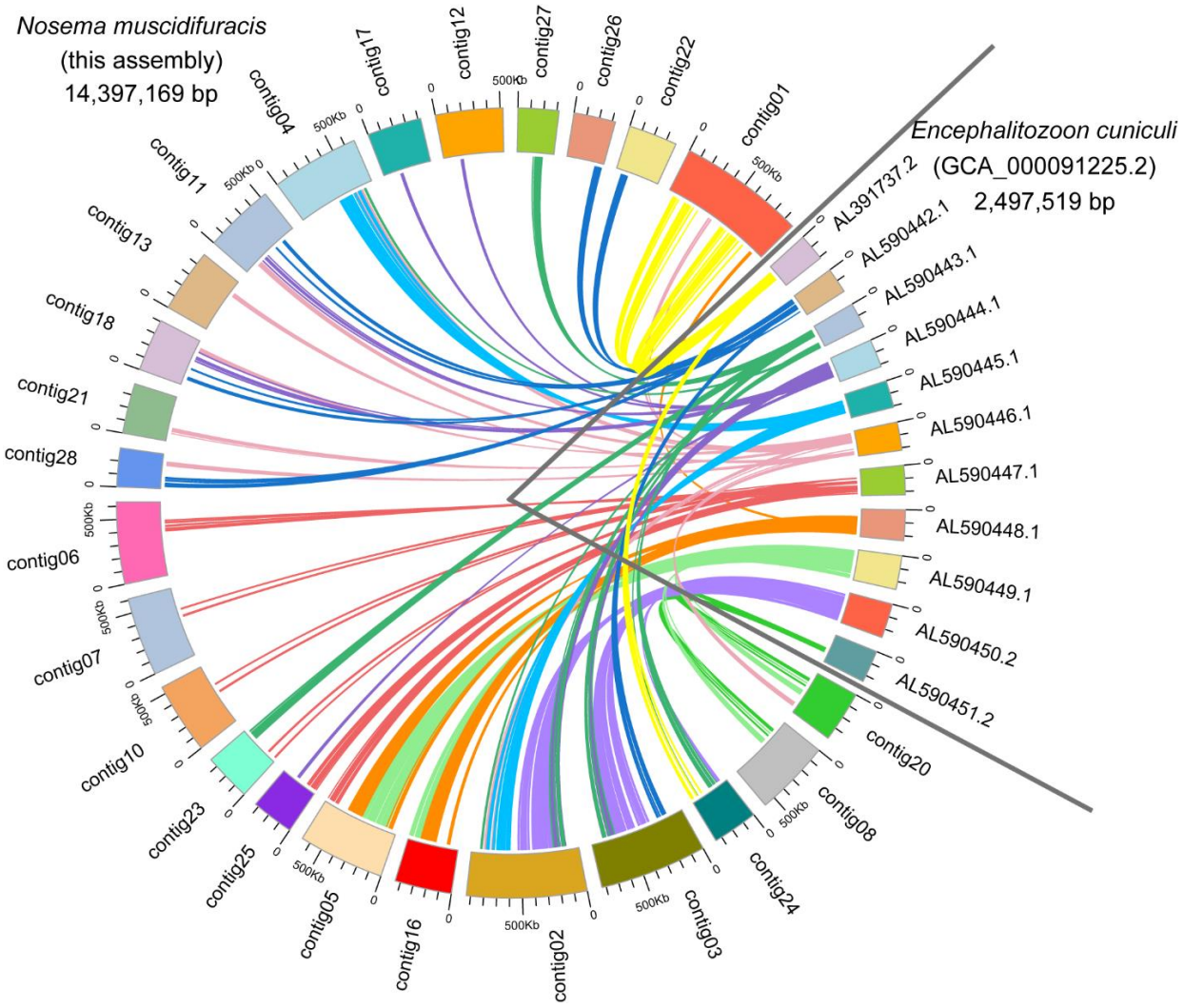


Figure 31. Genome comparisons between *Nosema muscidifuracis* and *Encephalitozoon cuniculi*.

A total of 24 contigs of *N. muscidifuracis* (86.6% of assembly in this research) show a one-to-one relationship with 11 scaffolds in the *E. cuniculi* genome (100% of assembly GCA_000091225.2). The contigs on the left of the circle represent *N. muscidifuracis* contigs, and the scaffolds on the right represent *E. cuniculi* scaffolds.

4.4.8 Genome annotation reveals extensive gene duplication events in *Nosema muscidifuracis*

In the *N. muscidifuracis* genome, the majority of genes were duplicated with two or more copies (N = 2,526 genes in 1,147 paralogous groups). In sharp contrast, only a small number of genes (N = 256) are single-copy (Figure 32A). This is not the case in *N. ceranae* or *E. cuniculi* genomes (Figure 32A). The pattern is very complex (Figure 32B), indicating extensive duplications and rearrangements within the genome. For *N. muscidifuracis* with two copies, I aligned all 919 pairs and found that 859 of them are the same length, and only 60 pairs have different gene sizes. The average pairwise identity for the 859 gene pairs with the same length is 99.7%, and that of the remaining 60 gene pairs is 94.1%. This indicates relative recent duplications in *N. muscidifuracis*. I computed the number of substitutions for the 859 gene pairs with the same size (average gene length = 1,027 bp), and on average, 1.66 synonymous and 1.09 nonsynonymous substitutions were observed (2.75 total). I next examined the syntenic pattern between homologous genes within *N. muscidifuracis*. A self-circos plot (syntenic relationship among contigs determined by homolog of gene models) did not reveal whole contig duplications, but rather extensive duplications and rearrangements across the genome (Figure 32B). The majority of the duplicated gene pairs are on different contigs, except for 66 gene pairs on contig01 (Figure 32B). Contig01 is the largest contig/chromosome (982 Kb in length), with duplicated regions on contig16, contig21, and contig24, with short to moderate stretches of synteny (Figure 32B).

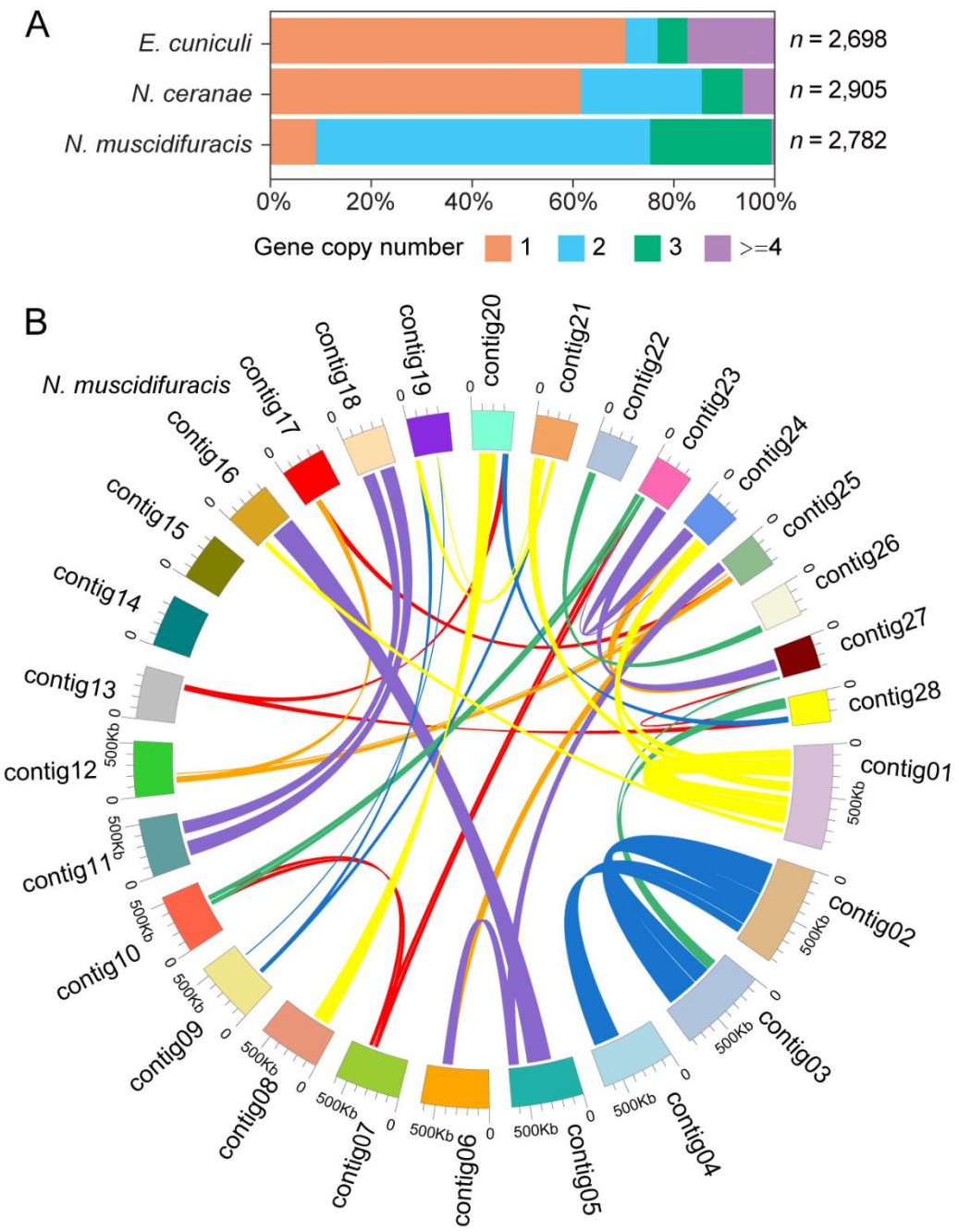

A
E. cuniculi
N. ceranae
N. muscidifuracis
n = 2,698
n = 2,905
n = 2,782
0%
20%
40%
60%
80%
100%
Gene copy number
1
2
3
>=4
B
N. muscidifuracis
contig01
contig02
contig03
contig04
contig05
contig06
contig07
contig08
contig09
contig10
contig11
contig12
contig13
contig14
contig15
contig16
contig17
contig18
contig19
contig20
contig21
contig22
contig23
contig24
contig25
contig26
contig27
contig28
500Kb

Figure 32. Histograms of gene copy number in *Nosema muscidifuracis, N. ceranae,* and *E. cuniculi* genomes and self-circos plot of 28 contigs in *N. muscidifuracis*.

(A) Proportion of annotated genes (*x*-axis) with different gene copy number in the genomes of *N. muscidifuracis, N. ceranae,* and *E. cuniculi* (*y*-axis). (B) The self-circos of *N. muscidifuracis* genome shows the linked relationship of 28 contigs.

I examined the read depth in single-copy gene regions and duplicated gene regions. There are significant differences in depth between single-copy genes and genes with two copies. For single-copy genes, the average coverage depth is 44.7 with a standard error of 1.6. In duplicated gene regions, the average coverage total depth is 72.0 with a standard error of 0.7, which is significantly higher than single-copy genes ($P < 2.2\times10^{-16}$, Mann-Whitney U test). To further confirm the gene duplication events, I quantified the relative abundance of a single-copy gene *bim1* (microtubule binding protein BIM1), a two-copy genes *mfs1* (major facilitator superfamily 1 nucleoside transporter), and a three-copy gene *tef1* (translation elongation factor 1 alpha), and compared them with the 18S rDNA gene (Figure 33). The *tef1* relative abundance is significantly higher than *mfs1* ($P < 0.05$, t test), and *mfs1* has significantly higher abundance that the one-copy *bim1* gene ($P < 0.05$, t test; Figure 33). The read depth and qPCR results confirmed the copy number differences of the *N. muscidifuracis* genes.

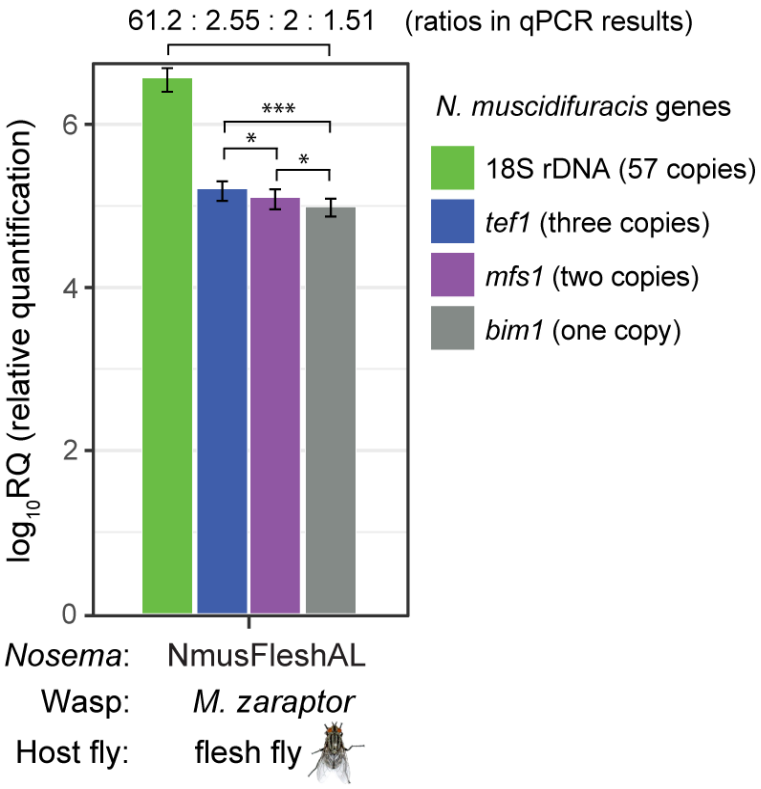


Figure 33. Quantification and confirmation of gene copy number in *Nosema muscidifuracis* using quantitative PCR approach.

Results of qPCR analysis of 18S rDNA gene (57 copies in the genome, in green), protein-coding genes *tef1* (three copies, in blue), *mfs1* (two copies, in purple), and *bim1* (single-copy, in grey). NmusFleshAL: *Nosema*-infected *Muscidifurax zaraptor* reared using flesh fly pupae at Auburn, Alabama. Statistical significance was assessed by t test (*, $P < 0.05$; ***, $P < 0.001$).

4.4.9 Severe genome reduction and lack of mitochondrial genes in *N. muscidifuracis*

Compared to the number of genes in yeast genome (approximately 6,000 genes in all) [292, 293], a total of 2,782 genes in *N. muscidifuracis* genome suggested that the *N. muscidifuracis* went through massive gene loss during the evolution. I proposed that the severe genome reduction in *N.*

muscidifuracis could be explained by a large set of gene loss. To confirm this hypothesis, I performed the functional annotation for the proteins to investigate the related KEGG pathways. In the *N. muscidifuracis* genome, gene functional annotation indicated that the genes associated with mitochondrial electron transport chains were absent (Table 22), and KEGG pathway analysis suggested that the F-type ATPase was completely missing in mitochondrial oxidative phosphorylation metabolic pathway (Table 22 and Figure 34). The results revealed the lack of mitochondrial and severe genome reduction in *Nosema muscidifuracis* genome. This will improve our understanding of gene loss and mitochondrial evolution in parasitic eukaryotes.

Table 22. The number of genes in major functional pathways identified in *Nosema muscidifuracis* and *Saccharomyces cerevisiae*

Pathway	*N. muscidifuracis* gene number	*S. cerevisiae* gene number
Oxidative phosphorylation	0	41
Endocytosis	11	50
mTOR signaling pathway	8	26
PI3K-Akt signaling pathway	7	19
Ribosome biogenesis in eukaryotes	23	56
Cellular senescence	8	19
Ubiquitin mediated proteolysis	17	39
Phagosome	6	13
Cell cycle	17	36
HIF-1 signaling pathway	6	12
Thyroid hormone signaling pathway	7	14

Protein export	8	16
Nucleotide excision repair	12	24
Proteasome	18	36
mRNA surveillance pathway	18	36
RNA degradation	15	30
Basal transcription factors	16	31
Mismatch repair	7	12
DNA replication	15	25
RNA polymerase	19	31
Tight junction	11	14
Wnt signaling pathway	8	10

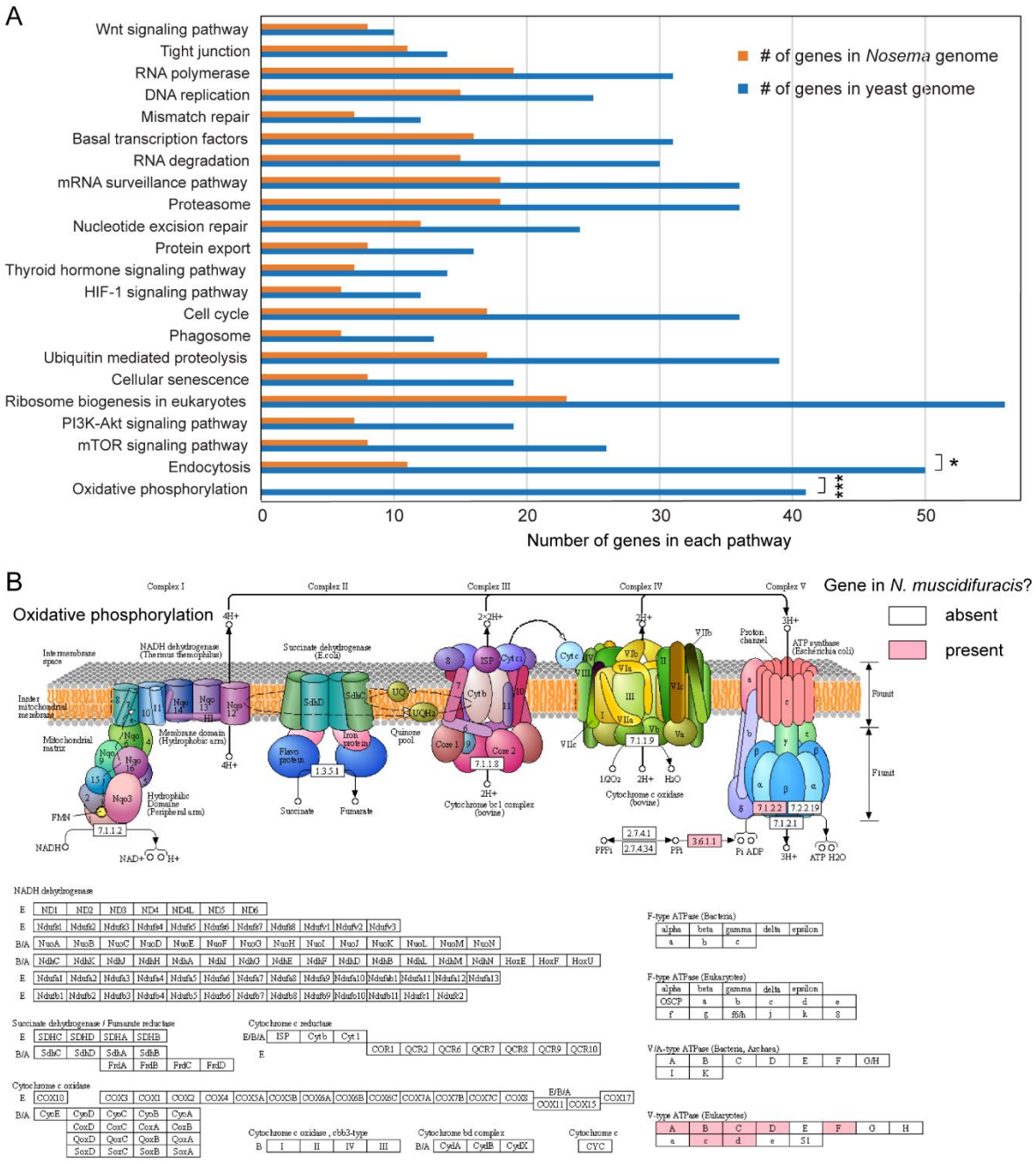


Figure 34. Functional pathway specific genome reduction in *Nosema muscidifuracis*.

(A) The number of genes enriched in 22 selected pathways in *Nosema muscidifuracis* and *Saccharomyces cerevisiae* (Chi-squared test, *, $P < 0.05$; ***, $P < 0.001$). (B) KEGG pathway

analysis of *N. muscidifuracis* mitochondria-related proteins suggested that the entire electron transport chain and eukaryotic F-type ATPase were completely missing in mitochondrial oxidative phosphorylation metabolic pathway. The enzymes/proteins that are present in *N. muscidifuracis* genome are shaded in red.

4.4.10 A conserved regulatory motif upstream of the translation start sites in *N. muscidifuracis*

A conserved sequence motif was discovered in the 200 bp region upstream of the start codon in *N. muscidifuracis*, which consists of a thymine homopolymer of seven Ts ('TTTTTTT') followed by an adenine ('A') and consecutive cytosine triplet of three Cs ('CCC') (Figure 35A). The cytosine triplet is conserved, and it's highly over-represented (E=1.6e-256) (Table 23). The sequence logo and the pattern of *N. muscidifuracis* motif are consistent with the previously published motif in *N. ceranae* [257]. To estimate the distribution of the motifs, the distance from the first nucleotide of motif to the start codon was calculated by MEME [286]. The motifs predominantly located within 20 bp upstream of the start codon, and occurred much more frequently than expected by chance (Figure 35B). Average RNA-seq coverage across 200 bp upstream (-200 bp) and 500 bp downstream (+500 bp) regions of the protein-coding genes confirmed that the predicted motif was not in the gene region (Figure 35C). The motif is present in 90% of highly expressed *N. muscidifuracis* genes with RNA-seq read depth greater than 300, in sharp contrast to ~20% in lowly expressed genes, suggesting it serves as a candidate *cis*-element for positive regulation of gene expression (Figure 35D).

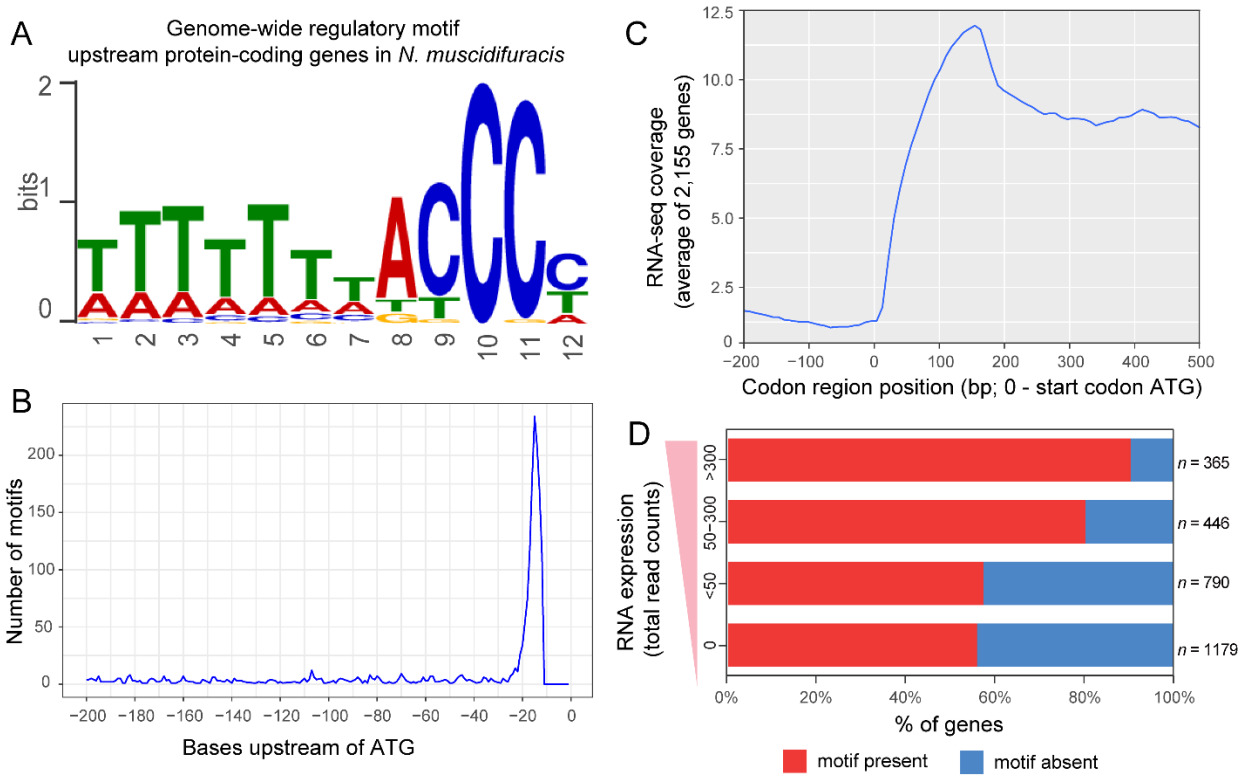


Figure 35. A motif associated with translation start sites and gene expression levels in *Nosema muscidifuracis*.

(A) A sequence motif enriched upstream *N. muscidifuracis* genes, containing a homopolymer of seven Ts followed by an A and three consecutive Cs. (B) Distribution of the motif upstream of the gene regions. The *x*-axis is the distance from the first nucleotide of the motif to the start codon in bases (5′ to 3′ with the ATG not shown at the 3′ end), and the *y*-axis is the number of detected motifs. (C) Average RNA-seq coverage across protein-coding gene regions in *N. muscidifuracis*. (D) Percentage of genes with the motif in gene groups with different expression levels.

I also searched for novel 5′ motifs in other *Nosema* genomes by MEME [286]. Compared with other predicted genes, the signal of the conserved motifs is notably significant using 449 shared orthologous genes in seven *Nosema* species and *E. cuniculi* (Table 23 and Table 24). The

comparison of the motifs in *Nosema* genomes suggested that a conserved cytosine triplet 'CCC' was found in *N. apis*, *N. ceranae*, *N. muscidifuracis*, and NosYNPr genomes. However, the motif of *N. bombycis*, *N. antheraea*, and *N. granulosis* was characterized by a cytosine tetramer 'CCCC', and followed by the thymine homopolymer (Table 23 and Table 24). The similar (C)3 and (C)4 motifs were also discovered in other distantly related *Nosema* species, suggesting a conserved *cis*-regulatory mechanism.

Table 23. Motif sequences and significance identified with 449 shared genes in *Nosema* species

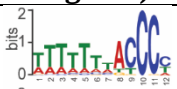
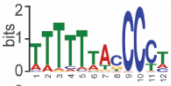
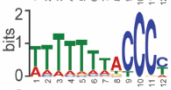
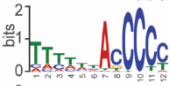
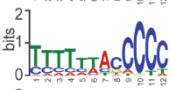
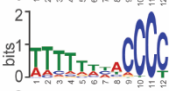
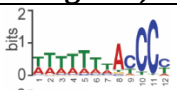
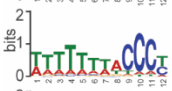
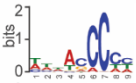
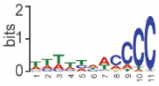
Species (strain)	Predicted motif	E-value	Sites	Motif sequence logo (449 genes)
N. muscidifuracis	TTTTTTTACCCC	1.6e-256	342	
N. apis (BRL01)	ACCCT	7.4e-94	289	
N. ceranae (BRL)	TTTTTTACCCCT	9.5e-174	227	
Nosema sp. (YNPr)	TTTTTTACCCC	2.6e-202	271	
N. antheraeae (YY)	TTTTTTACCCCC	5.4e-155	226	
N. bombycis (CQ1)	TTTTTTACCCCC	1.6e-72	203	
N. granulosis (Ou3-Ou53)	TTTTTTTACCCC	9.4e-172	273	
E. cuniculi (GB-M1)	TCTTTTTCTCCA	2.4e-18	449	

Table 24. Motif sequences and significance identified with other predicted genes in *Nosema* species

Species (strain)	Predicted motif	E-value	Sites	Motif sequence logo (other genes)
N. muscidifuracis	TTTTTTTACCCC	4.8e-155	2263	
N. apis (BRL01)	ACCCT	2.7e+175	2220	
N. ceranae (BRL)	TTTTTTACCCCT	3.7e-21	1828	
Nosema sp. (YNPr)	TTTTTTTACCCT	5.9e+217	1212	
N. antheraeae (YY)	TTTACCCCC	8.0e+108	2354	
N. bombycis (CQ1)	CCCCA	5.9e+217	3692	
N. granulosis (Ou3-Ou53)	TTTTTAACCCC	2.4e+125	2819	
E. cuniculi (GB-M1)	TCTTCTTCTCCA	1.0e-6	1706	

4.4.11 Phylogenomic analysis with other *Nosema* genomes revealed a host switch event between wasps and bees

To establish the phylogenetic relationship between *N. muscidifuracis* and other *Nosema* species in arthropods. 1,387 orthologs in *N. ceranae* PA08 1199 strain were extracted from OrthoDB v10.1 (https://www.orthodb.org/). 449 orthologous genes were identified in 10 *Nosema* strains (*N. ceranae* BRL 01, *N. ceranae* PA08 1199, *N. ceranae* BRL, *N. apis* BRL 01, *N. bombycis* CQ1, *N. granulosis* Ou3-Ou53, *N. antheraeae* YY, *Nosema sp.* YNPr, *N. muscidifuracis* Mzar and *N. muscidifuracis* Mrap) and the outgroup species *E. cuniculi* with the protein sequence identity greater than 20% and the E-value smaller than 10e-5. A maximum-likelihood phylogenetic tree of *N. muscidifuracis* and other *Nosema* species was constructed based on a concatenated alignment of the 449 proteins with the JTT model of protein evolution. *N. muscidifuracis* clustered between

two *Nosema* species *N. ceranae* and *N. apis* in honey bees (Figure 36). The *Nosema* species tree does not match the host tree, which indicates historical host switch events between wasps and bees. The newly available *N. muscidifuracis* genome will facilitate evolutionary and comparative genomic studies in the microsporidian pathogen genus *Nosema*. The sequence motif immediately upstream of the translation start site was identified in all *Nosema* species and *E. cuniculi*. The similar (C)3 and (C)4 motifs suggested a conserved *cis*-regulatory mechanism (Figure 36).

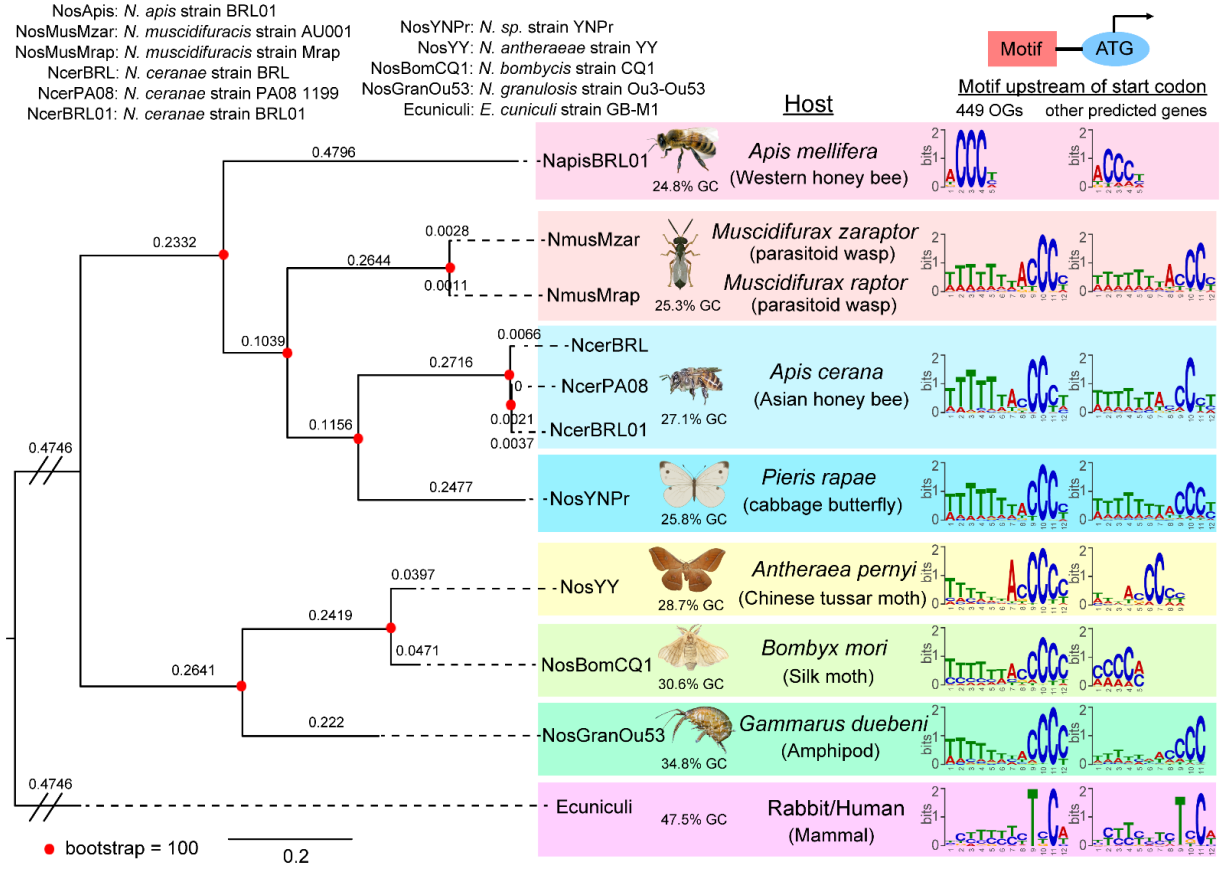


Figure 36. Phylogenomic analysis revealed a host switch event and conserved sequence motif in *Nosema*.

A maximum-likelihood phylogenetic tree of *N. muscidifuracis* with other *Nosema* species was constructed based on 449 shared proteins. The *Encephalitozoon cuniculi* (*E. cuniculi*) was used as the outgroup. The bootstrap value is labeled by dots, with red supported at 100/100. The length of each branch is shown under the branches. The sequence logos showed the conserved motifs upstream of the start codons predicted by MEME using 449 shared orthologous genes, and other predicted genes in 7 *Nosema* species and *E. cuniculi*.

4.4.12 Cytogenetics of reproductive tissues

N. muscidifuracis shows evidence of vertical transmission. I therefore investigated the ovaries of three species, *Muscidifurax zaraptor* and *M. raptor*, which are infected with *Nosema*, and the closely related species *M. uniraptor*, which is *Nosema*-free despite being reared in the same research environment. Staining with the DNA intercalating agent DAPI revealed puncta of ~2mM in diameter, much smaller than nurse cell or follicle cell nuclei, in ovarioles from *M. zaraptor* and *M. raptor* but not *M. uniraptor*. These puncta, which are consistent in appearance with previous observations of *N. muscidifuracis* [294], are associated with developing egg chambers at all stages and can be readily distinguished inside late-stage oocytes (Figure 37). Infected ovarioles also demonstrate signs of infection in other cell types, including nurse cells and follicle epithelial cells (Figure 37). This confirms heavy infections of microsporidia in the ovaries, and likely supports cytoplasmic transmission of the microbe through eggs.

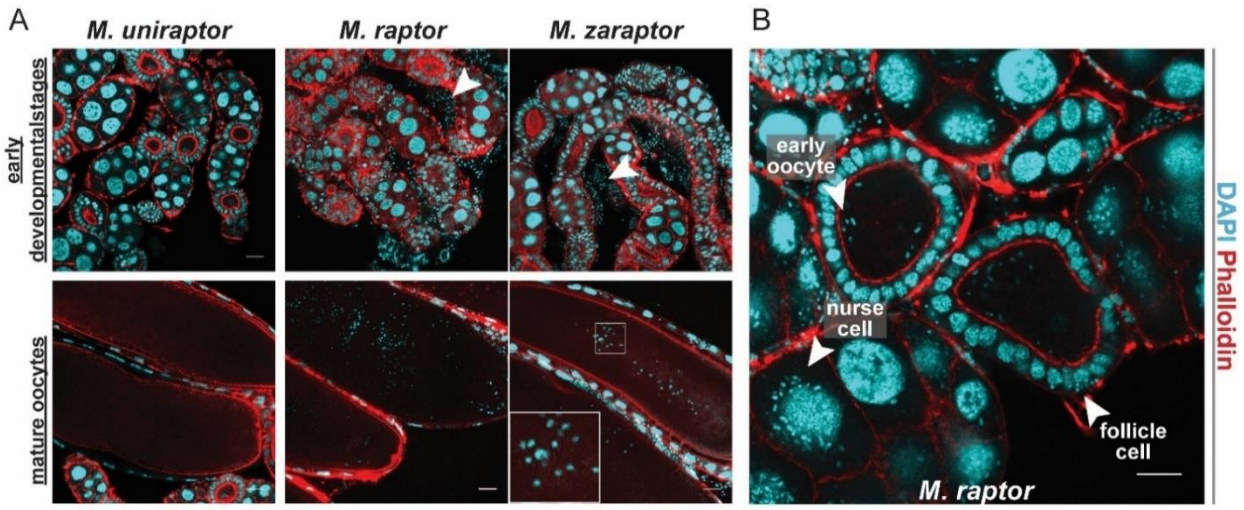


Figure 37. Cytogenetic analysis of *Nosema muscidifuracis* distribution in developing and mature oocytes of three *Muscidifurax* species.

(A) Staining of ovaries from adult female *M. uniraptor*, *M. raptor*, and *M. zaraptor* (from ***left*** to ***right***) showing early developmental stages (***top***) and mature ovaries (***bottom***). ***Red***, actin staining using Rhodamine Phalloidin. ***Blue***, DNA staining using DAPI. The length of the scale bar is 20 µm. (B) Staining showing the distribution of *N. muscidifuracis* in *M. raptor* early oocyte, follicle cells, and nurse cells (white arrows). The length of the scale bar is 20 µm.

4.5 Discussion

4.5.1 A high-quality genome assembly of *Nosema muscidifuracis*

Nosema (Microsporidia: Nosematidae) is one of the most widespread unicellular parasites belonging to the microsporidia group. Previously, microsporidia were classified as protozoan, but recently, they were recognized as a group of fungi. As a genus in microsporidia, *Nosema* infect insects and other arthropods using a specialized organelle known as the polar filament, which is coiled inside its spores. Compared to other microsporidians, *Nosema* has a much lower GC content,

which makes it very difficult to assemble its genome at high levels of completeness and continuity. In this study, I sequenced and annotated the genome of *Nosema muscidifuracis*, which is the *Nosema* species infecting the parasitoid wasp species *Muscidifurax zaraptor* and *M. raptor*. This is the first genome assembly of *N. muscidifuracis*, with the highest completeness (97.0% BUSCO score) and continuity (contig N50 = 544 Kb) among all available *Nosema* genomes (Table 14). The high-quality and well-annotated genome of *N. muscidifuracis* is crucial for the comparative genomic analysis and evolutionary study in *Nosema*, and will facilitate future genome manipulation in this fungal pathogen to control nosemosis.

4.5.2 Variation in genome size among *Nosema* species

Microsporidia usually have a severely reduced genome, such as a 2.5 Mb genome in *E. cuniculi* [75, 76]. In *Nosema*, the smallest reported genome is *Nosema sp.* YNPr (3.6 Mb), which infects the cabbage butterfly *Pieris rapae* [71]. However, the BUSCO completeness score was 84.5%, raising the question of whether sizeable numbers of genes were missed in that assembly (Table 14). The smaller genome may be partly due to incomplete genome assembly, or loss of some conserved genes in microsporidia. Four *Nosema* species have a reported genome size of 7~9 Mb, including 8.8 Mb in *N. ceranae* BRL [70], 8.6 Mb in *N. apis* [69], 7.8 Mb in *N. antheraeae* infecting the Chinese tussar moth *Antheraea pernyi* [73], and 8.6 Mb *N. granulosis* infecting the Amphipod *Gammarus duebeni* [74]. The 15.7 Mb *N. bombycis* [72] and 14.4 Mb *N. muscidifuracis* (this research) assemblies are a bit less than doubled genome size of other *Nosema* species (Table 14), indicating changes in the genome size in *Nosema*. A larger genome harbors more genes for adapting to different conditions and host environments, but this increase in capacity could also be achieved by polyploidy, which is quite common in intracellular parasites [295]. The evolution of

genome size in *Nosema* and other microsporidia warrants further study. Furthermore, little is known about the sexual cycle of *Nosema*, and how this may relate to life cycle changes in ploidy levels.

4.5.3 A novel composite form telomere in *Nosema muscidifuracis*

Previous studies have shown that the telomeric repeat motif (TRM) found in most species followed the classical form TxAyGz, like TAGG, TTAGG, TTAGGG, and TTTAGGG [296]. In this study, I predicted the potential telomeric repeat motifs using the TRIP pipeline [267] with the publicly available short-read sequencing data of *Nosema* (Data accession numbers are provided in Table 15). Two types of TRMs, (TAGG)n and (TTAGG)n take the largest proportion. Besides, my high-quality *N. muscidifuracis* assembly allows me to detect the telomeric repeat motifs at the ends of chromosomes directly, which benefited from the PacBio long-read sequencing technology. The combination of TRIP prediction and the telomeric sequence alignments revealed a novel composite 4-bp (TAGG)n and 5-bp (TTAGG)n telomeric repeat motif discovered at the ends of chromosomes in *Nosema* species. The telomeric repeat motif is highly conserved and plays an important role in protecting the chromosome termini. The novel composite form and repeat abundance of telomere in *Nosema* species provided a new insight into telomere evolution.

4.5.4 Extensive gene duplication events in *Nosema muscidifuracis*

I observed a unique feature in the *N. muscidifuracis* genome, which is the majority of protein-coding genes have two copies (66.2% of genes) or multiple copies (8.1%). One possible reason for this observation is polyploidy, which is a phenomenon that organisms possess more than two complete sets of chromosomes in individual cells [297]. However, the duplication of

genomic regions in *N. muscidifuracis* is not consistent with complete chromosome duplication (Figure 4). Based on my results, the most plausible scenario would be extensive gene duplications affecting multiple regions in the genome. In fungi, alternating haploid and diploid phases are present in many species [298], and it is not clear whether such a process exists in *Nosema*. Polyploidy has been suggested in the natural *N. ceranae* population (4N or more), based on the distribution of genetic variation [79]. A recent study of multiple fungal genomes identified that most zoosporic fungi, including microsporidia, are diplontic (diploid-dominant) based on SNP density [299]. The density of pairwise substitutions in *N. muscidifuracis* duplicated genes is 2.68×10^{-3} in this study, which would be classified as diplontic according to the criteria for diploid genomes (mean = 2.04×10^{-3}) in [299]. However, not all *N. muscidifuracis* genes are duplicated (only 66%), and my long-read assembled results do not support complete chromosome duplication. It is likely to be a complex history of gene duplication and rearrangement in *N. muscidifuracis* evolution, which expanded its genome size and resulted in the gene copy number profile I observed in this study.

4.5.5 Lack of mitochondria

Mitochondria or mitochondrion-related organelles are widely believed to exist in almost all eukaryotic organisms. Mitochondria harbor key metabolic genes and play a variety of roles in respiration, stress responses and cell metabolism. A dominant function for the mitochondria is the production of ATP [300, 301]. *Monocercomonoides* sp. is the first example of a eukaryotic microorganism absolutely lacking mitochondria. The complete absence of mitochondria is not an ancestral feature but has occurred secondarily [302]. This indicates that mitochondria are not a strictly essential organelle for the survival of eukaryotic cells [302, 303]. The mitochondrial

genome contains genes involved in basic cell metabolism. During eukaryotic evolution, the mitochondrial genome was not expected to be affected by parasitism. However, gene loss and evolution of the mitochondrial genome were observed in parasitic plants [304]. In some hemiparasitic Viscum species (Viscaceae), a great proportional reduction in the mitochondrial genome has been found. These species experienced massive gene loss in mitochondrial genes, while the remaining genes underwent very rapid evolution [305, 306].

When it comes to the mitochondrial genome reduction in eukaryotes, it's known that numerous genes have been transferred to the host nucleus. However, there is no clear pattern of mitochondrial gene loss in parasitic organisms. Compared to the yeast genome [292, 293], the number of genes has been greatly decreased in the *Nosema muscidifuracis* genome. In this study, I generated sufficient sequencing data to capture genes both from the nuclear genome and mitochondrial genome. Therefore, I can dismiss the possibility that genes are not truly lost from the mitochondrial genome but just not detected. Functional annotation and KEGG pathway analysis indicated that the F-type ATPase was completely lost in the mitochondrial oxidative phosphorylation metabolic pathway, suggesting the lack of mitochondria in the *Nosema muscidifuracis* genome (Figure 34B). These findings show that *Nosema* may be a good model for investigating how eukaryotic organisms survive with reduced or absent mitochondrial genomes, especially prior to establishing contact with their hosts. *Nosema* may shed additional light on the pattern of gene loss and the evolution of mitochondria in parasitic eukaryotes.

4.5.6 Transmission of *Nosema* in parasitoid wasps

Transovarial transmission of unicellular parasites has been observed and reported in microsporidia [294, 307-309]. In this study, I established vertical transmission of *N. muscidifuracis*

in the parasitoid wasps *Muscidifurax zaraptor* and *M. raptor* through straining experiments in the ovaries of infected females. However, whether this is the exclusive or primary mode of transmission remains to be determined. In addition, maternal transmission by injection of microsporidia during stinging of the host is also possible, which could also result in transfer between lineages when two females parasitize the same host. Additional studies are needed to characterize the transmission mechanism(s), which will be relevant to attempts to maintain *Nosema*-free colonies of these biological control agents.

4.5.7 Toward an ultimate cure for Nosemosis

Parasitoid wasps in the *Muscidifurax* genus have great potential as an alternative biological control agent that is more environment-friendly and sustainable. The production of parasitoids is affected by disease caused by persistent infection of *N. muscidifuracis*, which significantly reduces fitness and fecundity [38]. A heat shock treatment approach was claimed to be effective in controlling *Nosema* disease, resulting in a 100% cure rate and significantly improved fitness [249]. In 2019, the USDA *M. zaraptor* colony appeared to be cured by heat treatment. However, the *Nosema* bounced back to a high level within a short period of time, based on microscopic inspection, and qPCR quantification. The *Nosema* titer in the USDA *M. zaraptor* is about a third of the untreated lines (AUB), although the titer is high in both colonies. The USDA *M. raptor* colony was started from *Nosema*-free founders based on microscopic examinations. However, infection was observed after one year, and the *M. raptor* colony was treated using a series of heat treatment procedures in 2017. Parasitoids from the colony appeared to still be uninfected after three months, but infection reappeared after several years of rearing without regular inspection or further treatment for infection. These results demonstrate that visual inspection for patent

infections is an insufficiently sensitive metric for eliminating *Nosema* disease from colonies of *M. raptor*. In the qPCR results, the average Ct value was 11.23, which is orders of magnitude higher than uninfected controls (Ct value = 38; Figure 29B). Therefore, the carefully designed extensive heat treatment approach failed to eradicate *Nosema*, and the infection came back and reached high titer. The vertical transmission I showed may play a role in the rapid reoccurrence of *Nosema* because some spores could escape the treatment and be transmitted in the eggs. To combat this highly infectious parasite, molecular approaches, such as RNA interference or *Nosema* genome manipulation, need to be considered. My high-quality genome assembly and annotation serve as a first step to providing the necessary genome toolkit for these approaches, and for a better understanding of the mechanisms and etiology of Nosemosis. The findings in this study will promote the development of effective curing methods in *Nosema* infection, thereby significantly improving fitness in parasitoid wasps and enhancing their application in pest control.

Chapter 5 Conclusions and future directions

Jewel wasps in the genus of *Nasonia* are ideal models for comprehensive studies in insect genetics, genomics, epigenetics, development, and evolution. *N. vitripennis* (*Nv*) and *N. giraulti* (*Ng*) are closely related species that can be intercrossed, particularly after removal of the intracellular bacterium *Wolbachia*, serving as a powerful tool to map and positionally clone morphological, behavioral, gene expression and DNA methylation phenotypes. Wasps in the genus of *Muscidifurax* have a close evolutionary relationship to the model parasitoid genus *Nasonia*. The parasitoid wasp *Muscidifurax raptorellus* is a gregarious species with high reproductive capacity and ease of rearing. It has gained extensive attention for its potential in biological control against filth flies. Thus, it's urgent to assemble high-quality reference genomes in parasitoid wasp genera *Nasonia* and *Muscidifurax*. The intracellular bacteria *Wolbachia* are widespread and mediate many important biological processes in arthropod species. *Nasonia* has been an excellent model for intracellular bacteria *Wolbachia* research with 11 *Wolbachia* strains identified in four *Nasonia* species. The fungal symbiont *Nosema* is a diverse genus of microsporidian parasites in insects and other arthropods. *Nosema muscidifuracis* infects parasitoid wasp species of *Muscidifurax zaraptor* and *M. raptor*, causing a reduction in both longevity and fecundity.

To advance the uses of the parasitoid wasp model system, the initial goals of this project were to assemble and improve the reference genomes, as described in chapter two. I reported the first *de novo Ng* assembly using 10× Genomics linked-reads technology with a genome size of 259 Mbp and BUSCO completeness score of 98.6%. The *M. raptorellus* genome was assembled using the combination of PacBio long-read sequencing and 10× Genomics sequencing technologies with the genome size of 314 Mbp and BUSCO completeness score of 97.9%. The final genome has 226

contigs with contig N50 of 4.67 Mbp. My high-quality *Ng* and *M. raptorellus* assemblies provide reference genomes for comparative and evolutionary genomic analysis in the parasitoid wasp *Nasonia*/*Muscidifurax* model system.

In chapter three, I described the patterns of genome evolution and recombination in *Wolbachia* endosymbionts. A total of 210 conserved single-copy genes were identified in 33 genome-sequenced *Wolbachia* strains in the A–F supergroups. Phylogenomic analyses with these core genes indicate that all 33 *Wolbachia* strains maintain the supergroup relationship, which was classified previously based on the multilocus sequence typing (MLST) genes. 14 inter-supergroup recombination events (9 A-B events and 5 A-E events) were discovered using an interclade recombination screening method in six genes (2.9%) among 210 single-copy orthologs, suggesting a relatively low frequency of intergroup recombination in *Wolbachia*. The phylogenomic analysis and the identified core gene set in this study will serve as a valuable foundation for strain identification and investigating evolution and adaptations in *Wolbachia* genomes after host switching from Dipteran flies to Hymenopteran wasps.

The short generation time, easy laboratory maintenance, effective curing methods to generate uninfected controls (temporarily), and the availability of a well-assembled reference genome, make the *Muscidifurax* genus an ideal model for *Nosema* genomic research. In chapter four, I reported chromosome-level genome assembly of *N. muscidifuracis* (14.4 Mbp in 28 contigs, 97.0% BUSCO completeness score) in the parasitoid wasps *M. zaraptor* and *M. raptor* using PacBio long-read sequencing technology, and determined the genetic architecture, gene regulation, and genome evolution of the fungal pathogen genus *Nosema*. In *N. muscidifuracis*, a novel composite 4-bp (TAGG)n and 5-bp (TTAGG)n telomeric repeat motif was discovered at the ends

of chromosomes. The genome exhibits extensive gene duplications and rearrangements, with high similarity in duplicated genes. Comparative phylogenomic analyses revealed incongruency in *Nosema* and host species trees, indicating a host switch event between parasitoid wasps and bees. A highly significant ACCC motif was found within 20 bp upstream of the start codon. This motif is present in 90% of highly expressed genes, in sharp contrast to ~20% in lowly expressed genes, and similar (C)3 and (C)4 motifs were also discovered in other distantly related *Nosema* species, suggesting a conserved *cis*-regulatory mechanism. The parasitoid-*Nosema* system is laboratory tractable, and, therefore, can serve as a model to inform future genome manipulations of the *Nosema*-host system for investigations of Nosemosis.

PacBio long-read sequencing technology generates highly accurate long reads, known as HiFi reads, with base-level resolution and 99.9% single-molecule read accuracy. These HiFi reads offer sufficient overlap between individual reads, even in regions of high homology. The length and accuracy of HiFi reads simplify the analysis of whole genome sequencing data, allowing assembly software such as Hifiasm to reconstruct genomes with fewer errors and areas of uncertainty. Taking advantage of these strengths, researchers can produce complete, continuous, accurate, and phased *de novo* assemblies of complex genomes and achieve closed chromosomes even in repeat-dense and GC-rich genomes.

The genomes of the parasitoid wasp *M. raptorellus* and the fungal pathogen *N. muscidifuracis* were sequenced using PacBio long-read sequencing technology. HiFi sequencing reads enable me to generate *de novo* assemblies efficiently and affordably. I successfully achieved the first genome assembly of *M. raptorellus* in the *Muscidifurax* genus, which will serve as the reference genome for future genomic studies. Additionally, the GC-poor genome of *N.*

muscidifuracis was assembled at the chromosome level. The quality of assembled genomes is high in both completeness and continuity.

Future directions of this project involve characterizing genomes in the *Nasonia* and *Muscidifurax* genera using PacBio long-read sequencing technology to promote scientific research in the parasitoid wasp model system. For *Wolbachia* bacteria, the identification of specific genes such as *cif* gene implicated in host switching and recombination is crucial to elucidate genome evolution and adaptations in endosymbiont *Wolbachia* genomes. In *N. muscidifuracis*, it's necessary to validate the regulatory effect that the conserved motif has on gene expression. Comparative genomic and molecular evolution analyses of *Nosema* genomes in different hosts are also needed to determine candidate *Nosema* genes involved in host-pathogen interaction. Furthermore, the exploration includes the identification of additional symbionts, such as gut microbes in parasitoid wasps, and an investigation into the roles these symbionts play in the host.

To summarize, I sequenced and assembled the parasitoid wasp *N. giraulti* and *M. raptorellus* reference genomes, and the fungal pathogen *N. muscidifuracis* genome in the parasitoid wasp species *M. zaraptor* and *M. raptor* using the combination of 10× Genomics linked-reads sequencing and PacBio long-read sequencing technologies. Phylogenomic analyses of *Wolbachia* strains revealed patterns of genome evolution and recombination in *Wolbachia* endosymbionts. My dissertation work provides novel insights into the genetic architecture, gene regulation, and genome evolution of parasitoid wasps, intracellular bacteria *Wolbachia*, and fungal pathogen *Nosema*, which will build the foundation for the study of comparative genomics and host-parasite interactions in the parasitoid wasp *Nasonia* and *Muscidifurax* genera, as well as possible future biocontrol applications.

References

1. Zhu, Z., et al., Nasonia–microbiome associations: a model for evolutionary hologenomics research. Trends in Parasitology, 2022.
2. Paolucci, S., L. van de Zande, and L.W. Beukeboom, Adaptive latitudinal cline of photoperiodic diapause induction in the parasitoid Nasonia vitripennis in Europe. Journal of Evolutionary Biology, 2013. 26(4): p. 705-718.
3. Danneels, E.L., D.B. Rivers, and D.C. De Graaf, Venom proteins of the parasitoid wasp Nasonia vitripennis: recent discovery of an untapped pharmacopee. Toxins, 2010. 2(4): p. 494-516.
4. Martinson, E.O., et al., Nasonia vitripennis venom causes targeted gene expression changes in its fly host. Mol Ecol, 2014. 23(23): p. 5918-30.
5. Mrinalini, et al., Parasitoid venom induces metabolic cascades in fly hosts. Metabolomics, 2015. 11: p. 350-366.
6. Darling, D.C. and J.H. Werren, Biosystematics of Nasonia (Hymenoptera, Pteromalidae) - 2 New Species Reared from Birds Nests in North-America. Annals of the Entomological Society of America, 1990. 83(3): p. 352-370.
7. Raychoudhury, R., et al., Behavioral and genetic characteristics of a new species of Nasonia. Heredity, 2010. 104(3): p. 278-288.
8. Whiting, A.R., Biology of Parasitic Wasp Mormoniella Vitripennis [=Nasonia Brevicornis] (Walker). Quarterly Review of Biology, 1967. 42(3): p. 333-&.
9. Werren, J.H., et al., Functional and evolutionary insights from the genomes of three parasitoid Nasonia species. Science, 2010. 327(5963): p. 343-348.
10. Lynch, J.A., The expanding genetic toolbox of the wasp Nasonia vitripennis and its relatives. Genetics, 2015. 199(4): p. 897-904.
11. Werren, J.H. and D.W. Loehlin, The parasitoid wasp Nasonia: an emerging model system with haploid male genetics. Cold Spring Harb Protoc, 2009. 2009(10): p. pdb emo134.
12. Breeuwer, J.A.J. and J.H. Werren, Microorganisms Associated with Chromosome Destruction and Reproductive Isolation between 2 Insect Species. Nature, 1990. 346(6284): p. 558-560.

13. Bordenstein, S.R., F.P. O'Hara, and J.H. Werren, Wolbachia-induced incompatibility precedes other hybrid incompatibilities in Nasonia. Nature, 2001. 409(6821): p. 707-710.
14. Jin, B. and K.D. Robertson, DNA methyltransferases, DNA damage repair, and cancer. Adv Exp Med Biol, 2013. 754: p. 3-29.
15. Wang, X., et al., Function and Evolution of DNA Methylation in Nasonia vitripennis. Plos Genetics, 2013. 9(10).
16. Wang, X., J.H. Werren, and A.G. Clark, Genetic and epigenetic architecture of sex-biased expression in the jewel wasps Nasonia vitripennis and giraulti. Proceedings of the National Academy of Sciences of the United States of America, 2015. 112(27): p. E3545-E3554.
17. Wang, X., J.H. Werren, and A.G. Clark, Allele-Specific Transcriptome and Methylome Analysis Reveals Stable Inheritance and Cis-Regulation of DNA Methylation in Nasonia. Plos Biology, 2016. 14(7).
18. Werren, J.H. and D.W. Loehlin, The parasitoid wasp Nasonia: an emerging model system with haploid male genetics. Cold Spring Harbor Protocols, 2009. 2009(10): p. pdb. emo134.
19. Whiting, A.R., The biology of the parasitic wasp Mormoniella vitripennis [= Nasonia brevicornis](Walker). The Quarterly Review of Biology, 1967. 42(3): p. 333-406.
20. Beukeboom, L. and C. Desplan, Nasonia. Current Biology, 2003. 13(22): p. R860.
21. Brooks, A.W., et al., Phylosymbiosis: relationships and functional effects of microbial communities across host evolutionary history. PLoS biology, 2016. 14(11): p. e2000225.
22. Bell, K. and S.R. Bordenstein, A margulian view of symbiosis and speciation: the Nasonia wasp system. Symbiosis, 2022. 87(1): p. 3-10.
23. Girault, A. and G.E. Sanders, The chalcidoid parasites of the common house or typhoid fly (Musca domestica Linn.) and its allies. Psyche, 1910. 17(1): p. 9-28.
24. Kogan, M. and E. Legner, A BIOSYSTEMATIC REVISION OF THE GENUS MUSCIDIFURAX (HYMENOPTERA: PTEROMALIDAE) WITH DESCRIPTIONS OF FOUR NEW SPECIES1. The Canadian Entomologist, 1970. 102(10): p. 1268-1290.
25. Taylor, D.B., et al., Mitochondrial DNA variation among Muscidifurax spp.(Hymenoptera: Pteromalidae), pupal parasitoids of filth flies (Diptera). Annals of the Entomological Society of America, 1997. 90(6): p. 814-824.

26. Xiao, H., S.Y. Zhou, and Y.F. Tong, A taxonomic study of Muscidifurax Girault & Sanders from China (Hymenoptera, Chalcidoidea, Pteromalidae). Zookeys, 2018(776): p. 91-103.
27. Geden, C.J. and R.D. Moon, Host ranges of gregarious muscoid fly parasitoids: Muscidifurax raptorellus (Hymenoptera: Pteromalidae), Tachinaephagus zealandicus (Hymenoptera: Encyrtidae), and Trichopria nigra (Hymenoptera: Diapriidae). Environ Entomol, 2009. 38(3): p. 700-7.
28. Seidl, S.E. and B. King, Sex-Ratio Manipulation by the Parasitoid Wasp Muscidifurax Raptor in Response to Host Size. Evolution, 1993. 47(6): p. 1876-1882.
29. Legner, E., Reproductive isolation and size variation in the Muscidifurax raptor complex. Annals of the Entomological Society of America, 1969. 62(2): p. 382-385.
30. Petersen, J. and D. Currey, Reproduction and development ofMuscidifurax raptorellus (Hymenoptera: Pteromalidae), a parasite offilth flies. J. Agric. Entomol, 1996. 13(2).
31. McKay, T. and A.B. Broce, Discrimination of Self-Parasitized Hosts by the Pupal Parasitoid Muscidifurax zaraptor (Hymenoptera: Pteromalidae). Annals of the Entomological Society of America, 2004. 97(3): p. 592-599.
32. Newton, I.L.G., et al., Comparative genomics of two closely related Wolbachia with different reproductive effects on hosts. Genome Biology and Evolution, 2016. 8(5): p. 1526-1542.
33. Zchori-Fein, E., Y. Gottlieb, and M. Coll, Wolbachia density and host fitness components in Muscidifurax uniraptor (Hymenoptera: pteromalidae). J Invertebr Pathol, 2000. 75(4): p. 267-72.
34. Geden, C.J. and J.A. Hogsette, Suppression of house flies (Diptera: Muscidae) in Florida poultry houses by sustained releases of Muscidifurax raptorellus and Spalangia cameroni (Hymenoptera: Pteromalidae). Environmental Entomology, 2006. 35(1): p. 75-82.
35. Heraty, J., Parasitoid Biodiversity and Insect Pest Management, in Insect Biodiversity. 2009. p. 445-462.
36. Martinson, E.O., et al., The Evolution of Venom by Co-option of Single-Copy Genes. Curr Biol, 2017. 27(13): p. 2007-2013 e8.
37. Kaur, R., et al., Living in the endosymbiotic world of Wolbachia: A centennial review. Cell Host & Microbe, 2021. 29(6): p. 879-893.

38. Geden, C.J., et al., Nosema disease of the parasitoid Muscidifurax raptor (Hymenoptera: Pteromalidae): prevalence, patterns of transmission, management, and impact. Biological Control, 1995. 5(4): p. 607-614.
39. Wang, S., et al., Assessment of Water Mobility in Surf Clam and Soy Protein System during Gelation Using LF-NMR Technique. Foods, 2020. 9(2): p. 213.
40. Dittmer, J. and R.M. Brucker, When your host shuts down: larval diapause impacts host-microbiome interactions in Nasonia vitripennis. Microbiome, 2021. 9(1): p. 1-19.
41. Hoffmann, A., Wolbachia. Current Biology, 2020. 30(19): p. R1113-R1114.
42. Mukai, A., et al., Juvenile hormone as a causal factor for maternal regulation of diapause in a wasp. Insect Biochemistry and Molecular Biology, 2022. 144: p. 103758.
43. Fenn, K. and M. Blaxter, Wolbachia genomes: revealing the biology of parasitism and mutualism. Trends in Parasitology, 2006. 22(2): p. 60-65.
44. Werren, J.H., L. Baldo, and M.E. Clark, Wolbachia: master manipulators of invertebrate biology. Nature Reviews Microbiology, 2008. 6(10): p. 741-751.
45. Werren, J.H., Biology of Wolbachia. Annu Rev Entomol, 1997. 42: p. 587-609.
46. Hilgenboecker, K., et al., How many species are infected with Wolbachia?--A statistical analysis of current data. FEMS Microbiol Lett, 2008. 281(2): p. 215-20.
47. Zug, R. and P. Hammerstein, Still a host of hosts for Wolbachia: analysis of recent data suggests that 40% of terrestrial arthropod species are infected. PLoS One, 2012. 7(6): p. e38544.
48. Bailly-Bechet, M., et al., How Long Does Wolbachia Remain on Board? Mol Biol Evol, 2017. 34(5): p. 1183-1193.
49. Klopfstein, S., et al., Wolbachia infections in Australian ichneumonid parasitoid wasps (Hymenoptera: Ichneumonidae): evidence for adherence to the global equilibrium hypothesis. Biological Journal of the Linnean Society, 2018. 123(3): p. 518-534.
50. Werren, J.H. and D.M. Windsor, Wolbachia infection frequencies in insects: evidence of a global equilibrium? Proc Biol Sci, 2000. 267(1450): p. 1277-85.
51. Stouthamer, R., J.A.J. Breeuwer, and G.D.D. Hurst, Wolbachia pipientis: Microbial manipulator of arthropod reproduction. Annual Review of Microbiology, 1999. 53: p. 71-102.

52. Breeuwer, J. and J.H. Werren, Cytoplasmic incompatibility and bacterial density in Nasonia vitripennis. Genetics, 1993. 135(2): p. 565-574.
53. Hedges, L.M., et al., Wolbachia and virus protection in insects. Science, 2008. 322(5902): p. 702-702.
54. Hosokawa, T., et al., Wolbachia as a bacteriocyte-associated nutritional mutualist. Proceedings of the National Academy of Sciences of the United States of America, 2010. 107(2): p. 769-774.
55. Heath, B.D., et al., Horizontal transfer of Wolbachia between phylogenetically distant insect species by a naturally occurring mechanism. Current Biology, 1999. 9(6): p. 313-316.
56. Werren, J.H. and Z.L.R.G. Wan, Evolution and Phylogeny of Wolbachia: Reproductive Parasites of Arthropods. Proceedings Biological Sciences, 1995. 261(1360): p. 55-63.
57. Raychoudhury, R., et al., Modes of Acquisition of Wolbachia: Horizontal Transfer, Hybrid Introgression, and Codivergence in the Nasonia Species Complex. Evolution, 2009. 63(1): p. 165-183.
58. Baldo, L. and J.H. Werren, Revisiting Wolbachia supergroup typing based on WSP: spurious lineages and discordance with MLST. Current microbiology, 2007. 55(1): p. 81-87.
59. Baldo, L., et al., Multilocus sequence typing system for the endosymbiont Wolbachia pipientis. Appl Environ Microbiol, 2006. 72(11): p. 7098-110.
60. Bordenstein, S.R., J.J. Uy, and J.H. Werren, Host genotype determines cytoplasmic incompatibility type in the haplodiploid genus Nasonia. Genetics, 2003. 164(1): p. 223-233.
61. Perrot-Minnot, M.-J., L.R. Guo, and J.H. Werren, Single and double infections with Wolbachia in the parasitic wasp Nasonia vitripennis effects on compatibility. Genetics, 1996. 143(2): p. 961-972.
62. Raychoudhury, R., et al., Modes of acquisition of Wolbachia: horizontal transfer, hybrid introgression, and codivergence in the Nasonia species complex. Evolution, 2009. 63(1): p. 165-183.

63. Didier, E.S., K.F. Snowden, and J.A. Shadduck, Biology of microsporidian species infecting mammals. Advances in parasitology, 1998. 40: p. 283-320.
64. Becnel, J.J. and T.G. Andreadis, Microsporidia in insects. The microsporidia and microsporidiosis, 1999: p. 447-501.
65. Williams, B.A., et al., A high frequency of overlapping gene expression in compacted eukaryotic genomes. Proceedings of the National Academy of Sciences, 2005. 102(31): p. 10936-10941.
66. Corradi, N., et al., The complete sequence of the smallest known nuclear genome from the microsporidian Encephalitozoon intestinalis. Nature communications, 2010. 1(1): p. 1-7.
67. Keeling, P.J. and N.M. Fast, Microsporidia: biology and evolution of highly reduced intracellular parasites. Annual Reviews in Microbiology, 2002. 56(1): p. 93-116.
68. Keeling, P.J., et al., The reduced genome of the parasitic microsporidian Enterocytozoon bieneusi lacks genes for core carbon metabolism. Genome Biology and Evolution, 2010. 2: p. 304-309.
69. Chen, Y.p., et al., Genome sequencing and comparative genomics of honey bee microsporidia, Nosema apis reveal novel insights into host-parasite interactions. BMC genomics, 2013. 14(1): p. 1-16.
70. Huang, Q., et al., Genome and evolutionary analysis of Nosema ceranae: a microsporidian parasite of honey bees. Frontiers in microbiology, 2021. 12: p. 1303.
71. Xu, J., et al., The genome of Nosema sp. isolate YNPr: a comparative analysis of genome evolution within the Nosema/Vairimorpha clade. PLoS One, 2016. 11(9): p. e0162336.
72. Pan, G., et al., Comparative genomics of parasitic silkworm microsporidia reveal an association between genome expansion and host adaptation. BMC genomics, 2013. 14(1): p. 1-14.
73. Li, T., et al., SilkPathDB: a comprehensive resource for the study of silkworm pathogens. Database, 2017. 2017.
74. Cormier, A., et al., Comparative genomics of strictly vertically transmitted, feminizing microsporidia endosymbionts of amphipod crustaceans. Genome Biology and Evolution, 2021. 13(1): p. evaa245.

75. Katinka, M.D., et al., Genome sequence and gene compaction of the eukaryote parasite Encephalitozoon cuniculi. Nature, 2001. 414(6862): p. 450-453.
76. Peyretaillade, E., et al., Identification of transcriptional signals in Encephalitozoon cuniculi widespread among Microsporidia phylum: support for accurate structural genome annotation. BMC genomics, 2009. 10(1): p. 1-13.
77. Fries, I., et al., Nosema ceranae n. sp.(Microspora, Nosematidae), morphological and molecular characterization of a microsporidian parasite of the Asian honey bee Apis cerana (Hymenoptera, Apidae). European Journal of Protistology, 1996. 32(3): p. 356-365.
78. Stanimirović, Z., et al., Looking for the causes of and solutions to the issue of honey bee colony losses. Acta Veterinaria, 2019. 69(1): p. 1-31.
79. Pelin, A., et al., Genome analyses suggest the presence of polyploidy and recent human-driven expansions in eight global populations of the honeybee pathogen Nosema ceranae. Environmental Microbiology, 2015. 17(11): p. 4443-4458.
80. Higes, M., et al., Honeybee colony collapse due to Nosema ceranae in professional apiaries. Environmental Microbiology Reports, 2009. 1(2): p. 110-113.
81. Rinderer, T.E. and H.A. Sylvester, Variation in response to Nosema apis, longevity, and hoarding behavior in a free-mating population of the honey bee. Annals of the Entomological Society of America, 1978. 71(3): p. 372-374.
82. Malone, L., H. Giacon, and M. Newton, Comparison of the responses of some New Zealand and Australian honey bees (Apis mellifera L) to Nosema apis Z. Apidologie, 1995. 26(6): p. 495-502.
83. Chen, Y., et al., Asymmetrical coexistence of Nosema ceranae and Nosema apis in honey bees. Journal of invertebrate pathology, 2009. 101(3): p. 204-209.
84. Antúnez, K., et al., Immune suppression in the honey bee (Apis mellifera) following infection by Nosema ceranae (Microsporidia). Environmental microbiology, 2009. 11(9): p. 2284-2290.
85. Dussaubat, C., et al., Gut pathology and responses to the microsporidium Nosema ceranae in the honey bee Apis mellifera. PloS one, 2012. 7(5): p. e37017.

86. Costa, C., et al., Negative correlation between Nosema ceranae spore loads and deformed wing virus infection levels in adult honey bee workers. Journal of invertebrate pathology, 2011. 108(3): p. 224-225.
87. Gómez-Moracho, T., et al., Artificial diets modulate infection rates by nosema ceranae in bumblebees. Microorganisms, 2021. 9(1): p. 158.
88. Cox-Foster, D.L., et al., A metagenomic survey of microbes in honey bee colony collapse disorder. Science, 2007. 318(5848): p. 283-287.
89. Forsgren, E. and I. Fries, Comparative virulence of Nosema ceranae and Nosema apis in individual European honey bees. Veterinary parasitology, 2010. 170(3-4): p. 212-217.
90. Botías, C., et al., Nosema spp. infection and its negative effects on honey bees (Apis mellifera iberiensis) at the colony level. Veterinary research, 2013. 44(1): p. 1-15.
91. Fries, I., G. Ekbohm, and E. Villumstad, Nosema apis, sampling techniques and honey yield. Journal of Apicultural Research, 1984. 23(2): p. 102-105.
92. Fries, I., Comb replacement and Nosema disease (Nosema apis Z.) in honey bee colonies. Apidologie, 1988. 19(4): p. 343-354.
93. Anderson, D.L. and H. Giacon, Reduced pollen collection by honey bee (Hymenoptera: Apidae) colonies infected with Nosema apis and sacbrood virus. Journal of Economic Entomology, 1992. 85(1): p. 47-51.
94. Farrar, C., Nosema losses in package bees as related to queen supersedure and honey yields. Journal of economic entomology, 2014. 40(3): p. 333-338.
95. Zchori-Fein, E., C.J. Geden, and D.A. Rutz, Microsporidioses of Muscidifurax raptor (Hymenoptera: Pteromalidae) and other pteromalid parasitoids of muscoid flies. Journal of Invertebrate Pathology, 1992. 60(3): p. 292-298.
96. Becnel, J.J. and C. Geden, Description of a new species of microsporidia from Muscidifurax raptor (Hymenoptera: Pteromalidae), a pupal parasitoid of muscoid flies. Journal of Eukaryotic Microbiology, 1994. 41(3): p. 236-243.
97. Chen, Y.P., et al., Morphological, molecular, and phylogenetic characterization of Nosema ceranae, a microsporidian parasite isolated from the European honey bee, Apis mellifera. Journal of Eukaryotic Microbiology, 2009. 56(2): p. 142-147.

98. Werren, J.H., et al., Functional and evolutionary insights from the genomes of three parasitoid Nasonia species (vol 327, pg 343, 2010). Science, 2010. 327(5973): p. 1577-1577.
99. Hoedjes, K.M., et al., Introgression study reveals two quantitative trait loci involved in interspecific variation in memory retention among Nasonia wasp species. Heredity (Edinb), 2014. 113(6): p. 542-50.
100. Raychoudhury, R., et al., Behavioral and genetic characteristics of a new species of Nasonia. Heredity (Edinb), 2010. 104(3): p. 278-88.
101. Loehlin, D.W. and J.H. Werren, Evolution of shape by multiple regulatory changes to a growth gene. Science, 2012. 335(6071): p. 943-7.
102. Niehuis, O., et al., Behavioural and genetic analyses of Nasonia shed light on the evolution of sex pheromones. Nature, 2013. 494(7437): p. 345-8.
103. Verhulst, E.C., L.W. Beukeboom, and L. van de Zande, Maternal control of haplodiploid sex determination in the wasp Nasonia. Science, 2010. 328(5978): p. 620-3.
104. Wang, X., J.H. Werren, and A.G. Clark, Genetic and epigenetic architecture of sex-biased expression in the jewel wasps Nasonia vitripennis and giraulti. Proceedings of the National Academy of Sciences, 2015. 112(27): p. E3545-E3554.
105. Wang, X., J.H. Werren, and A.G. Clark, Allele-specific transcriptome and methylome analysis reveals stable inheritance and cis-regulation of DNA methylation in Nasonia. PLoS biology, 2016. 14(7): p. e1002500.
106. Rago, A., J.H. Werren, and J.K. Colbourne, Sex biased expression and co-expression networks in development, using the hymenopteran Nasonia vitripennis. PLoS genetics, 2020. 16(1): p. e1008518.
107. Martinson, E.O., et al., The evolution of venom by co-option of single-copy genes. Current Biology, 2017. 27(13): p. 2007-2013. e8.
108. Wang, X., et al., Genome report: Whole genome sequence and annotation of the parasitoid jewel wasp Nasonia giraulti laboratory strain RV2X [u]. G3: Genes, Genomes, Genetics, 2020. 10(8): p. 2565-2572.
109. Leung, K., et al., Next-generation biological control: the need for integrating genetics and genomics. Biol Rev Camb Philos Soc, 2020. 95(6): p. 1838-1854.

110. Legner, E.F., E.C. Bay, and E.B. White, Activity of Parasites from Diptera: Musca domestica, Stomoxys calcitrans, Fannia canicularis, and F. femoralis, at Sites in the Western Hemisphere1. Annals of the Entomological Society of America, 1967. 60(2): p. 462-468.
111. Kaufman, P.E., et al., Parasitism Rates of Muscidifurax raptorellus and Nasonia vitripennis (Hymenoptera: Pteromalidae) After Individual and Paired Releases in New York Poultry Facilities. Journal of Economic Entomology, 2001. 94(2): p. 593-598.
112. Andrews, S., FastQC: a quality control tool for high throughput sequence data. 2010, Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom.
113. Weisenfeld, N.I., et al., Direct determination of diploid genome sequences. Genome Res, 2017. 27(5): p. 757-767.
114. Li, D.H., et al., MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics, 2015. 31(10): p. 1674-1676.
115. Bolger, A.M., M. Lohse, and B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 2014. 30(15): p. 2114-2120.
116. Zerbino, D.R. and E. Birney, Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res, 2008. 18(5): p. 821-9.
117. Wences, A.H. and M.C. Schatz, Metassembler: merging and optimizing de novo genome assemblies. Genome Biology, 2015. 16.
118. Kent, W.J., BLAT - The BLAST-like alignment tool. Genome Research, 2002. 12(4): p. 656-664.
119. Bernt, M., et al., MITOS: Improved de novo metazoan mitochondrial genome annotation. Molecular Phylogenetics and Evolution, 2013. 69(2): p. 313-319.
120. Trapnell, C., L. Pachter, and S.L. Salzberg, TopHat: discovering splice junctions with RNA-Seq. Bioinformatics, 2009. 25(9): p. 1105-1111.
121. Seppey, M., M. Manni, and E.M. Zdobnov, BUSCO: Assessing Genome Assembly and Annotation Completeness. Methods Mol Biol, 2019. 1962: p. 227-245.
122. Andrews, S., et al., FastQC. 2010.

123. Weisenfeld, N.I., et al., Direct determination of diploid genome sequences. Genome research, 2017. 27(5): p. 757-767.
124. Cheng, H., et al., Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nature Methods, 2021. 18(2): p. 170-175.
125. Nurk, S., et al., HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. Genome research, 2020. 30(9): p. 1291-1305.
126. Koren, S., et al., Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res, 2017. 27(5): p. 722-736.
127. Zheng, G.X., et al., Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. Nature biotechnology, 2016. 34(3): p. 303-311.
128. McKenna, A., et al., The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research, 2010. 20(9): p. 1297-1303.
129. DePristo, M.A., et al., A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature genetics, 2011. 43(5): p. 491-498.
130. Quinlan, A.R., BEDTools: the Swiss-army tool for genome feature analysis. Current protocols in bioinformatics, 2014. 47(1): p. 11.12. 1-11.12. 34.
131. Chakraborty, M., et al., Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. Nucleic acids research, 2016. 44(19): p. e147-e147.
132. Walker, B.J., et al., Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PloS one, 2014. 9(11): p. e112963.
133. Seppey, M., M. Manni, and E.M. Zdobnov, BUSCO: assessing genome assembly and annotation completeness. Methods in molecular biology (Clifton, NJ), 2019. 1962: p. 227-245.
134. Martinson, E.O., et al., The Evolution of Venom by Co-option of Single-Copy Genes. Current Biology, 2017. 27(13): p. 2007-+.
135. Haas, B.J., et al., De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature protocols, 2013. 8(8): p. 1494-1512.
136. Trapnell, C., et al., Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature protocols, 2012. 7(3): p. 562-578.

137. Flynn, J.M., et al., RepeatModeler2 for automated genomic discovery of transposable element families. Proceedings of the National Academy of Sciences, 2020. 117(17): p. 9451-9457.
138. Bao, Z. and S.R. Eddy, Automated de novo identification of repeat sequence families in sequenced genomes. Genome research, 2002. 12(8): p. 1269-1276.
139. Price, A.L., N.C. Jones, and P.A. Pevzner, De novo identification of repeat families in large genomes. Bioinformatics, 2005. 21(suppl_1): p. i351-i358.
140. Benson, G., Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Research, 1999. 27(2): p. 573-580.
141. Tarailo-Graovac, M. and N. Chen, Using RepeatMasker to identify repetitive elements in genomic sequences Curr Protoc Bioinformatics. 2009. Chapter.
142. Cantarel, B.L., et al., MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Research, 2008. 18(1): p. 188-196.
143. Haas, B.J., et al., De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols, 2013. 8(8): p. 1494-1512.
144. Trapnell, C., et al., Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc, 2012. 7(3): p. 562-78.
145. Kim, D., et al., TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome biology, 2013. 14(4): p. 1-13.
146. Cantarel, B.L., et al., MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome research, 2008. 18(1): p. 188-196.
147. Korf, I., Gene finding in novel genomes. BMC Bioinformatics, 2004. 5: p. 59.
148. Stanke, M. and S. Waack, Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics, 2003. 19 Suppl 2: p. ii215-25.
149. Stanke, M., et al., Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics, 2006. 7(1): p. 62.
150. Keilwagen, J., F. Hartung, and J. Grau, GeMoMa: homology-based gene prediction utilizing intron position conservation and RNA-seq data. Gene prediction: Methods and protocols, 2019: p. 161-177.

151. Rago, A., et al., OGS2: genome re-annotation of the jewel wasp Nasonia vitripennis. BMC genomics, 2016. 17(1): p. 1-25.
152. Dalla Benetta, E., et al., Genome elimination mediated by gene expression from a selfish chromosome. Science advances, 2020. 6(14): p. eaaz9808.
153. Kurtz, S., et al., Versatile and open software for comparing large genomes. Genome Biol, 2004. 5(2): p. R12.
154. Emms, D.M. and S. Kelly, OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol, 2019. 20(1): p. 238.
155. Benetta, E.D., et al., Genome Elimination Mediated by Gene Expression from a Selfish Chromosome. 2019: p. 793273.
156. Wang, Y., et al., MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic acids research, 2012. 40(7): p. e49-e49.
157. Krzywinski, M., et al., Circos: an information aesthetic for comparative genomics. Genome research, 2009. 19(9): p. 1639-1645.
158. Adams, M.D., et al., The genome sequence of Drosophila melanogaster. Science, 2000. 287(5461): p. 2185-95.
159. International Aphid Genomics, C., Genome sequence of the pea aphid Acyrthosiphon pisum. PLoS Biol, 2010. 8(2): p. e1000313.
160. Honeybee Genome Sequencing, C., Insights into social insects from the genome of the honeybee Apis mellifera. Nature, 2006. 443(7114): p. 931-49.
161. Colbourne, J.K., et al., The ecoresponsive genome of Daphnia pulex. Science, 2011. 331(6017): p. 555-61.
162. Kirkness, E.F., et al., Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. Proc Natl Acad Sci U S A, 2010. 107(27): p. 12168-73.
163. Lawniczak, M.K., et al., Widespread divergence between incipient Anopheles gambiae species revealed by whole genome sequences. Science, 2010. 330(6003): p. 512-4.
164. Xia, Q., et al., A draft sequence for the genome of the domesticated silkworm (Bombyx mori). Science, 2004. 306(5703): p. 1937-40.

165. Emms, D.M. and S. Kelly, OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol, 2015. 16: p. 157.
166. Katoh, K. and D.M. Standley, MAFFT: iterative refinement and additional methods. Methods Mol Biol, 2014. 1079: p. 131-46.
167. Darriba, D., et al., ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics, 2011. 27(8): p. 1164-5.
168. Stamatakis, A., RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics, 2014. 30(9): p. 1312-3.
169. Kriventseva, E.V., et al., OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic acids research, 2019. 47(D1): p. D807-D811.
170. Helmkampf, M., et al., Draft Genome of the Rice Coral Montipora capitata Obtained from Linked-Read Sequencing. Genome Biol Evol, 2019. 11(7): p. 2045-2054.
171. Lin, Z.J., et al., Comparative analysis reveals the expansion of mitochondrial DNA control region containing unusually high GC tandem repeat arrays in Nasonia vitripennis. International Journal of Biological Macromolecules, 2021. 166: p. 1246-1257.
172. Gokhman, V.E. and M. Westendorff, The Chromosomes of three species of the Nasonia complex (Hymenoptera, Pteromalidae). Beiträge zur Entomologie= Contributions to Entomology, 2000. 50(1): p. 193-198.
173. Goodpasture, C.E., Cytological data and classification of the Hymenoptera. 1974: University of California, Davis.
174. Silva-Junior, J., S. Pompolo, and L. Campos. Cytogenetics of some species of parasitic wasps of the families Pteromalidae and Eulophidae. in Abstracts. XXI International Congress of Entomology. Brazil, August. 2000.
175. Niehuis, O., et al., Recombination and its impact on the genome of the haplodiploid parasitoid wasp Nasonia. PLoS One, 2010. 5(1): p. e8597.
176. Desjardins, C.A., et al., Fine-scale mapping of the Nasonia genome to chromosomes using a high-density genotyping microarray. G3 (Bethesda), 2013. 3(2): p. 205-15.

177. Rago, A., et al., OGS2: genome re-annotation of the jewel wasp Nasonia vitripennis. BMC Genomics, 2016. 17: p. 678.
178. Taylor, D.B., R.D. Moon, and D.R. Mark, Economic impact of stable flies (Diptera: Muscidae) on dairy and beef cattle production. J Med Entomol, 2012. 49(1): p. 198-209.
179. Oyarzún, M., A. Quiroz, and M. Birkett, Insecticide resistance in the horn fly: alternative control strategies. Medical and veterinary entomology, 2008. 22(3): p. 188-202.
180. Oyarzún, M., A. Li, and C. Figueroa, High levels of insecticide resistance in introduced horn fly (Diptera: Muscidae) populations and implications for management. Journal of economic entomology, 2011. 104(1): p. 258-265.
181. Guglielmone, A.A., et al., Dynamics of cypermethrin resistance in the field in the horn fly, Haematobia irritans. Medical and veterinary entomology, 2002. 16(3): p. 310-315.
182. Petersen, J. and D. Currey, Reproduction and development of Muscidifurax raptorellus (Hymenoptera: Pteromalidae), a parasite of filth flies. Journal of agricultural entomology (USA), 1996.
183. Geden, C.J. and R.D. Moon, Host ranges of gregarious muscoid fly parasitoids: Muscidifurax raptorellus (Hymenoptera: Pteromalidae), Tachinaephagus zealandicus (Hymenoptera: Encyrtidae), and Trichopria nigra (Hymenoptera: Diapriidae). Environmental entomology, 2009. 38(3): p. 700-707.
184. Machtinger, E.T. and C.J. Geden, Biological control with parasitoids, in Pests and vector-borne diseases in the livestock industry. 2018. p. 299-335.
185. Scott, J.G., et al., Insecticide resistance in house flies from the United States: resistance levels and frequency of pyrethroid resistance alleles. Pesticide biochemistry and physiology, 2013. 107(3): p. 377-384.
186. Wylie, H., CONTROL OF EGG FERTILIZATION BY NASONIA VITRIPENNIS (HYMENOPTERA: PTEROMALIDAE) WHEN LAYING ON PARASITIZED HOUSE FLY PUPAE1. The Canadian Entomologist, 1973. 105(5): p. 709-718.
187. Rivers, D.B. and D.L. Denlinger, Developmental fate of the flesh fly, Sarcophaga bullata, envenomated by the pupal ectoparasitoid, Nasonia vitripennis. Journal of insect physiology, 1994. 40(2): p. 121-127.

188. Yoder, J., G. Theriot, and D. Rivers, Venom from Nasonia vitripennis alters water loss from the flesh fly, Sarcophaga bullata. Entomologia experimentalis et applicata, 1996. 81(2): p. 235-238.
189. Rivers, D.B., J. Zdarek, and D.L. Denlinger, Disruption of pupariation and eclosion behavior in the flesh fly, Sarcophaga bullata Parker (Diptera: Sarcophagidae), by venom from the ectoparasitic wasp Nasonia vitripennis (Walker)(Hymenoptera: Pteromalidae). Archives of Insect Biochemistry and Physiology: Published in Collaboration with the Entomological Society of America, 2004. 57(2): p. 78-91.
190. Bishop, D., A. Heath, and N. Haack, Distribution, prevalence and host associations of Hymenoptera parasitic on Calliphoridae occurring in flystrike in New Zealand. Medical and Veterinary Entomology, 1996. 10(4): p. 365-370.
191. Caleffe, R.R.T., et al., Biological control of diptera calliphoridae: A Review. Journal of the Entomological Research Society, 2019. 21(2): p. 144-155.
192. Wu, M., et al., Phylogenomics of the reproductive parasite Wolbachia pipientis wMel: A streamlined genome overrun by mobile genetic elements. Plos Biology, 2004. 2(3): p. 327-341.
193. Foster, J., et al., The Wolbachia genome of Brugia malayi: endosymbiont evolution within a human pathogenic nematode. PLoS Biol, 2005. 3(4): p. e121.
194. Klasson, L., et al., Genome evolution of Wolbachia strain wPip from the Culex pipiens group. Mol Biol Evol, 2008. 25(9): p. 1877-87.
195. Klasson, L., et al., The mosaic genome structure of the Wolbachia wRi strain infecting Drosophila simulans. Proc Natl Acad Sci U S A, 2009. 106(14): p. 5725-30.
196. Kent, B.N., et al., Complete bacteriophage transfer in a bacterial endosymbiont (Wolbachia) determined by targeted genome capture. Genome Biol Evol, 2011. 3: p. 209-18.
197. Mavingui, P., et al., Whole-genome sequence of Wolbachia strain wAlbB, an endosymbiont of tiger mosquito vector Aedes albopictus. J Bacteriol, 2012. 194(7): p. 1840.
198. Saha, S., et al., Survey of endosymbionts in the Diaphorina citri metagenome and assembly of a Wolbachia wDi draft genome. PLoS One, 2012. 7(11): p. e50067.

199. Darby, A.C., et al., Analysis of gene expression from the Wolbachia genome of a filarial nematode supports both metabolic and defensive roles within the symbiosis. Genome Res, 2012. 22(12): p. 2467-77.
200. Ellegaard, K.M., et al., Comparative genomics of Wolbachia and the bacterial species concept. PLoS Genet, 2013. 9(4): p. e1003381.
201. Siozios, S., et al., Draft Genome Sequence of the Wolbachia Endosymbiont of Drosophila suzukii. Genome Announc, 2013. 1(1).
202. Duplouy, A., et al., Draft genome sequence of the male-killing Wolbachia strain wBol1 reveals recent horizontal gene transfers from diverse sources. BMC Genomics, 2013. 14: p. 20.
203. Desjardins, C.A., et al., Genomics of Loa loa, a Wolbachia-free filarial parasite of humans. Nat Genet, 2013. 45(5): p. 495-500.
204. Pinto, S.B., et al., Transcriptional Regulation of Culex pipiens Mosquitoes by Wolbachia Influences Cytoplasmic Incompatibility. Plos Pathogens, 2013. 9(10).
205. Brelsfoard, C., et al., Presence of extensive Wolbachia symbiont insertions discovered in the genome of its host Glossina morsitans morsitans. PLoS Negl Trop Dis, 2014. 8(4): p. e2728.
206. Nikoh, N., et al., Evolutionary origin of insect-Wolbachia nutritional mutualism. Proc Natl Acad Sci U S A, 2014. 111(28): p. 10257-62.
207. Sutton, E.R., et al., Comparative genome analysis of Wolbachia strain wAu. BMC Genomics, 2014. 15: p. 928.
208. Derks, M.F., et al., The Genome of Winter Moth (Operophtera brumata) Provides a Genomic Perspective on Sexual Dimorphism and Phenology. Genome Biol Evol, 2015. 7(8): p. 2321-32.
209. Newton, I.L., et al., Comparative Genomics of Two Closely Related Wolbachia with Different Reproductive Effects on Hosts. Genome Biol Evol, 2016. 8(5): p. 1526-42.
210. Lindsey, A.R., et al., Comparative Genomics of a Parthenogenesis-Inducing Wolbachia Symbiont. G3 (Bethesda), 2016. 6(7): p. 2113-23.
211. Chung, M., et al., Draft genome sequence of the Wolbachia endosymbiont of Wuchereria bancrofti wWb. Pathog Dis, 2017. 75(9).

212. Faddeeva-Vakhrusheva, A., et al., Coping with living in the soil: the genome of the parthenogenetic springtail Folsomia candida. Bmc Genomics, 2017. 18.
213. Badawi, M., et al., Investigating the Molecular Genetic Basis of Cytoplasmic Sex Determination Caused by Wolbachia Endosymbionts in Terrestrial Isopods. Genes (Basel), 2018. 9(6).
214. Conner, W.R., et al., Genome comparisons indicate recent transfer of wRi-like Wolbachia between sister species Drosophila suzukii and D.subpulchrella. Ecology and Evolution, 2017. 7(22): p. 9391-9404.
215. Ramirez-Puebla, S.T., et al., Genomes of Candidatus Wolbachia bourtzisii wDacA and Candidatus Wolbachia pipientis wDacB from the Cochineal Insect Dactylopius coccus (Hemiptera: Dactylopiidae). G3-Genes Genomes Genetics, 2016. 6(10): p. 3343-3349.
216. Woolfit, M., et al., Genomic Evolution of the Pathogenic Wolbachia Strain, wMelPop. Genome Biology and Evolution, 2013. 5(11): p. 2189-2204.
217. Gerth, M. and C. Bleidorn, Comparative genomics provides a timeframe for Wolbachia evolution and exposes a recent biotin synthesis operon transfer. Nature Microbiology, 2017. 2(3).
218. Wang, X., et al., Genome assembly of the A-group Wolbachia in Nasonia oneida using linked-reads technology. Genome biology and evolution, 2019. 11(10): p. 3008-3013.
219. Jiggins, F.M., et al., Recombination confounds interpretations of Wolbachia evolution. Proc Biol Sci, 2001. 268(1474): p. 1423-7.
220. Jiggins, F.M., The rate of recombination in Wolbachia bacteria. Mol Biol Evol, 2002. 19(9): p. 1640-3.
221. Baldo, L., N. Lo, and J.H. Werren, Mosaic nature of the Wolbachia surface protein. Journal of Bacteriology, 2005. 187(15): p. 5406-5418.
222. Duron, O., et al., Transposable element polymorphism of Wolbachia in the mosquito Culex pipiens: evidence of genetic diversity, superinfection and recombination. Molecular Ecology, 2005. 14.
223. Werren, J.H. and J.D. Bartos, Recombination in Wolbachia. Current Biology, 2001. 11(6): p. 431-435.

224. Chafee, M.E., et al., Lateral phage transfer in obligate intracellular bacteria (Wolbachia): verification from natural populations. Molecular biology and evolution, 2009. 27(3): p. 501-505.
225. Verne, S., et al., Evidence for recombination between feminizing Wolbachia in the isopod genus Armadillidium. Gene, 2007. 397(1-2): p. 58-66.
226. Ros, V.I., et al., Diversity and recombination in Wolbachia and Cardinium from Bryobia spider mites. BMC Microbiol, 2012. 12 Suppl 1: p. S13.
227. Baldo, L., et al., Widespread recombination throughout Wolbachia genomes. Molecular biology and evolution, 2006. 23(2): p. 437-449.
228. Reuter, M. and L. Keller, High levels of multiple Wolbachia infection and recombination in the ant Formica exsecta. Mol Biol Evol, 2003. 20(5): p. 748-53.
229. Ilinsky, Y. and O.E. Kosterin, Molecular diversity of Wolbachia in Lepidoptera: Prevalent allelic content and high recombination of MLST genes. Mol Phylogenet Evol, 2017. 109: p. 164-179.
230. Foster, J., et al., Recombination in wolbachia endosymbionts of filarial nematodes? Appl Environ Microbiol, 2011. 77(5): p. 1921-2.
231. Suyama, M., D. Torrents, and P. Bork, PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res, 2006. 34(Web Server issue): p. W609-12.
232. Rambaut, A. FigTree v1.4.4, A Graphical Viewer of Phylogenetic Trees. 2018 [cited 2020 4/5/2020]; Available from: https://github.com/rambaut/figtree/.
233. Kosakovsky Pond, S.L., et al., Automated phylogenetic detection of recombination using a genetic algorithm. Molecular biology and evolution, 2006. 23(10): p. 1891-1901.
234. Jolley, K.A. and M.C. Maiden, BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics, 2010. 11: p. 595.
235. Tamura, K., M. Nei, and S. Kumar, Prospects for inferring very large phylogenies by using the neighbor-joining method. Proceedings of the National Academy of Sciences of the United States of America, 2004. 101(30): p. 11030-11035.
236. Kumar, S., G. Stecher, and K. Tamura, MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol, 2016. 33(7): p. 1870-4.

237. Harrell Jr, F.E. and M.F.E. Harrell Jr, Package 'Hmisc'. CRAN2018, 2019: p. 235-6.
238. Wang, X., et al., Genome Assembly of the A-Group Wolbachia in Nasonia oneida Using Linked-Reads Technology. Genome Biol Evol, 2019. 11(10): p. 3008-3013.
239. Wolfgang, S., FtsH – a single-chain charonin? Fems Microbiology Reviews, 1999(1): p. 1.
240. Vladimirov, et al., Identification of 50S components neighboring 23S rRNA nucleotides A2448 and U2604 within the. Biochemistry, 2000.
241. Oguiza, J.A., et al., A gene encoding arginyl-tRNA synthetase is located in the upstream region of the lysA gene in Brevibacterium lactofermentum: regulation of argS-lysA cluster expression by arginine. Journal of bacteriology, 1993. 175(22): p. 7356-7362.
242. Bleidorn, C. and M. Gerth, A critical re-evaluation of multilocus sequence typing (MLST) efforts in Wolbachia. FEMS Microbiol Ecol, 2018. 94(1).
243. Bordenstein, S.R. and S.R. Bordenstein, Eukaryotic association module in phage WO genomes from Wolbachia. Nat Commun, 2016. 7: p. 13155.
244. Wang, G.H., et al., Bacteriophage WO Can Mediate Horizontal Gene Transfer in Endosymbiotic Wolbachia Genomes. Front Microbiol, 2016. 7: p. 1867.
245. Czarnetzki, A.B. and C.C. Tebbe, Detection and phylogenetic analysis of Wolbachia in Collembola. Environ Microbiol, 2004. 6(1): p. 35-44.
246. Fountain, M.T. and S.P. Hopkin, Folsomia candida (Collembola): a "standard" soil arthropod. Annu Rev Entomol, 2005. 50: p. 201-22.
247. Vandekerckhove, T.T.M., et al., Phylogenetic analysis of the 16S rDNA of the cytoplasmic bacterium Wolbachia from the novel host Folsomia candida (Hexapoda, Collembola) and its implications for wolbachial taxonomy. Fems Microbiology Letters, 1999. 180(2): p. 279-286.
248. Ma, Y., et al., Revisiting the phylogeny of Wolbachia in Collembola. Ecol Evol, 2017. 7(7): p. 2009-2017.
249. Boohene, C., C. Geden, and J. Becnel, Evaluation of remediation methods for Nosema disease in Muscidifurax raptor (Hymenoptera: Pteromalidae). Environmental entomology, 2003. 32(5): p. 1146-1153.
250. Steinhaus, E., Insect Pathology V1: An Advanced Treatise. Vol. 1. 2012: Elsevier.

251. Williams, B.A., et al., Genome sequence surveys of Brachiola algerae and Edhazardia aedis reveal microsporidia with low gene densities. BMC genomics, 2008. 9: p. 1-9.
252. Tsaousis, A.D., et al., A novel route for ATP acquisition by the remnant mitochondria of Encephalitozoon cuniculi. Nature, 2008. 453(7194): p. 553-556.
253. Peyretaillade, E., et al., Microsporidia, amitochondrial protists, possess a 70-kDa heat shock protein gene of mitochondrial evolutionary origin. Molecular biology and evolution, 1998. 15(6): p. 683-689.
254. Hirt, R.P., et al., A mitochondrial Hsp70 orthologue in Vairimorpha necatrix: molecular evidence that microsporidia once contained mitochondria. Current Biology, 1997. 7(12): p. 995-998.
255. Germot, A., H. Philippe, and H. Le Guyader, Evidence for loss of mitochondria in Microsporidia from a mitochondrial-type HSP70 in Nosema locustae. Molecular and biochemical parasitology, 1997. 87(2): p. 159-168.
256. Germot, A., H. Philippe, and H. Le Guyader, Evidence for loss of mitochondria in Microsporidia from a mitochondrial-type HSP70 in Nosema locustae1Note: Nucleotide sequence data reported in this paper has been submitted to the GenBank™ data base under the accession number U97520.1. Molecular and Biochemical Parasitology, 1997. 87(2): p. 159-168.
257. Cornman, R.S., et al., Genomic analyses of the microsporidian Nosema ceranae, an emergent pathogen of honey bees. PLoS pathogens, 2009. 5(6): p. e1000466.
258. de Vienne, D.M., et al., Cospeciation vs host-shift speciation: methods for testing, evidence from natural associations and relation to coevolution. New Phytologist, 2013. 198(2): p. 347-385.
259. Quiles, A., et al., Microsporidian infections in the species complex Gammarus roeselii (Amphipoda) over its geographical range: evidence for both host–parasite co-diversification and recent host shifts. Parasites & Vectors, 2019. 12(1): p. 327.
260. Quiles, A., et al., Dictyocoela microsporidia diversity and co-diversification with their host, a gammarid species complex (Crustacea, Amphipoda) with an old history of divergence and high endemic diversity. BMC Evolutionary Biology, 2020. 20(1): p. 149.

261. Kogan, M. and E. Legner, A biosystematic revision of the genus Muscidifurax (Hymenoptera: Pteromalidae) with descriptions of four new species. The Canadian Entomologist, 1970. 102(10): p. 1268-1290.
262. Geden, C.J., et al., Effect of Fluctuating High Temperatures on House Flies (Diptera: Muscidae) and Their Principal Parasitoids (Muscidifurax spp. and Spalangia spp. [Hymenoptera: Pteromalidae]) From the United States. J Med Entomol, 2019. 56(6): p. 1650-1660.
263. Xiong, X., et al., Long-read assembly and annotation of the parasitoid wasp Muscidifurax raptorellus, a biological control agent for filth flies. Front Genet, 2021. 12: p. 748135.
264. Seppey, M., M. Manni, and E.M. Zdobnov, BUSCO: assessing genome assembly and annotation completeness, in Gene prediction. 2019, Springer. p. 227-245.
265. Amoroso, E., The evolution of viviparity. Proceedings of the Royal Society of Medicine, 1968. 61(11P2): p. 1188-1200.
266. Vurture, G.W., et al., GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics, 2017. 33(14): p. 2202-2204.
267. Zhou, Y., et al., Profiles of telomeric repeats in Insecta reveal diverse forms of telomeric motifs in Hymenopterans. Life science alliance, 2022. 5(7).
268. Nawrocki, E.P. and S.R. Eddy, Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics, 2013. 29(22): p. 2933-2935.
269. Nawrocki, E.P., et al., Rfam 12.0: updates to the RNA families database. Nucleic Acids Research, 2014. 43(D1): p. D130-D137.
270. Min, B., I.V. Grigoriev, and I.-G. Choi, FunGAP: Fungal Genome Annotation Pipeline using evidence-based gene model evaluation. Bioinformatics, 2017. 33(18): p. 2936-2937.
271. Stanke, M., et al., AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic acids research, 2006. 34(suppl_2): p. W435-W439.
272. Katoh, K. and D.M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular biology and evolution, 2013. 30(4): p. 772-780.

273. Needleman, S.B. and C.D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of molecular biology, 1970. 48(3): p. 443-453.
274. Wang, D., et al., How do variable substitution rates influence Ka and Ks calculations? Genomics, proteomics & bioinformatics, 2009. 7(3): p. 116-127.
275. Wang, D., et al., KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. Genomics, proteomics & bioinformatics, 2010. 8(1): p. 77-80.
276. Yang, Z. and R. Nielsen, Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Molecular biology and evolution, 2000. 17(1): p. 32-43.
277. Wang, D.-P., et al., γ-MYN: a new algorithm for estimating Ka and Ks with consideration of variable substitution rates. Biology Direct, 2009. 4(1): p. 1-18.
278. Salazar, A.N., et al., Nanopore sequencing enables near-complete de novo assembly of Saccharomyces cerevisiae reference strain CEN.PK113-7D. FEMS Yeast Res, 2017. 17(7).
279. Kanehisa, M. and Y. Sato, KEGG Mapper for inferring cellular functions from protein sequences. Protein Sci, 2020. 29(1): p. 28-35.
280. Kanehisa, M., et al., KEGG for taxonomy-based analysis of pathways and genomes. Nucleic Acids Research, 2022. 51(D1): p. D587-D592.
281. Suzuki, S., et al., GHOSTX: an improved sequence homology search algorithm using a query suffix array and a database suffix array. PloS one, 2014. 9(8): p. e103833.
282. Kanehisa, M., et al., Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic acids research, 2014. 42(D1): p. D199-D205.
283. Kanehisa, M., Y. Sato, and K. Morishima, BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. Journal of Molecular Biology, 2016. 428(4): p. 726-731.
284. Corradi, N., A. Gangaeva, and P.J. Keeling, Comparative profiling of overlapping transcription in the compacted genomes of microsporidia Antonospora locustae and Encephalitozoon cuniculi. Genomics, 2008. 91(4): p. 388-393.
285. Gill, E.E., J.J. Becnel, and N.M. Fast, ESTs from the microsporidian Edhazardia aedis. BMC genomics, 2008. 9(1): p. 1-12.

286. Bailey, T.L., et al., MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic acids research, 2006. 34(suppl_2): p. W369-W373.
287. Pelin, A., et al., Genome analyses suggest the presence of polyploidy and recent human-driven expansions in eight global populations of the honeybee pathogen N osema ceranae. Environmental Microbiology, 2015. 17(11): p. 4443-4458.
288. ping Chen, Y., et al., Genome sequencing and comparative genomics of honey bee microsporidia, Nosema apis reveal novel insights into host-parasite interactions. BMC genomics, 2013. 14(1): p. 1-16.
289. Shen, W. and H. Ren, TaxonKit: a practical and efficient NCBI taxonomy toolkit. Journal of Genetics and Genomics, 2021. 48(9): p. 844-850.
290. Katoh, K. and D.M. Standley, MAFFT: iterative refinement and additional methods, in Multiple sequence alignment methods. 2014, Springer. p. 131-146.
291. Stamatakis, A., RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics, 2014. 30(9): p. 1312-1313.
292. Goffeau, A., et al., Life with 6000 genes. Science, 1996. 274(5287): p. 546-567.
293. Eldarov, M.A., et al., Whole-genome analysis of three yeast strains used for production of sherry-like wines revealed genetic traits specific to flor yeasts. Frontiers in microbiology, 2018. 9: p. 965.
294. Pei, B., et al., The First Report on the Transovarial Transmission of Microsporidian Nosema bombycis in Lepidopteran Crop Pests Spodoptera litura and Helicoverpa armigera. Microorganisms, 2021. 9(7).
295. Nakabachi, A. and N.A. Moran, Extreme Polyploidy of Carsonella, an Organelle-Like Bacterium with a Drastically Reduced Genome. Microbiol Spectr, 2022. 10(3): p. e0035022.
296. Peska, V. and S. Garcia, Origin, diversity, and evolution of telomere sequences in plants. Frontiers in plant science, 2020. 11: p. 117.
297. Van de Peer, Y., E. Mizrachi, and K. Marchal, The evolutionary significance of polyploidy. Nature Reviews Genetics, 2017. 18(7): p. 411-424.
298. Valero, M., et al., Evolution of alternation of haploid and diploid phases in life cycles. Trends Ecol Evol, 1992. 7(1): p. 25-9.

299. Amses, K.R., et al., Diploid-dominant life cycles characterize the early evolution of Fungi. Proc Natl Acad Sci U S A, 2022. 119(36): p. e2116841119.
300. McBride, H.M., M. Neuspiel, and S. Wasiak, Mitochondria: More Than Just a Powerhouse. Current Biology, 2006. 16(14): p. R551-R560.
301. Vowinckel, J., et al., The metabolic growth limitations of petite cells lacking the mitochondrial genome. Nature Metabolism, 2021. 3(11): p. 1521-1535.
302. Karnkowska, A., et al., A Eukaryote without a Mitochondrial Organelle. Current Biology, 2016. 26(10): p. 1274-1284.
303. Yahalomi, D., et al., A cnidarian parasite of salmon (Myxozoa: Henneguya) lacks a mitochondrial genome. Proceedings of the National Academy of Sciences of the United States of America, 2020. 117(10): p. 5358-5363.
304. Zervas, A., G. Petersen, and O. Seberg, Mitochondrial genome evolution in parasitic plants. BMC Evolutionary Biology, 2019. 19(1): p. 87.
305. Petersen, G., et al., Massive gene loss in mistletoe (Viscum, Viscaceae) mitochondria. Scientific reports, 2015. 5(1): p. 1-7.
306. Skippington, E., et al., Miniaturized mitogenome of the parasitic plant Viscum scurruloideum is extremely divergent and dynamic and has lost all nad genes. Proceedings of the National Academy of Sciences, 2015. 112(27): p. E3515-E3524.
307. Dunn, A.M., R.S. Terry, and J.E. Smith, Transovarial transmission in the microsporidia. Adv Parasitol, 2001. 48: p. 57-100.
308. Terry, R.S., A.M. Dunn, and J.E. Smith, Cellular distribution of a feminizing microsporidian parasite: a strategy for transovarial transmission. Parasitology, 1997. 115 (Pt 2): p. 157-63.
309. Becnel, J.J., Life cycles and host-parasite relationships of Microsporidia in culicine mosquitoes. Folia Parasitol (Praha), 1994. 41(2): p. 91-6.