The first draft genomes of the ant *Formica exsecta,* and its *Wolbachia* endosymbiont reveal extensive gene transfer from endosymbiont to host

Authors: Kishor Dhaygude[1], Abhilash Nair[1], Helena Johansson[1], Yannick Wurm[2], and Liselotte Sundström[1, 3]

1) Organismal and Evolutionary Biology Research Programme, Faculty of Biological and environmental sciences P.O. Box 65, FI-00014 University of Helsinki, Finland

2) Organismal Biology Department, School of Biological and Chemical Sciences, Queen Mary University of London, Mile End Road, London E1 4NS, United Kingdom

3) Tvärminne Zoological Station, University of Helsinki, J.A. Palménin tie 260, FI-10900 Hanko, Finland

Abstract:

The wood ant *Formica exsecta* (Formicidae; Hymenoptera), is a common ant species throughout the Palearctic region. The species is a well established model for studies of ecological characteristics and evolutionary conflict. In this study, we sequenced and assembled draft genomes for *Formica exsecta* and its endosymbiont *Wolbachia*. The draft *F. exsecta* genome is 277.7 Mb long; we identify 13,767 protein coding genes for which we provide gene ontology, and protein domain annotations. This is also the first report of a *Wolbachia* genome from ants, and provides insights into the phylogenetic position of this endosymbiont. We also identified multiple horizontal gene transfer events (HGTs) from *Wolbachia* to *F. exsecta*. Some of these HGTs have also occurred in parallel in multiple other insect genomes, highlighting the extent of HGTs in eukaryotes. We expect that the *F. exsecta* genome will be valuable resource in further exploration of the molecular basis of the evolution of social organization.

Key words: *Formica exsecta*, genome, endosymbionts, transposons, horizontal gene transfer, *Wolbachia*

Introduction

Adapting to changes in the environment is the foundation of species survival, and is usually thought to be a gradual process. Genomic changes, such as single nucleotide substitutions play key roles in adaptive evolution, although few mutations are beneficial. Besides nucleotide substitutions, other structural and regulatory units, such as transposable elements (TEs) and epigenetic modifications, can also act as drivers in adaptation (González et al., 2010; Rostant, Wedell & Hosken, 2012; Casacuberta & González, 2013). Genetic material can also be acquired from other organisms by means of horizontal gene transfer (HGTs), and this can also lead to novel adaptive traits (Schönknecht, Weber & Lercher, 2014; Wybouw et al., 2016). Both mutations and HGTs can drive rapid genome evolution (Dunning Hotopp, 2011; Boto, 2014). Horizontal gene transfers have been reported in many taxa, most commonly from bacteria to eukaryotes (Dunning Hotopp, 2011), plants (Yue et al., 2012; Matveeva & Lutova, 2014), fungi (Rolland et al., 2009; Fitzpatrick, 2012; Bruto et al., 2014), but the underlying mechanisms that underpin horizontal gene transfer events, and mode by which bacterial genetic material is integrated into the eukaryote genome are not well understood.

Many cases of horizontal gene transfer from bacteria to eukaryotes involve intracellular endosymbionts, which are maternally transmitted through oocytes (Werren, 1997; Ferree et al., 2005). The most common examples of endosymbiont to host horizontal gene transfers involve the bacterium *Wolbachia*, a well described intracellular, maternally inherited gram-negative bacterium known to infect over 40% of the investigated insect species (Werren, 1997; Werren, Baldo & Clark, 2008). *Wolbachia* infection is also prevalent in filarial nematodes, crustaceans, and arachnids (Cordaux, Michel-Salzat & Bouchon, 2001; Fenn et al., 2006; Goodacre et al., 2006). *Wolbachia*- host interactions can be mutualistic or pathogenic (Moya et al., 2008). A number of ecdysozoan genomes have been reported to contain chromosomal insertions originating from *Wolbachia*, including the mosquito *Aedes aegypti* (Klasson et al., 2009a; Woolfit et al., 2009), the longhorn beetle *Monochamus alternatus* (Aikawa et al., 2009), filarial nematodes of the genera *Onchocerca*, *Brugia*, and *Dirofilaria* (Fenn et al., 2006; Hotopp et al., 2007), parasitoid wasps of the genus *Nasonia*, the fruit fly *Drosophila ananassae*, the pea aphid *Acythosiphon pisum* (Nikoh & Nakabachi, 2009; Nikoh et al., 2010), and the bean beetle *Callosobruchus chinensis* (Kondo et al., 2002). Although most of the transferred DNA is probably nonfunctional in the host genome (Kondo et al., 2002; Hotopp et al., 2007; Nikoh et al., 2008), some of the transferred genes are functional (Klasson et al., 2009a). These genes are expressed in specific tissues, are subject to purifying selection, and are involved in processes such as protein synthesis inhibition, membrane transport and metabolism (Hotopp et al., 2007; Woolfit et al., 2009; McNulty et al., 2013).

Infection with *Wolbachia* is widespread in Hymenoptera. Most hymenopteran *Wolbachia* infections have the cytoplasmic incompatibility phenotype (Werren & Windsor, 2000), which leads to reproductive incompatibility between infected sperm and uninfected eggs. Wenseleers et al. (1998) showed that 25 out of 50 species of ants in Java and Sumatra screened positive for one strain of *Wolbachia*. By contrast, a study on a single Swiss population of the ant *Formica exsecta*, found that all the ants tested were infected with four or five different strains of *Wolbachia* (Keller et al., 2001; Reuter & Keller, 2003).

The aims of this study are to test whether horizontally transferred genetic elements exist in the genome of the ant *Formica exsecta*, and to describe the genomic organization of any such elements. The genus *Formica* is listed by the Global Ant Genome Alliance (GAGA) as one of the high-priority ant taxons to be sequenced (Boomsma et al., 2017; http://antgenomics.dk/), owing to its key taxonomic position, and the ecological and behavioral data that are available for the species. To date, no genome sequence is available for this genus.

Our study population of *F. exsecta*, located on the Hanko peninsula, Southwestern Finland, has been monitored since 1994, and data on demography, genetic structure, and ecology are available (Sundström, Chapuisat & Keller, 1996; Sundström, Keller & Chapuisat, 2003; Haag-Liautard et al., 2009; Vitikainen, Haag-Liautard & Sundström, 2015). Based on genetic data on colony kin structure most (97%) of the approximately 200 colonies are known to have a single reproductive queen, mated to one or more (usually two) males (Sundström, Chapuisat & Keller, 1996; Sundström, Keller & Chapuisat, 2003; Haag-Liautard et al., 2009; Vitikainen, Haag-Liautard & Sundström, 2015). We report the whole genome sequencing of this species, and the draft genome sequence of its associated cytoplasmic *Wolbachia* endosymbiont (wFex). We further report the presence of multiple extensive insertions of *Wolbachia* genetic material in the host genome, and compare the HGTs insertions discovered in the assembled draft genome to other genomes, to understand the pattern of HGT events between endosymbiont and host. We analyze in detail the genomic features of *F. exsecta* along with its endosymbiont *Wolbachia*, and discuss our findings in the light of genome evolution in *Wolbachia* and its host.

Materials and Methods

Sample collection and genome sequencing.

We selected one single-queen colony from our study population on the island Furuskär (F162), and collected 200 adult males from this colony. We used males because in Hymenoptera these arise through arrhenotoky (Normark, 2003) and are haploid (Crozier, 1975), meaning that a pool of males together are representative of the diploid genome of their mother. DNA extraction was done from testis, which contains sperm cells and organ tissue, to avoid contamination by gut microbiota. We used a Qiagen Genomic-tip 20/G extraction kit according to the manufacturer's protocol. For Illumina sequencing we constructed three small insert paired-end libraries (insert sizes of 200 bp, 500 bp, 800 bp), and four mate pair (large insert paired-end) libraries (insert sizes of 2 kb, 5 kb, 10 kb and 20 kb), each containing DNA from 15-50 pooled males. Libraries were prepared using protocols recommended by the manufacturers. Sequencing was done at the Beijing Genomics Institute (BGI) using HiSeq2000, which produced a total of 99.97 GB of raw data (Table 1).

Genome assembly

We assembled the *F. exsecta* genome using SOAPdenovo2 version 2.04 (Xie et al., 2014) in three main steps. First, a de Bruijn graph was constructed using short length insert library reads with default parameters (k-mer value of 45), to construct the contigs. The initial contig assembly contained 104,190 contigs with an N50 size of 22,328 bp, and total length of 276.23 Mb of sequence, at an average depth of coverage of 47.37X. Second, all individual reads were realigned onto the contigs. Because reads are paired, they can aid with scaffolding: The number of reads supporting the adjacency of each pair of contigs was calculated and weighted by the ratio between consistent and conflicting paired ends. Scaffolds were constructed in a stepwise manner using libraries of increasing sizes from 500bp insert size paired-end reads up to mate-pair of 5 kb insert size. 80,473 contigs could not be placed in scaffolds. These are highly similar repetitive sequences, since the cd-hit-est tool (Huang et al., 2010) showed that 43% of these contigs clustered together at 80% of the sequence length. Third, sequencing gaps in the scaffolds were closed with the two mate-pair libraries (Insert size 10 kb and 20 kb). Overall, these steps produced an initial assembly with an N50 scaffold length of 949,634 bp, and a total length of 289,843,734 bp with each scaffold longer than 200 bp.

We used blobology v1.0 (Kumar et al., 2013) to generate taxon-annotated GC-coverage (TAGC) plots of scaffolds in the genome assembly, which can help to identify bacterial contamination (Supplementary Figure S1). The scaffolds for the TAGC plot were successfully annotated to the taxonomic order based on the best blast match to the NCBI nt database (O'Leary et al., 2016). This analysis revealed that 74 scaffolds matched the endosymbiotic bacterium *Wolbachia*. Sixty-nine of these scaffolds were removed as we concluded that they are part of the *Wolbachia* genome (see analysis below), but five contigs were retained in the final assembly for *F. exsecta* as they contained both *Wolbachia* and ant sequences. Following this curation, the final draft genome assembly was 277.7 Mb long with an N50 value of 997,654 bp and 36% Guanine-cytosine (GC) content (Table 2).

Genome assembly of *Wolbachia*

All 25 published *Wolbachia* genomes were obtained from the NCBI database (O'Leary et al., 2016). We aligned the 74 scaffolds from the initial *F. exsecta* assembly that matched with *Wolbachia* against these genomes using MUMmer 3.23 (Kurtz et al., 2004), and inspected the alignments manually. Sixty-nine of the 74 scaffolds matched completely to *Wolbachia* genomic regions. These 69 scaffolds represented 3.09 Mb total, with a N50 value of 104,167 bp, henceforth we refer to this group of scaffolds as "the *Wolbachia* endosymbiont genome of *F. exsecta*" (wFex).

The remaining five scaffolds each contained several interspersed fragments with similarity to *Wolbachia* genomes, whereas other parts of these scaffolds had high similarity to genomes of ants. Furthermore, the sequencing coverage of these scaffolds was similar to the *F. exsecta* scaffolds, rather than to the *Wolbachia* scaffolds. Finally, detailed inspection of these scaffolds in a genome browser showed no change in sequencing depth where we identify the interspersed fragments with similarity to *Wolbachia*, which would be expected for erroneous chimeric assembly (Lasken & Stockwell, 2007). These data thus suggest that fragments of *Wolbachia* were horizontally transferred to the *F. exsecta* genome. To corroborate these results with independent approaches, we re-assembled the raw sequencing data with two additional independent algorithms that we expect would make different types of assembly errors than SOAPdenovo. The first software, Velvet version 1.2.09 (Zerbino & Birney, 2008), is also based on a de Bruijn graph; the second, SGA version 0.10.5 (Simpson & Durbin, 2012) is based on a string graph. Both resulting assemblies confirmed the patterns we had seen, and validate the idea that the five SOAPdenovo scaffolds containing sequence with similarity to both ants, and *Wolbachia* represent horizontal gene transfers from *Wolbachia* to *F. exsecta*.

We further compared the sequences of the horizontally transferred fragments in the five SOAPdenovo scaffolds against the NCBI (nr/nt) database (O'Leary et al., 2016), using blast 2.2.27 (Altschul et al., 1990) to determine whether these fragments may have also undergone horizontal gene transfer in other arthropod genomes. We performed analogous searches on ant genomes present in the NCBI, and the Fourmidable databases (Wurm et al., 2009). When a positive match with any other ant or arthropod genomes was found, the exact location of the insertion was determined, and compared with that of *F. exsecta*. Finally, the five scaffolds were also compared to the *F. exsecta* transcriptome (Dhaygude et al., 2017), using blastn 2.2.27, to assess similarity with expressed sequences.

Quantitative assessment of genome assemblies

The quality of the genome assembly is crucial, as it defines the quality of all subsequent analyses that are based on the genome sequences. We explored multiple assembly options (data not shown), and used two methods to assess assembly quality and robustness in order to select the highest quality assembly. First, we evaluated genome contiguity (number and length of contigs) using Quast 3.2 (Gurevich et al., 2013) to assess whether our newly assembled draft genome is comparable to published ant genomes (Favreau et al., 2018) based on assembly statistics (N50,N90).

Second, we used core gene content-based quality assessment using CEGMA 2.4 (Parra et al., 2007) to ascertain that the 248 most highly conserved eukaryotic proteins are present in our genome assembly. We also compared genes present in our genome assembly to single-copy orthologs across four lineage-specific sets (Eukaryota (303 genes), Insecta (1,658 genes), Arthropoda (2,675 genes), and Hymenoptera (4,415 genes)) using the BUSCO 1.1(Simão et al., 2015). In addition, we compared the *F. exsecta* genome with 13 other ant genomes, *Camponotus floridanus, Atta cephalotes, Acromyrmex echinatior*, *Cardiocondyla obscurior, Cerapachys biroi, Lasius niger*, *Linepithema humile, Monomorium pharaonis, Pogonomyrmex barbatus, Vollenhovia emeryi, Wasmannia auropunctata, Harpegnathos saltator*, and *Solenopsis invicta* (Wurm et al., 2009), using BUSCO. We report BUSCO quality metrices for the *F. exsecta* genome. (Table 3).

The quality of the *Wolbachia* endosymbiont genome was quantified with a similar approach, where we used BUSCO to examine the presence of Universal Single-Copy Orthologs of the Bacteria (148 genes), and the Proteobacteria (221 genes) lineages (Table 3). We also used BUSCO to compare the wFex genome with four other *Wolbachia* genomes, including the *Wolbachia* endosymbionts of *Drosophila simulans (wRi), Culex quinquefasciatus (wPip), Drosophila melanogaster (wMel),* and *Drosophila simulans (wNo).*

Gene prediction

We combined several publicly available data sets and computational gene prediction tools to establish an Official Gene Set (OGS) for the *F. exsecta* genome. First, we used the MAKER version 2.28 pipeline (Cantarel et al., 2008; Holt & Yandell, 2011), to derive consensus gene models from Augustus version 3.1.0 (Stanke & Morgenstern, 2005), SNAP version 2016-07-28 (Korf, 2004), and Exonerate version 2.2.0 (Slater & Birney, 2005). For this MAKER prediction we used as input datasets the *F. exsecta* transcriptome (ESTs) (Bioproject ID: PRJNA213662, (Dhaygude et al., 2017)), and the proteomes of all available ant species (Uniprot download on 20-04-2015). The longest protein at each genomic locus was retained, resulting in a set of 23,517 gene models. Because samples may have different sets of transcripts, owing to different biological conditions or developmental stages (Dhaygude et al., 2017), we additionally made a separate transcript-spliced assembly using RNA sequences generated from separate libraries for different life stages (Dhaygude et al., 2017), using the Tophat version 2.1.0 (Trapnell, Pachter & Salzberg, 2009), and Cufflinks version 2.2.1 (Trapnell et al., 2010). The assemblies from the different samples were then merged using cuffmerge (Trapnell et al., 2010). We further obtained separate Augustus version 3.1.0 (Stanke & Morgenstern, 2005), and Glimmer version 3.02 (Salzberg et al., 1998) gene models with default settings (Augustus: --species=fly --genemodel=partial, --strand=both, Glimmer: +f, +s, -g 60). The gene sets and gene models from MAKER and from other programs were then merged. Redundancy was removed by favoring for each transcript the longest prediction starting with a methionine. If several transcripts had the same length we retained the one which had the best support from the cufflinks transcript assembly. This redundancy removal resulted in a final set of 13,637 protein coding gene models (final OGS), which contained 33,121 transcripts.

Genome Annotation

We analyzed the complete official gene sets (OGS) of *F. exsecta* to identify sequence and functional similarity by comparing with different sequence databases using blast. By using a ribosomal database, we were able to annotate both the large (LSU), and the small (SSU) subunit ribosomal RNAs. The remaining gene sequences were used for retrieving functional information from other databases (SwissProt, Pfam, PROSITE, and COG). Gene sequences were considered to be coding if they had a strong unique hit to the SwissProt protein database (Magrane & Consortium, 2011; The Uniprot Consortium, 2017), or appeared to be orthologs of known predicted protein-coding genes from ant species based on TrEMBL (Translation of EMBL nucleotide sequence database). We also assigned putative metabolic pathways, functional classes, enzyme classes, GeneOntology terms, and locus names with the AutoFact tool (Koski et al., 2005). To further improve annotation, and for assigning biological function (e.g. gene expression, metabolic pathways), we also did orthologous searches by comparing with other Hymenoptera sequences (Wurm et al., 2009). To quantify variation in the numbers of protein family members, we performed Pfam (version 24.0) (Bateman et al., 2004) and PROSITE profile (Sigrist et al., 2010) analyses on proteins obtained from the *F. exsecta* gene set. Our final annotation included gene sequences with retrieved protein-related names, functional domains, and expression in other organisms along with enzyme commission (EC) numbers, pathway information, Cluster of Orthologous Groups (COG), functional classes, and Gene Ontology terms.

Orthology and evolutionary rates

Comparative genome-wide analysis of orthologous genes was performed with OrthoVenn (Wang et al., 2015) to compare the predicted *F. exsecta* protein sequences with those of four other ant species, *Camponotus floridanus*, *Lasius niger*, *Solenopsis invicta*, and *Cerapachys biroi,* all of which were downloaded from their respective public NCBI repositories. The predicted proteins of *F. exsecta* and the other four species were uploaded into the OrthoVenn web server for identification and comparison of orthologous clusters (Wang et al., 2015). Following clustering, orthAgogue was used for the identification of putative orthology and inparalogy relationships. To deduce the putative function of each ortholog, the first protein sequence from each cluster was searched against the non-redundant protein database UniProt using blastp 2.2.27. Pairwise sequence similarities among protein sequences were determined for all species with a blastp 2.2.27 (E-value cut-off of 10^{-5}, and an inflation value of 1.5 for MCL). Finally, an interactive Venn diagram, summary counts, and functional summaries of clusters shared between species were visualized using OrthoVenn.

To identify genes under positive or relaxed purifying selection in *F. exsecta*, we estimated the rates of non-synonymous to synonymous changes for core orthologous genes (3,156) from five ant species (*F. exsecta*, *Camponotus floridanus*, *Lasius niger*, *Solenopsis invicta*, and *Cerapachys biroi*). For this we only included orthologous groups with one ortholog for each species (no paralogous genes were included) in the analysis. We extracted coding and protein sequences for 3,156 orthologous groups from the respective public NCBI repositories for the species included. We then aligned all protein sequences using Clustal Omega (Sievers & Higgins, 2014), and then converted them to nucleotide sequences with PAL2NAL version 14 (Yang, 1997). We then ran CODEML version 4.9e (Yang, 1997), using the branch site model with *F. exsecta* as foreground branch, and the other five ant species as background lineages. The Bayes empirical method (Yang et al. 2005) was used to estimate the posterior probabilities, which were then used to identify sites under selection. We additionally estimated pairwise dN/dS ratios for orthologous genes (5,148 genes) between *Camponotus floridanus* and *F. exsecta* in CODEML.

We also ran an orthology analysis between the proteins from three *Wolbachia* species published previously (wRi, wDac, wNo; (Klasson et al., 2009b; Ellegaard et al., 2013; Ramirez-Puebla et al., 2016)), to find similarity with the predicted protein sets of the newly assembled wFex genome. Orthologs were identified using OrthoVenn (E-value cut-off of 10^{-5} and inflation value 1.5). In addition, we analyzed the paralogous genes within the wFex genome, to help understand the increased genome size in comparison to other *Wolbachia* genomes.

Discovery and annotation of transposable elements

We used RepeatMasker version 4.0.7 (Smit. et al., 2015), and the TransposonPSI version 08-22-2010 (Brian J. Haas, 2011) to detect repetitive elements in the genome. To retrieve and mask repetitive elements, we downloaded files from the Repbase and Dfam databases, and aligned each of them with the *F. exsecta* genome sequences as query sequences. Positive alignments were regarded as repetitive regions and extracted for further analysis. To identify genome sequence region homology to proteins encoded by different families of transposable elements, we used the TransposonPSI analysis tool. This tool uses PSI-blast, with a collection of retro-transposon ORF homology profiles to identify statistically significant alignments.

***Wolbachia* phylogeny**

We analysed the phylogeny of *Wolbachia* in MrBayes v3.2.6 x64 (Ronquist & Huelsenbeck, 2003), using a concatenated sequences of 35 genes. For this analysis, each gene was considered as a different partition, and the most fitting nucleotide substitution model was chosen for each gene, using the bayesian information criterion (BIC) in the program jMODELTEST (Posada, 2008). The partitioned dataset was run for 200,000 generations, sampling at every 100th generation with each partition unlinked for the substitution parameters. Convergence of the runs was confirmed by checking that the potential scale reduction factor was ~1.0 for all model parameters, and by ensuring that an average split frequency of standard deviations < 0.01 was reached (Ronquist & Huelsenbeck, 2003). The first 25% of the trees were discarded as burn-in, and the remaining trees were used to create a 50% majority-rule consensus tree, and to estimate the posterior probabilities. To check for consistency of the phylogeny, Markov chain Monte Carlo (MCMC) runs were repeated to get a similar 50% majority-rule consensus tree with high posterior probabilities. The phylogenetic tree generated was visualized using Figtree v1.4.2 (Rambaut, 2012).

Results & Discussion

Assembly of the *Formica exsecta* genome

We created Illumina sequencing libraries from DNA extracted from testes of males of a *F. exsecta* colony to obtain >99 gigabases of Illumina sequence data. The final *F. execta* genome resulting from assembly of this data was 277.7 megabases (Mb) long, encompassing 14,617 scaffolds (Figure 1) with a N50 scaffold length of 997.7 kb (Table 2). The number of scaffolds is higher than the number of chromosomes reported for *F. exsecta* (n=26; Agosti & Hauschteck-Jungen, 1987; Rosengren, Rosengren & Söderlund, 2009). Similarly, the *F. exsecta* genome assembly is somewhat shorter than genome size estimates obtained by flow cytometry for species in the subfamily Formicinae (range: 296-385 Mb; Tsutsui et al., 2008). These discrepancies are unsurprising given the difficulty of assembling highly repetitive gene content from short sequencing reads (Henson, Tischler & Ning, 2012). In line with this, the genome assembly length metrics are similar to those of the 23 ant genomes that have been published. The raw data, gapped scaffolds, and annotations underpinning this assembly are deposited into public databases under BioProject PRJNA393850 (accession NPMM00000000).

Quantitative assessment of genome assembly

Based on scaffold N50 and N75 statistics, contig size, and GC content, the *F. exsecta* genome assembly is comparable in quality and completeness to other sequenced ant genomes (Supplementary Table S1). All the 248 CEGMA eukaryotic core genes were found, and 241 of these genes were complete in length. Similarly, 98.5% of 1634 BUSCO Insecta genes were complete in the genome (Table 3). These results held with other BUSCO analysis levels including Eukaryota, Arthropoda, and Hymenoptera, with low duplication levels (2.2% to 5.3%), and few missing genes (0.6% to 1.27%; all details in Table 3). Such discrepancies can be due to technical artifacts such as sequencing biases or assembly difficulties, as well as to true differences between our *F. exsecta* sample and the BUSCO and CEGMA datasets. To further evaluate genome completeness, we compared the independently generated *F. exsecta* transcriptome (Dhaygude et al., 2017) to the genome reported here. More than 98.75 % of the 10.999 assembled ESTs mapped unambiguously to the genome (blastn $E < 10^{-50}$). Together, these analyses show that the genome assembly has high completeness.

Gene Content in the *Formica exsecta* genome

We identified 13,637 protein coding genes by combining *ab initio*, EST-based, and sequence similarity based gene predictions methods. The GC content was higher in exons (41.6%) than in introns (30.6%), a pattern similar to that reported in the honey bee, *Apis mellifera*, and the fire ant, *Solenopsis invicta* (Weinstock et al., 2006; Wurm et al., 2011). Despite this, as in other ant genomes (Schrader et al., 2014; Boomsma et al., 2017), overall GC content in genes (35.1%) was similar to the rest of the genome (36.0%).

We used blast and orthology analyses to characterize *F. exsecta* genes. The vast majority (88%; 12,050) of these had the highest blastp similarity to genes in other ants. A further 0.4% had the highest similarity to Apidae, and 0.6% to Braconidae, Amniota, and *Wolbachia* (the latter probably due to HGT; see below and Figure 2). The remaining 3.09% belong to other taxa not included in Figure 2 because they had fewer than 20 hits. The remaining genes (7.91%, n= 1,080) lacked clear sequence similarity [cutoff for blastx $E < 10^{-3}$] to known protein sequences or protein domains. Some of these may represent erroneous gene predictions (Drăgan et al., 2016), however 994 of them are ≥1000 bp and include an open reading frame >300 amino acids long, which is unlikely to occur by chance. Importantly, although only a single pooled transcriptome library, prepared from different developmental life stage samples, was available for *F. exsecta*, 235 of the genes are expressed (FPKM ≥ 1; Dhaygude et al., 2017). It is thus likely that a high proportion of the 1,080 genes are taxonomically restricted genes unique to the *F. exsecta* lineage.

The total genes of *F. exsecta* (n=13,637) were grouped into 7,727 orthologous clusters (Figure 3). Comparative analysis of the *F. exsecta* genes with the closely related species *C. floridanus* and *L. niger*, and the more distantly related *S. invicta* and *C. biroi* revealed, that 4,685 orthologous clusters out of 7,727 are shared between all five species. In addition, we found 102 gene clusters that were exclusive to three Formicinae genomes (*F. exsecta*, *C. floridanus* and *L. niger;* Supplementary Table S2). Such genes are important candidates that could be involved in the evolution of this subfamily. Many of the genes in these clusters had no detectable relation to existing genes outside the Formicinae; those that did included GO annotations such as glycerate kinase, transferase activity, deoxyribonucleoside diphosphate metabolic process.

Interestingly, 633 of the *F. exsecta*-specific genes could be grouped into 197 ortholog clusters of 2 or more genes (Supplementary Table S3), suggesting not only newly evolved genes, but also potential gene duplication and subfunctionalisation. Previous comparative genome studies have indicated that 10-20% of genes lack recognizable homologs in other species in every taxonomic group so far studied (Wilson et al., 2007; Khalturin et al., 2009; Johnson & Tsutsui, 2011; Tautz & Domazet-Lošo, 2011). Our lower percentage of orphan genes could be due to our hierarchical approach to annotation, the wide range of databases used, and the large amounts of ant genomic data generated over the past years (Favreau et al., 2018).

Genes with signatures of evolution under positive selection

We performed analyses to detect genes with signatures of positive selection in *F. exsecta*. First, selection analysis (dN/dS ratio estimations) on 3,157 single-copy genes shared between the five core ant species (without paralogous genes), revealed that 500 genes have signatures of positive selection in the lineage leading to *F. exsecta*. These include genes involved in fatty acid metabolism, lipid catabolism, and chitin metabolism (Supplementary Table S4). Interestingly, previous studies on ants, bees, and flies also provide evidence for positive selection on genes in similar functional categories as in our study (Roux et al., 2014). For example, genes involved in biological functions such as carbohydrate metabolic processes, lipid metabolic processes, cytoskeleton organization, cell surface receptor signaling pathways, and RNA processing were overrepresented in the enrichment analysis, and such genes were also previously reported as positively selected genes in ants, bees, and flies (Viljakainen et al., 2009; Roux et al., 2014).

To perform a similar analysis on a larger number of genes, we used a second approach based on pairwise comparisons between *F. exsecta* and *C. floridanus*. Out of 5,148 one-to-one- orthologs, 29 showed dN/dS > 1 (P < 0.005; Supplementary Table S5). Although some of these putative genes could be artefactual or non-coding, they all include an open reading frame of > 100 amino acids. Five (17%) out of 29 genes are likely linked to transposon activity as they are transposase-like or have EpsG domains. Among the other genes, only a few are annotated: the Icarapin-like protein is a venom gene, and such genes have been shown to be under positive selection in wasps (Werren et al., 2010). Perhaps more surprisingly we found high dN/dS for the Homeobox protein gene orthopedia which is involved in early embryonic development (Mackenzie et al., 1991).

Repetitive elements

Repetitive elements comprised 15.88% (44.10 Mb) of the *F. exsecta* assembly. This proportion is similar to that found in other ants (16.5-31.5% (Schrader et al., 2014). This is probably an underestimate because (i) genomic regions that cannot be assembled are enriched with such repeats, (ii) multiple copies of a repetitive element are often collapsed into a single copy during genome assembly, and (iii) only a portion of repetitive elements in *F. exsecta* will have similarity to sequences in standard repeat databases. Overall, 3.18% (8.8 Mb) of the assembly was composed of simple repeats, whereas 12.73% (35.34 Mb) comprised interspersed repeats, most of which (53.73%) could not be classified. Among those that could be classified, 10,542 retro element fragments represented 2.74% of the genome, and 53,438 DNA transposons represented 4.23% of the genome. The *F. exsecta* genome contains copies of the piggyBac transposon (23 in total, and 7 within intact ORFs). Higher numbers (234) of piggyBac transposons have been found in *C. floridanus*, yet only 6 of these were found within ORFs (Bonasio et al., 2010).

The *Wolbachia* endosymbiont genome of *Formica exsecta*

The assembly of the *Wolbachia* endosymbiont, wFex, was 3.09 Mb long, encompassing 69 scaffolds with a N50 scaffold length of 104,167 nt, and a GC content of 35.13% (Table 2; GenBank: RCIU00000000, Bioproject: PRJNA436771). This assembly of wFex shows extensive nucleotide similarity with the complete genome of the *Wolbachia* endosymbiont of *Drosophila simulans*, wNo (GenBank ID:

NC_021084), and covers approximately 84% of its length (Supplementary Figure S2). We determined that 549 genes are present as a single copy in the *Wolbachia* genomes most closely related to wFex ((Lindsey et al., 2016) see below); 537 (99.6%) out of these 539 core genes are present in the wFex genome, suggesting high completeness.

However, the wFex genome is considerably larger (3.09 Mb) than the *Wolbachia* genomes reported previously (range: 0.95 to 1.66 Mb; Sun et al., 2001), and includes a greater number of open reading frames (1,796 ORFs) than other published *Wolbachia* genomes [range: 644 to 1,275 genes]. *Formica exsecta* is known to harbor more than one *Wolbachia* strain (Reuter & Keller, 2003), thus these patterns could be due to the presence of multiple endosymbiont strains. Two additional lines of evidence support this idea. First, 212 genes (11.80 %), that are present as single-copy genes in the wMel, wRi and wDac genomes (Klasson et al., 2009b; Ellegaard et al., 2013; Ramirez-Puebla et al., 2016), are duplicated in our assembly (Supplementary Table S6). Furthermore, 92 (12%) of the 775 genes present as a single copy in wFex, included genetic variation within our sample, including in the cytochrome oxidase subunit I; no such variation is normally expected. Despite extensive attempts, we were unable to disentangle the two or more *Wolbachia* strains – this is likely because differences in synteny between the strains cannot be resolved using short-read sequence data. Similar assembly artifacts, due to multiple *Wolbachia* strains, have also been reported by other studies (Ramírez-Puebla et al., 2016).

To determine how wFex is related to other *Wolbachia*, we used Bayesian phylogenetic analysis based on 35 conserved genes (Supplementary Table S7) from the 25 available *Wolbachia* genomes from the NCBI database. The analysis revealed three distinct monophyletic clades, all with posterior probabilities >0.9. Each of these clades represent one super group of *Wolbachia* (Figure 4). Of these three supergroups, two have been found only in arthropods (super groups A and B), and the third super group is found only in filarial nematodes (super group C; Werren, Baldo & Clark, 2008). In the phylogenetic analysis, wFex clustered with the *Wolbachia* strains within super group A, and most closely matched the strain that infects the scale insect, *Dactylopius coccus*, (wDacA). This is consistent with earlier studies on *Wolbachia* in ants, which also found supergroup A in the majority of the infected ants (Werren & Windsor, 2000).

Given that wFex affiliates with the supergroup A in our phylogenetic analysis, we investigated the extent to which its gene content aligned with that of other *Wolbachia* genomes in the same supergroup. We found that 525 genes were shared across all strains in this supergroup, including wFex (Figure 5). About 20% of these genes had no match to known proteins, whereas the remaining genes matched a wide range of predicted functions (Ellegaard et al., 2013; Lindsey et al., 2016). We also found strain-specific genes (wFex - 50 genes, wMel - 4 genes, wRi - 3 genes, wDac - 9 genes). The wFex-specific genes included inferred annotations including Ankyrin repeat protein, ATP synthase, and chromosome partition protein (Supplementary Table S8). These strain-specific genes can provide an interesting snapshot of the evolutionary dynamics of a species. For example, ankyrin repeat proteins are involved in numerous functional processes, and have been suggested to play an important role in host-symbiont interactions (Li, Mahajan & Tsai, 2006). Comparative analyses suggest that they may be involved in host communication and reproductive phenotypes (Voronin & Kiseleva, 2008).

To explore differences in gene content between CI-inducing and mutualist strains of *Wolbachia*, homologous genes in six CI-inducing strains, and three mutualist strains were aligned and compared (Lindsey et al., 2016). The mutualist *Wolbachia* strains (range: 644-805 genes) had fewer genes than the CI-inducing ones (range: 911-1,275 genes). The CI-inducing strains shared 84 genes not found in the mutualist strains. We found 80 (95.23%) of these 84 genes in wFex (Supplementary Figure S3), suggesting that wFex may be CI-inducing.

Horizontal gene transfers, and functional novelty

Intracellular symbionts can contribute new genes or fragments of genes to the host genome via horizontal gene transfer (Keeling & Palmer, 2008; Werren, Baldo & Clark, 2008; Dunning Hotopp, 2011). We found evidence for ancestral horizontal transfer from *Wolbachia* to the host *F. exsecta* in five scaffolds (scaffold83, scaffold233, scaffold574, scaffold707, scaffold741). The four largest transfers are 13 to 47 kb long, and include 83 putative functional protein coding genes, whereas the fifth and smallest insertion (475 bp) lacks protein coding genes other than a degenerate *Wolbachia* transposase. This transposase is present in 7 out of 29 published *Wolbachia* genomes. Our analysis shows that similar transfer events of this homologous fragment apparently also have occurred from *Wolbachia* to the genomes of the ants *Vollenhovia emeryi* (gene: LOC105557741), and *Cardiocondyla obscurior* (scaffolds scf7180001101632 and scf7180001108526), as well as the microfilarial nematode *Brugia pahangi*, the Arizona spittle bug *Clastoptera arizona*, and the parasitoid wasp *Diachasma alloeum*.

One-third of invertebrate genomes are thought to contain recent *Wolbachia* gene insertions, ranging in size from short segments (<600 bp), to nearly the entire genome (Hotopp et al., 2007; Werren, Baldo & Clark, 2008). Most of these transferred fragments contained transposable elements, as well as some other functional genes from the *Wolbachia* genome. The HGT events from *Wolbachia* to *F. exsecta* are located in or near regions with transposases. Our blast results suggest that four of the insert regions had *Wolbachia* transposases, whereas one insert region has a transposase of ant origin. Whether the presence of such transposases close to HGT sites facilitates insertions is unknown. Interestingly, the putative functional protein-coding genes of *Wolbachia* inserted in the *F. exsecta* genome are similar to the genes reported in similar HGTs events in other insect genomes (eg: ABC transporter, Ankyrin repeat containing protein (Table 4) (Brelsfoard et al., 2014; International Glossina Genome Initiative, 2014). This could indicate that some HGT events are either more likely to occur or to be retained for reasons that could be neutral or adaptive to the host or to the endosymbiont. The transcriptome of *F. exsecta* shows that at least 6 out of the 83 genes from the *Wolbachia* HGT regions are transcribed but with a low FPKM values (range 0.04 to 1.6). These low level transcription trait often observed in bacteria-eukaryote HGTs (Hotopp et al., 2007; Nikoh et al., 2008; Dunning Hotopp, 2011).

Conclusions

Here we present the first draft genome of the ant *F. exsecta*, and its *Wolbachia* endosymbiont. This is the first report of a *Wolbachia* genome from ants, and provides insights into its phylogenetic position. We further identified multiple HGT events from *Wolbachia* to *F. exsecta*. Some of these have also occurred in parallel in several other insect genomes, highlighting the extent of HGTs in eukaryotes. We expect that the *F. exsecta* genome will be a valuable resource in understanding the molecular basis of the evolution of social organization in ants: Recent genomic comparisons between *Formica selysi* and *S. invicta* have shown convergent evolution of a social chromosome, that underpins social organisation in these ants (Purcell et al., 2014). Additional comparison of these genomic regions with *F. exsecta* could provide valuable insights on the evolution of genomic architectures underlying social organization.

Acknowledgements

The authors thank Kalevi Trontti, Jenni Paviala and Minttu Ahjos for help with the laboratory work, Pekka Pamilo, Jonna Kulmuni for useful comments on an earlier draft of the manuscript. This work was funded by the Academy of Finland (Centre of Excellence in Biological Interactions, grants no. 252411 and 284666 to L. Sundström), the University of Helsinki (to L. Sundström), the Biotechnology and Biological Sciences Research Council (grant no. BB/K004204/1 to Yannick Wurm) and the Natural Environment Research Council (grant NE/L00626X/1 to Yannick Wurm).

References

Agosti D., Hauschteck-Jungen E. 1987. Polymorphism of males in *Formica exsecta* Nyl. (Hym.: Formicidae). Insectes Sociaux 34:280–290. DOI: 10.1007/BF02224360.

Aikawa et.al., 2009. Longicorn beetle that vectors pinewood nematode carries many *Wolbachia* genes on an autosome. Proceedings of the Royal Society B: Biological Sciences 276:3791–3798. DOI: 10.1098/rspb.2009.1022.

Altschul et.al., 1990. Basic local alignment search tool. Journal of Molecular Biology 215:403–410. DOI: 10.1016/S0022-2836(05)80360-2.

Bateman et.al., 2004. The Pfam protein families database. Nucleic Acids Research 32:D138-41. DOI: 10.1093/nar/gkh121.

Bonasio et.al 2010. Genomic comparison of the ants *Camponotus floridanus* and *Harpegnathos saltator*. Science 329:1068–1071. DOI: 10.1126/science.1192428.

Boomsma et.al., 2017. The Global Ant Genomics Alliance (GAGA). Myrmecological News 25:61–66.

Boto L. 2014. Horizontal gene transfer in the acquisition of novel traits by metazoans. Proceedings of the Royal Society B: Biological Sciences 281:20132450. DOI: 10.1098/rspb.2013.2450.

Brelsfoard et.al., 2014. Presence of extensive Wolbachia symbiont insertions discovered in the genome of its host *Glossina morsitans morsitans*. PLoS Neglected Tropical Diseases 8:e2728. DOI: 10.1371/journal.pntd.0002728.

Brian J. Haas. 2011. TransposonPSI. http://transposonpsi.sourceforge.net

Bruto et.al., 2014. Frequent, independent transfers of a catabolic gene from bacteria to contrasted filamentous eukaryotes. Proceedings of the Royal Society of London B: Biological Sciences. 281(1789): 20140848. DOI: 10.1098/rspb.2014.0848

Cantarel et.al., 2008. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Research 18:188–96. DOI: 10.1101/gr.6743907.

Casacuberta E., González J. 2013. The impact of transposable elements in environmental adaptation. Molecular Ecology 22:1503–1517. DOI: 10.1111/mec.12170.

Cordaux R., Michel-Salzat A., Bouchon D. 2001. Wolbachia infection in crustaceans: novel hosts and potential routes for horizontal transmission. Journal of Evolutionary Biology 14:237–243. DOI: 10.1046/j.1420-9101.2001.00279.x.

Crozier RH. 1975. Hymenoptera. Animal Cytogenetics:95. ISBN:9783443260040

Dhaygude et.al. 2017. Transcriptome sequencing reveals high isoform diversity in the ant *Formica exsecta*. PeerJ 5:e3998. DOI: 10.7717/peerj.3998.

Drăgan et.al., 2016. GeneValidator: identify problems with protein-coding gene predictions. Bioinformatics 32:1559–1561. DOI: 10.1093/bioinformatics/btw015.

Dunning Hotopp JC. 2011. Horizontal gene transfer between bacteria and animals. Trends in Genetics 27:157–163. DOI: 10.1016/j.tig.2011.01.005.

Ellegaard et.al., 2013. Comparative genomics of Wolbachia and the bacterial species concept. PLoS Genetics 9:e1003381. DOI: 10.1371/journal.pgen.1003381.

Favreau et.al., 2018. Genes and genomic processes underpinning the social lives of ants. Current Opinion in Insect Science 25:83–90. DOI: 10.1016/J.COIS.2017.12.001.

Fenn et.al., 2006. Phylogenetic relationships of the Wolbachia of nematodes and arthropods. PLoS Pathogens 2:e94. DOI: 10.1371/journal.ppat.0020094.

Ferree et al. 2005. Wolbachia utilizes host microtubules and dynein for anterior localization in the *Drosophila* oocyte. PLoS Pathogens 1:e14. DOI: 10.1371/journal.ppat.0010014.

Fitzpatrick DA. 2012. Horizontal gene transfer in fungi. FEMS Microbiology Letters 329:1–8. DOI: 10.1111/j.1574-6968.2011.02465.x.

González et.al., 2010. Genome-wide patterns of adaptation to temperate environments associated with transposable elements in Drosophila. PLoS Genetics 6:e1000905. DOI: 10.1371/journal.pgen.1000905.

Goodacre et.al. 2006. Wolbachia and other endosymbiont infections in spiders. Molecular Ecology 15:517–527. DOI: 10.1111/j.1365-294X.2005.02802.x.

Gurevich A., Saveliev V., Vyahhi N., Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics 29:1072–1075. DOI: 10.1093/bioinformatics/btt086.

Haag-Liautard et al., 2009. Fitness and the level of homozygosity in a social insect. Journal of Evolutionary Biology 22:134–42. DOI: 10.1111/j.1420-9101.2008.01635.x

Henson J., Tischler G., Ning Z. 2012. Next-generation sequencing and large genome assemblies. Pharmacogenomics 13:901–15. DOI: 10.2217/pgs.12.72.

Holt C., Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 12:491. DOI: 10.1186/1471-2105-12-491.

Hotopp et.al., 2007. Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. Science 317:1753–1756. DOI: 10.1126/science.1142490.

Huang Y., Niu B., Gao Y., Fu L., Li W. 2010. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics 26:680–682. DOI: 10.1093/bioinformatics/btq003.

International Glossina Genome Initiative. 2014. Genome sequence of the tsetse fly (Glossina morsitans): vector of African trypanosomiasis. Science (New York, N.Y.) 344:380–6. DOI: 10.1126/science.1249656.

Johnson BR., Tsutsui ND. 2011. Taxonomically restricted genes are associated with the evolution of sociality in the honey bee. BMC Genomics 12:164. DOI: 10.1186/1471-2164-12-164.

Keeling PJ., Palmer JD. 2008. Horizontal gene transfer in eukaryotic evolution. Nature Reviews Genetics 9:605–618. DOI: 10.1038/nrg2386.

Keller et.al., 2001. Sex ratio and Wolbachia infection in the ant *Formica exsecta*. Heredity 87:227–33.DOI:10.1046/j.1365-2540.2001.00918.x

Khalturin et.al. 2009. More than just orphans: are taxonomically-restricted genes important in evolution? Trends in Genetics 25:404–413. DOI: 10.1016/j.tig.2009.07.006.

Klasson et.al., 2009a. Horizontal gene transfer between *Wolbachia* and the mosquito *Aedes aegypti*. BMC Genomics 10:33. DOI: 10.1186/1471-2164-10-33.

Klasson et.al. 2009b. The mosaic genome structure of the *Wolbachia* wRi strain infecting *Drosophila simulans*. Proceedings of the National Academy of Sciences 106:5725–5730. DOI: 10.1073/pnas.0810753106.

Kondo et.al. 2002. Genome fragment of *Wolbachia* endosymbiont transferred to X chromosome of host insect. Proceedings of the National Academy of Sciences 99:14280–14285. DOI: 10.1073/pnas.222228199.

Korf I. 2004. Gene finding in novel genomes. BMC Bioinformatics 5:59. DOI: 10.1186/1471-2105-5-59.

Koski LB., Gray MW., Lang BF., Burger G. 2005. AutoFACT: an automatic functional annotation and classification tool. BMC Bioinformatics 6:151. DOI: 10.1186/1471-2105-6-151.

Kumar et.al. 2013. Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. Frontiers in Genetics 4:237. DOI: 10.3389/fgene.2013.00237.

Kurtz et.al. 2004. Versatile and open software for comparing large genomes. 5. Genome Biology 20045:R12. DOI:10.1186/gb-2004-5-2-r12

Lasken RS., Stockwell TB. 2007. Mechanism of chimera formation during the Multiple Displacement Amplification reaction. BMC Biotechnology 7:19. DOI: 10.1186/1472-6750-7-19.

Li J., Mahajan A., Tsai M-D. 2006. Ankyrin Repeat: A unique motif mediating protein−protein interactions. Biochemistry 45:15168–15178. DOI: 10.1021/bi062188q.

Lindsey ARI., Werren JH., Richards S., Stouthamer R. 2016. Comparative genomics of a parthenogenesis-inducing *Wolbachia* symbiont. G3 (Bethesda, Md.) 6:2113–23. DOI: 10.1534/g3.116.028449.

Mackenzie A., Leeming GL., Jowett AK., Ferguson MW., Sharpe PT. 1991. The homeobox gene Hox 7.1 has specific regional and temporal expression patterns during early murine craniofacial embryogenesis, especially tooth development in vivo and in vitro. Development (Cambridge, England) 111:269–85.

Magrane M., Consortium U. 2011. UniProt Knowledgebase: a hub of integrated protein data. Database 2011:bar009-bar009. DOI: 10.1093/database/bar009.

Matveeva T V., Lutova LA. 2014. Horizontal gene transfer from Agrobacterium to plants. Frontiers in Plant Science 5:326. DOI: 10.3389/fpls.2014.00326.

McNulty SN., Fischer K., Curtis KC., Weil GJ., Brattig NW., Fischer PU. 2013. Localization of Wolbachia-like gene transcripts and peptides in adult Onchocerca flexuosa worms indicates tissue specific expression. Parasites & Vectors 6:2. DOI: 10.1186/1756-3305-6-2.

Moya A., Peretó J., Gil R., Latorre A. 2008. Learning how to live together: genomic insights into prokaryote–animal symbioses. Nature Reviews Genetics 9:218–229. DOI: 10.1038/nrg2319.

Nikoh N., McCutcheon JP., Kudo T., Miyagishima S., Moran NA., Nakabachi A. 2010. Bacterial genes in the aphid genome: absence of functional gene transfer from Buchnera to its host. PLoS Genetics 6:e1000827. DOI: 10.1371/journal.pgen.1000827.

Nikoh N., Nakabachi A. 2009. Aphids acquired symbiotic genes via lateral gene transfer. BMC Biology 7:12. DOI: 10.1186/1741-7007-7-12.

Nikoh et.al. 2008. Wolbachia genome integrated in an insect chromosome: evolution and fate of laterally transferred endosymbiont genes. Genome Research 18:272–80. DOI: 10.1101/gr.7144908.

Normark BB. 2003. The evolution of alternative genetic systems in insects. Annual Review of Entomology 48:397–423. DOI: 10.1146/annurev.ento.48.091801.112703.

O'Leary et.al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Research 44:D733–D745. DOI: 10.1093/nar/gkv1189.

Parra G., Bradnam K., Korf I., Bateman A. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23:1061–1067. DOI: 10.1093/bioinformatics/btm071.

Purcell J., Brelsford A., Wurm Y., Perrin N., Chapuisat M. 2014. Convergent genetic architecture underlies social organization in ants. Current Biology□: CB 24:2728–32. DOI: 10.1016/j.cub.2014.09.071.

Rambaut A. 2012. Figtree Tool (http://tree.bio.ed.ac.uk/software/figtree/)

Ramírez-Puebla et.al., 2016. Genomes of Candidatus Wolbachia bourtzisii wDacA and Candidatus Wolbachia pipientis wDacB from the cochineal insect Dactylopius coccus (Hemiptera: Dactylopiidae). G3 (Bethesda, Md.) 6:3343–3349. DOI: 10.1534/g3.116.031237.

Reuter M., Keller L. 2003. High levels of multiple Wolbachia infection and recombination in the ant *Formica exsecta*. Molecular Biology and Evolution 20:748–753. DOI: 10.1093/molbev/msg082.

Rolland T., Neuvéglise C., Sacerdot C., Dujon B. 2009. Insertion of horizontally transferred genes within conserved syntenic regions of Yeast genomes. PLoS One 4:e6515. DOI: 10.1371/journal.pone.0006515.

Ronquist F., Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics (Oxford, England) 19:1572–4.

Rosengren M., Rosengren R., Söderlund V. 2009. Chromosome numbers in the genus Formica with special reference to the taxonomical position of *Formica uralensis Ruzsk*. and *Formica truncorum* Fabr. Hereditas 92:321–325. DOI: 10.1111/j.1601-5223.1980.tb01715.x.

Rostant WG., Wedell N., Hosken DJ. 2012. Transposable elements and insecticide resistance. Advances in Genetics 78:169–201. DOI: 10.1016/B978-0-12-394394-1.00002-X.

Roux J., Privman E., Moretti S., Daub JT., Robinson-Rechavi M., Keller L. 2014. Patterns of positive selection in seven ant genomes. Molecular Biology and Evolution 31:1661–1685. DOI: 10.1093/molbev/msu141.

Salzberg SL., Delcher AL., Kasif S., White O. 1998. Microbial gene identification using interpolated Markov models. Nucleic Acids Research 26:544–8.

Schönknecht G., Weber APM., Lercher MJ. 2014. Horizontal gene acquisitions by eukaryotes as drivers of adaptive evolution. BioEssays 36:9–20. DOI: 10.1002/bies.201300095.

Schrader et.al. 2014. Transposable element islands facilitate adaptation to novel environments in an invasive species. Nature Communications 5:5495. DOI: 10.1038/ncomms6495.

Sievers F., Higgins DG. 2014. Clustal Omega, accurate alignment of very large numbers of sequences. In: Methods in Molecular Biology (Clifton, N.J.). 105–116. DOI: 10.1007/978-1-62703-646-7_6.

Sigrist CJA., Cerutti L., de Castro E., Langendijk-Genevaux PS., Bulliard V., Bairoch A., Hulo N. 2010. PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Research 38:D161-6. DOI: 10.1093/nar/gkp885.

Simão et.al. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210–3212. DOI: 10.1093/bioinformatics/btv351.

Simpson JT., Durbin R. 2012. Efficient de novo assembly of large genomes using compressed data structures. Genome Research 22:549–556. DOI: 10.1101/gr.126953.111.

Slater GSC., Birney E. 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6:31. DOI: 10.1186/1471-2105-6-31.

Smit., AFA., Hubley R., Green P. 2015. RepeatMasker Open-4.0. http://www.repeatmasker.org/

Stanke M., Morgenstern B. 2005. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Research 33:W465-7. DOI: 10.1093/nar/gki458.

Sun L V., Foster JM., Tzertzinis G., Ono M., Bandi C., Slatko BE., O'Neill SL. 2001. Determination of Wolbachia genome size by pulsed-field gel electrophoresis. Journal of Bacteriology 183:2219–25. DOI: 10.1128/JB.183.7.2219-2225.2001.

Sundström L. 1994. Sex ratio bias, relatedness asymmetry and queen mating frequency in ants. Nature. DOI:10.1038/367266a0

Sundström L., Chapuisat M., Keller L. 1996. Conditional manipulation of sex ratios by ant workers: a test of kin selection theory. Science 274:993–995. DOI: 10.1126/science.274.5289.993

Sundström L., Keller L., Chapuisat M. 2003. Inbreeding and sex-biased gene flow in the ant Formica exsecta. Evolution; international journal of organic evolution 57:1552–61. DOI:10.1111/j.0014-3820.2003.tb00363.x

Tautz D., Domazet-Lošo T. 2011. The evolutionary origin of orphan genes. Nature Reviews Genetics 12:692–702. DOI: 10.1038/nrg3053.

The Uniprot Consortium. 2017. UniProt: the universal protein knowledgebase. Nucleic Acids Research 45:D158–D169. DOI: 10.1093/nar/gkw1099.

Trapnell C., Pachter L., Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25:1105–1111. DOI: 10.1093/bioinformatics/btp120.

Trapnell et.al. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnology 28:511–515. DOI: 10.1038/nbt.1621.

Tsutsui ND., Suarez A V., Spagna JC., Johnston JS. 2008. The evolution of genome size in ants. BMC Evolutionary Biology 8:64. DOI: 10.1186/1471-2148-8-64.

Viljakainen L., Evans JD., Hasselmann M., Rueppell O., Tingek S., Pamilo P. 2009. Rapid evolution of immune proteins in social insects. Molecular Biology and Evolution 26:1791–1801. DOI: 10.1093/molbev/msp086.

Vitikainen E., Haag-Liautard C., Sundström L. 2011. Inbreeding and reproductive investment in the ant *Formica exsecta*. Evolution 65. DOI: 10.1111/j.1558-5646.2011.01273.x.

Vitikainen EIK., Haag-Liautard C., Sundström L. 2015. Natal dispersal, mating patterns, and inbreeding in the ant *Formica exsecta*. The American naturalist 186:716–27. DOI: 10.1086/683799.

Voronin DA., Kiseleva E V. 2008. Functional role of proteins containing ankyrin repeats. Cell and Tissue Biology 2:1–12. DOI: 10.1134/S1990519X0801001X.

Wang Y., Coleman-Derr D., Chen G., Gu YQ. 2015. OrthoVenn: a web server for genome wide comparison and annotation of orthologous clusters across multiple species. Nucleic Acids Research 43:W78-84. DOI: 10.1093/nar/gkv487.

Weinstock et.al., 2006. Insights into social insects from the genome of the honeybee *Apis mellifera*. Nature 443:931–949. DOI: 10.1038/nature05260.

Wenseleers T., Ito F., Van Borm S., Huybrechts R., Volckaert F., Billen J. 1998. Widespread occurrence of the microorganism Wolbachia in ants. Proceedings of the Royal Society B: Biological Sciences 265:1447–1452. DOI: 10.1098/rspb.1998.0456.

Werren JH. 1997. Wolbachia run amok. Proceedings of the National Academy of Sciences of the United States of America 94:11154–5.

Werren JH., Baldo L., Clark ME. 2008. Wolbachia: master manipulators of invertebrate biology. Nature Reviews Microbiology 6:741–751. DOI: 10.1038/nrmicro1969.

Werren et.al. 2010. Functional and evolutionary insights from the genomes of three parasitoid Nasonia species. Science 327:343–348. DOI: 10.1126/science.1178028.

Werren JH., Windsor DM. 2000. Wolbachia infection frequencies in insects: evidence of a global equilibrium? Proceedings. Biological sciences 267:1277–85. DOI: 10.1098/rspb.2000.1139.

Wilson GA., Feil EJ., Lilley AK., Field D. 2007. Large-scale comparative genomic ranking of taxonomically restricted genes (TRGs) in bacterial and archaeal genomes. PLoS One 2:e324. DOI: 10.1371/journal.pone.0000324.

Woolfit M., Iturbe-Ormaetxe I., McGraw EA., O'Neill SL. 2009. An ancient horizontal gene transfer between mosquito and the endosymbiotic bacterium *Wolbachia pipientis*. Molecular Biology and Evolution 26:367–374. DOI: 10.1093/molbev/msn253.

Wurm et.al. 2009. Fourmidable: a database for ant genomics. BMC Genomics 10:5. DOI: 10.1186/1471-2164-10-5.

Wurm et.al., 2011. The genome of the fire ant *Solenopsis invicta*. Proceedings of the National Academy of Sciences 108:5679–5684. DOI: 10.1073/pnas.1009690108.

Wybouw N., Pauchet Y., Heckel DG., Van Leeuwen T. 2016. Horizontal gene transfer contributes to the evolution of arthropod herbivory. Genome Biology and Evolution 8:1785–801. DOI: 10.1093/gbe/evw119.

Xie et.al. 2014. SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. Bioinformatics 30:1660–1666. DOI: 10.1093/bioinformatics/btu077.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Computer Applications in the Biosciences□: CABIOS 13:555–6.

Yue J., Hu X., Sun H., Yang Y., Huang J. 2012. Widespread impact of horizontal gene transfer on plant colonization of land. Nature Communications 3:1152. DOI: 10.1038/ncomms2148.

Zerbino DR., Birney E. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Research 18:821–829. DOI: 10.1101/gr.074492.107.

Data Accessibility

The raw Illumina sequences of paired-end and mate-pair libraries are deposited on the National Center for Biotechnology Information (NCBI) under the bio-project number PRJNA393850, with the accession numbers SAMN07344805-SAMN07344811. The assembled genome sequence of *F. exsecta* is deposited on Genbank with the accession number NPMM00000000. Similarly, the draft genome assembly of wFex is deposited under the project number PRJNA436771.

List of Tables:

List of Figures:

Figure 1. *De novo* genome assembly of *F. exsecta* genome, summarized by the following metrics: a) Overall assembly length, b) Number of scaffolds/contigs, c) Length of the longest scaffold/contig, d) Scaffold/contig N50 and N90, e) Percentage GCs and percentage Ns , f) BUSCO completeness, g) Scaffold/contig length/count distribution.

Figure 2. Taxonomic distribution of the best blastp hits of *F. exsecta* proteins to the non-redundant (nr) protein database ($E < 10^{-5}$).

Figure 3. Venn diagram showing the distribution of gene families (orthologous clusters) among five ant species including three closely related members of the subfamily Formicinae (*Formica exsecta*, *Camponotus floridanus*, *Lasius niger*), and two distinctly related ants *(Solenopsis invicta* and *Cerapachys biroi).*

Figure 4: Phylogeny of the *Wolbachia* supergroups A, B, and C strains with the newly assembled wFex genome. The phylogenetic reconstructions are based on individual analyses of 35 core genes of 25 *Wolbachia* strains. The support values on the branch labels indicate Bayesian posterior probabilities. The letters A-C indicate the separate supergroups.

Figure 5. Venn diagram displaying the overlap in orthologous genes among four *Wolbachia* species including the newly assembled wFex strain and the wDac, wRi, wMel strains reported previously.

Supplementary Tables:

S1: Comparison of assembly statistics of the *F. exsecta* genome and 13 other published ant genomes.

S2: List of genes specific to the Formicinae as identified by OrthoVenn.

S3: List of species-specific genes in *F. exsecta,* as identified by OrthoVenn.

S4: List of *F. exsecta* genes under positive or relaxed purifying selection (dN/dS ratios > 1) in comparison to five other ant species (*Camponotus floridanus*, *Lasius niger*, *Solenopsis invicta* and *Cerapachys biroi*)

S5: List of *F. exsecta* genes showing dN/dS ratios > 1 in pairwise comparison to *Camponotus floridanus.*

S6: List of genes with paralogs in the wFex genome, which are present as single copies in the wMel, wRi, wDac genomes.

S7: List of conserved *Wolbachia* genes used for phylogenetic analysis.

S8: List of species-specific genes in wFEX genome, as identified by OrthoVenn.

Supplementary Figures:

S1. TAGC plot of *F. exsecta*, and its *Wolbachia* endosymbiont. The TAGC plots were taxonomically annotated, and the contigs with best similarity to Arthropoda and Proteobacteria are highlighted in color.

S2. Visualization of genome coverage of wFex against the *Wolbachia* endosymbiont of *Drosophila simulans* (wNo) genome, using the alignment software Mummer.

S3. Venn diagram displaying the overlap in orthologous genes across CI-inducing and mutualist *Wolbachia* species.

Figure 1 The de novo genome assembly of the *F. exsecta* genome summarized by different metrics: a) Overall assembly length, b) Number of scaffolds/contigs, c) Length of longest scaffold/contig, d) Scaffold/contig N50 and N90, e) Percentage GC and percentage Ns, f) BUSCO completeness. g) Scaffold/contig length/count distribution.

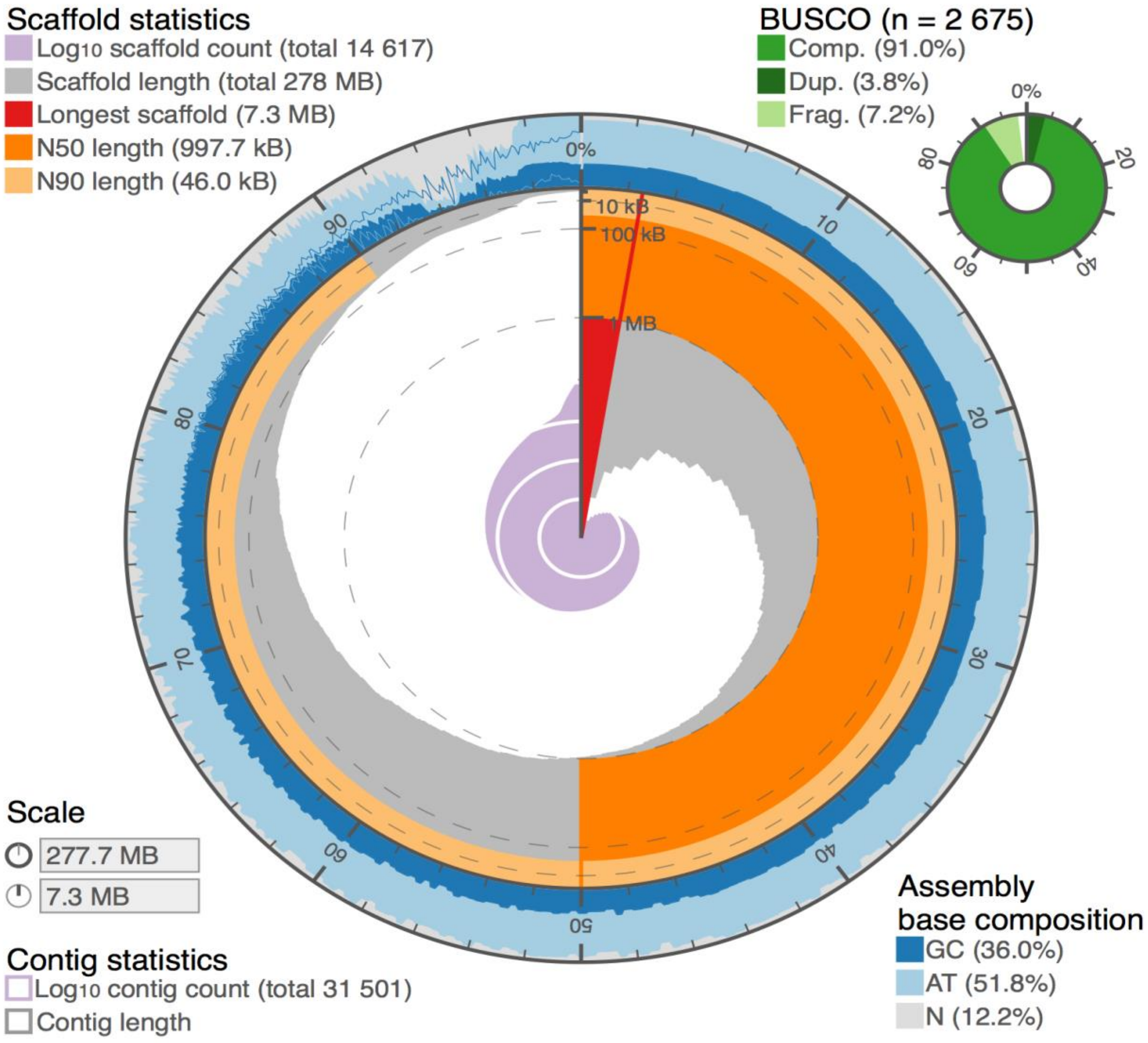

Figure 2. Taxonomic distribution of best blastP hits of *F. exsecta* proteins to the nonredundant (nr) protein database (E < 10−5).

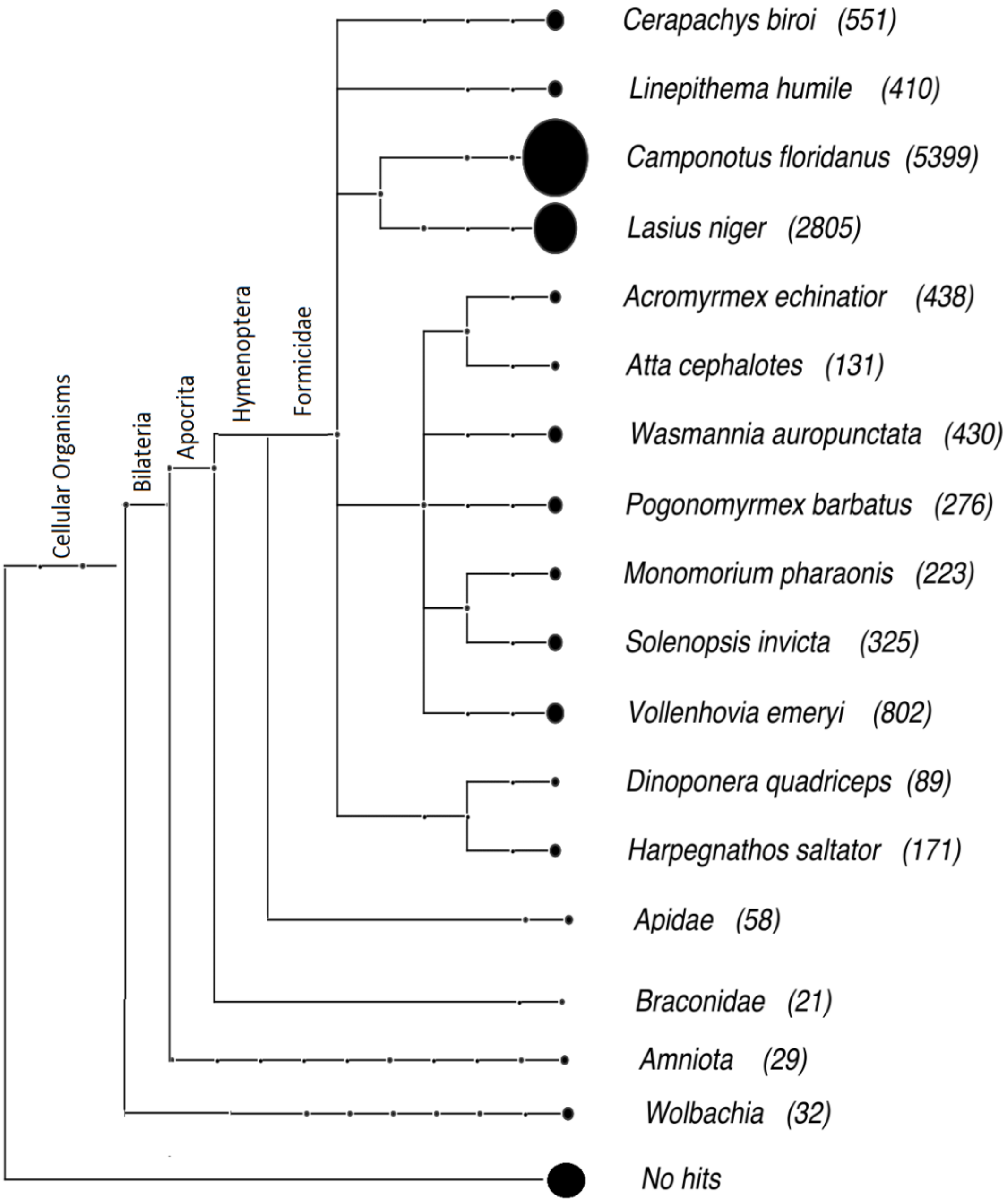

Figure 3. Venn diagram showing the distribution of gene families (orthologous clusters) among five ant species including closely related three members of subfamily Formicinae (*Formica exsecta, Camponotus floridanus, Lasius niger)* and other two distinctly related ants *(Solenopsis invicta* and *Cerapachys biroi)*.

-

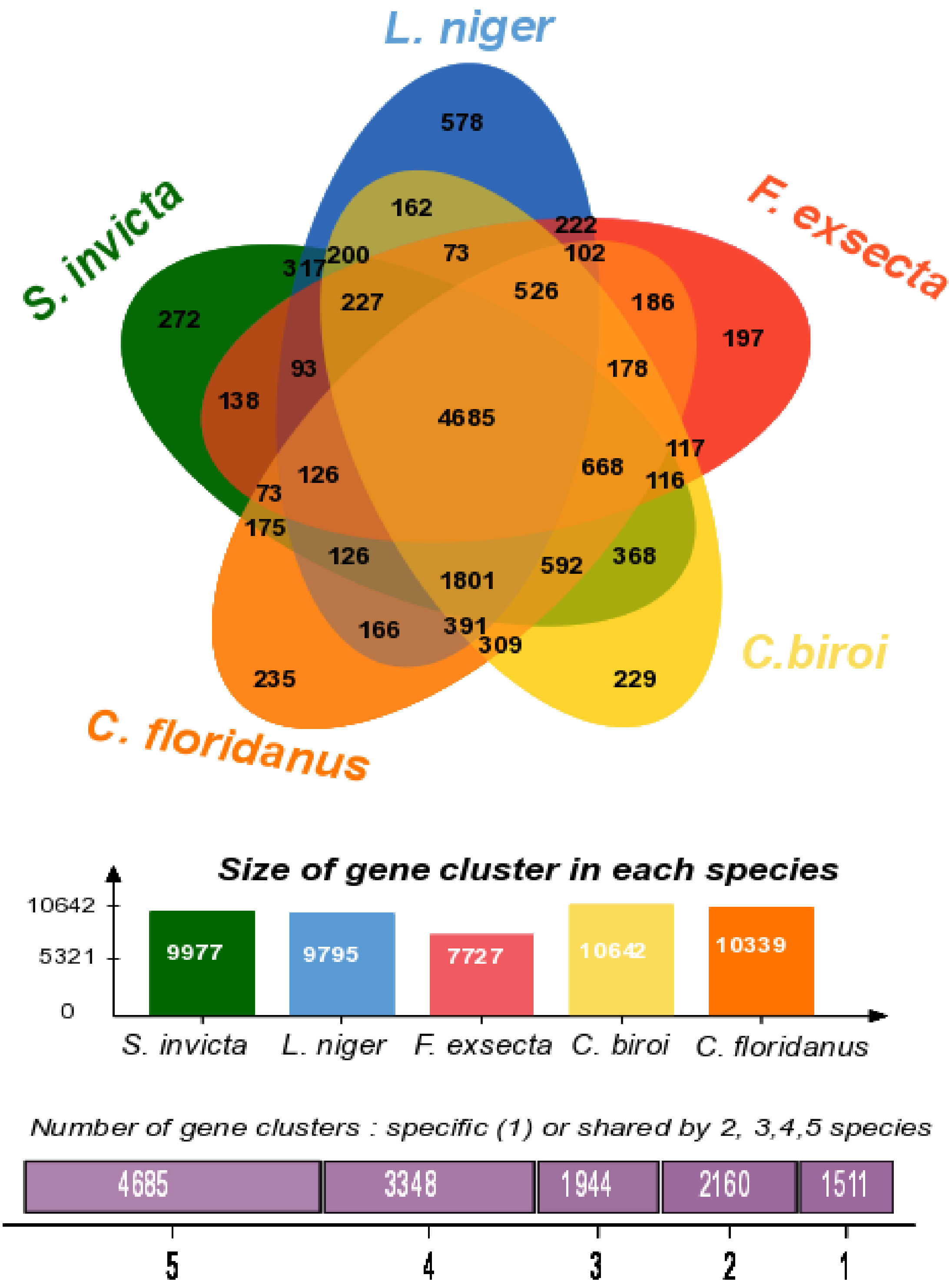

Figure 4. Phylogeny of the *Wolbachia* supergroups A, B, and C strains with the newly assembled wFex genome. Phylogenetic reconstructions based on individual analyses of 35 core gene of 25 *Wolbachia* strains. The numbers at the node indicate the posterior probabilities obtained from Bayesian phylogenetic analysis. Each Supergroup is labeled with letters A-C.

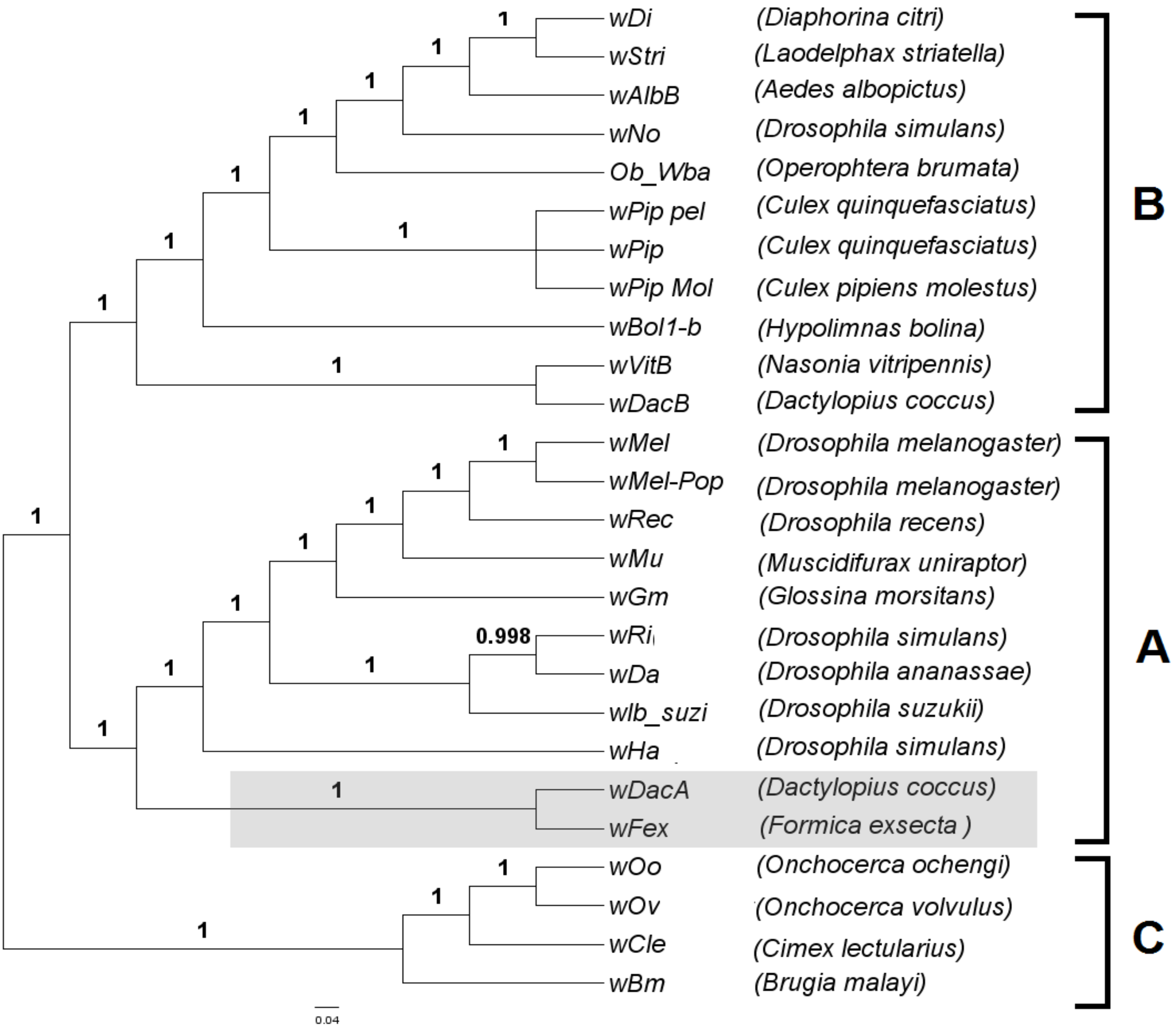
wDi (Diaphorina citri)
wStri (Laodelphax striatella)
wAlbB (Aedes albopictus)
wNo (Drosophila simulans)
Ob_Wba (Operophtera brumata)
wPip pel (Culex quinquefasciatus)
wPip (Culex quinquefasciatus)
wPip Mol (Culex pipiens molestus)
wBol1-b (Hypolimnas bolina)
wVitB (Nasonia vitripennis)
wDacB (Dactylopius coccus)
B
wMel (Drosophila melanogaster)
wMel-Pop (Drosophila melanogaster)
wRec (Drosophila recens)
wMu (Muscidifurax uniraptor)
wGm (Glossina morsitans)
wRi (Drosophila simulans)
wDa (Drosophila ananassae)
wIb_suzi (Drosophila suzukii)
wHa (Drosophila simulans)
wDacA (Dactylopius coccus)
wFex (Formica exsecta)
A
wOo (Onchocerca ochengi)
wOv (Onchocerca volvulus)
wCle (Cimex lectularius)
wBm (Brugia malayi)
C
0.998
1
0.04

Figure 5. Venn diagram displaying overlap in orthologous gene among four *Wolbachia* species including newly assembled wFex strain and previously reported wDac, wRi, wMel strains.

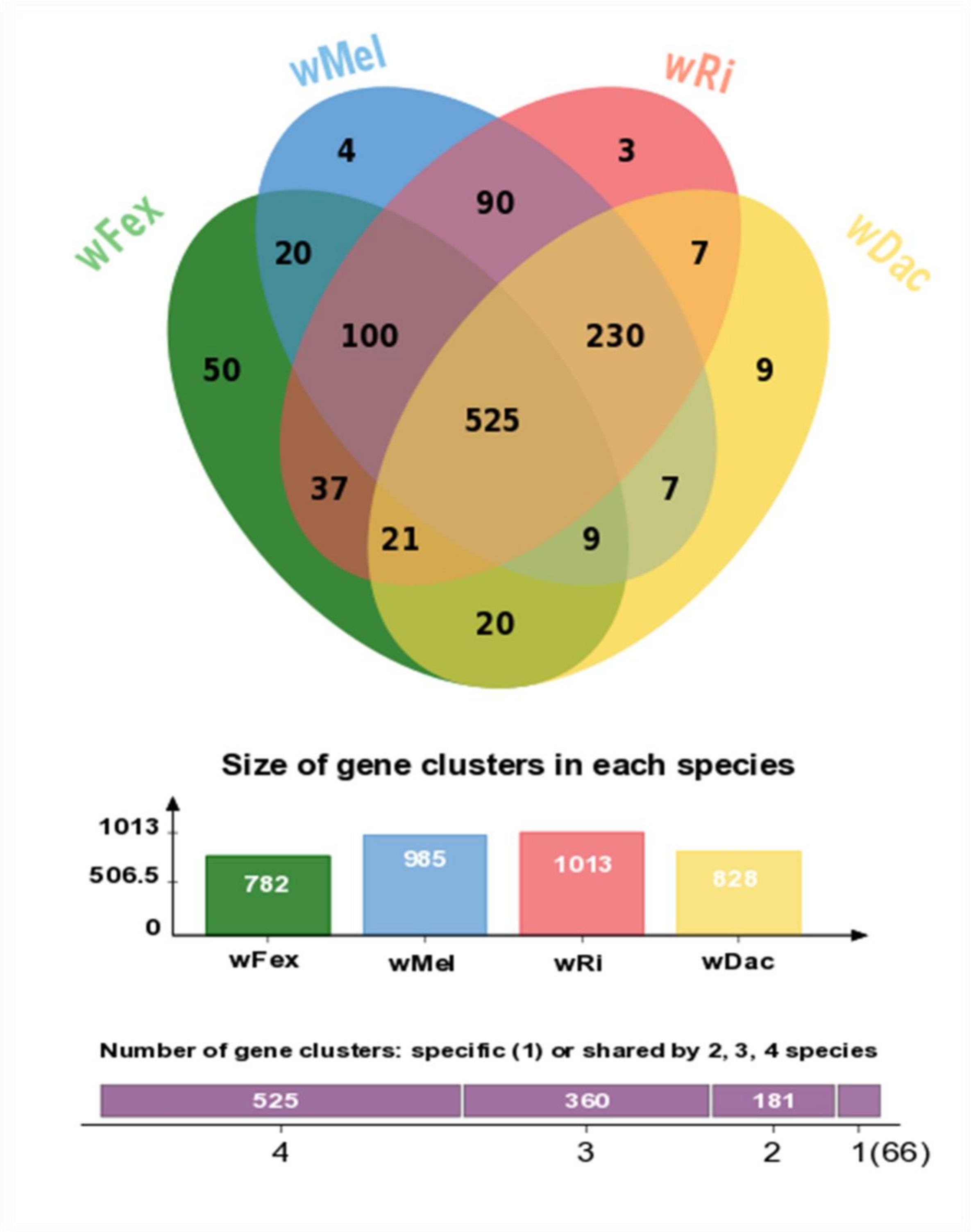

Table 1. Summary statistics of raw sequencing data before and after filtering the reads which was further used for the genome assembly. "Coverage depth" was calculated based on the assembled genome size (500 Mb).

Insert Size	Pair reads Length (bp)	Raw		After Filter	
		Total Data (G)	Sequence coverage (X)	Total Data (G)	Sequence coverage (X)
170bp	100 bp	22.68	45.36	20.96	41.93
500bp	100 bp	8.54	17.08	7.34	14.69
800bp	100 bp	8.84	17.69	5.14	10.29
2kb	100 bp	13.23	26.46	7.05	14.10
5kb	100 bp	14.51	29.02	4.74	9.49
10kb	100 bp	11.77	23.53	5.51	11.02
20kb	100 bp	20.40	40.81	2.91	5.81
Total	--	99.97	199.95	53.66	107.32

Table 2. Genome assembly statistics for *F. exsecta* and its *Wolbachia endosymbiont*.

Genome Assembly Stats	***Formica exsecta* Genome**	**FE *Wolbachia* endosymbiont Genome**
Total length	277719392 (277 MB)	3096460 (3.09 MB)
Total contigs	14617	69
Contigs (>= 1000 bp)	3136 (98.24% genome)	68(99.97% genome)
Contigs (>= 50000 bp)	545 (89.59% genome)	22(75.48% genome)
N50:	997654 bp	104167 bp
N75:	318356 bp	54296 bp
L50:	73	11
L75:	185	22
GC (%)	36.00	35.13

Table 3. BUSCO quality metrics for the *F. exsecta* genome and the *Wolbachia* endosymbiont *of F. exsecta (wFex)* genome assembly.

BUSCO metric	***F. exsecta* genome**				***FE Wolbachia endosymbiont Genome* (wFex)**	
	Eukaryota	**Insecta**	**Arthropoda**	**Hymenoptera**	**Bacteria**	**Proteobacteria**
Complete	299 (98.7%)	1634(98.5%)	2549(95.29%)	4249 (96.2%)	107 (72.30%)	158 (71.49%)
Complete and single copy	283 (93.4%)	1572 (94.8%)	2446(91.44%)	4151 (94.0%)	35 (23.65%)	55 (24.88)
Complete and duplicated	16 (5.3%)	62 (3.7%)	103 (3.86%)	98 (2.2%)	72 (48.65%)	103 (46.60%)
Fragmented	1 (0.3%)	15 (0.9%)	195 (7.29%)	123 (2.8%)	9 (6.08%)	11 (4.97%)
Missing	3 (1.0%)	9 (0.6%)	34 (1.27%)	43 (1.0%)	32 (21.62%)	52 (23.52%)
Total	303 (100%)	1658 (100%)	2675 (100%)	4415 (100%)	148 (100%)	221 (100%)

Table 4. HGT inserts from *Wolbachia* present in the genome of *F. exsecta* with details of its length and position in the *F. exsecta* genome. The presence of similar insert regions in other eukaryote genomes is also shown.

***Wolbachia* gene name**	**HGT region in *F. exsecta***	**Length HGT (bp)**	**Transposon region near HGT**	**Transposon Name**	**Observed in other species**	**Other Host Species name with position of similar insertion**
Transposase	scaffold83: 2271642-2272117	475	scaffold83:2271642-2272117	transposase	Complete	*Vollenhovia emeryi (LOC105557741), Cardiocondyla obscurior (genes: scf7180001101632 and scf7180001108526), Diachasma alloeum (LOC107035412),Brugia pahangi (BPAG_contig0001587),*
ABC transporter ATP-binding protein, porphobilinogen deaminase,D-alanine--D-alanine ligase, DNA processing protein DprA , triose-phosphate isomerase	scaffold233: 1712452-1725498	13046	scaffold233:1714122-1714241	transposase	Partial (few gene region)	*Vollenhovia emeryi(NW_011967015.1,NW_011967060.1),Wasmannia auropunctata (scf7180000683207,scf7180000730160),Rhagoletis zephyria(NW_016158779.1),Planococcus citri(KF021963.1) ,Ctenocephalides felis (KC177865.1)*
DNA repair protein RadC,transposase,DNA ligase,ABC transporter permease,ATP-dependent protease La	scaffold574: 102007-116197	14190	scaffold574:105963-106483	transposase	Partial (few gene region)	*Vollenhovia emeryi (LOC105557101,NW_011966940.1,NW_011966751.1), Monomorium pharaonis (scf7180001140281), Rhagoletis zephyria (LOC108377626),Parasteatoda tepidariorum(LOC107444616, LOC107450900)*
probable carboxypeptidase,type IV secretion system,conjugal transfer protein TrbL,lysyl-tRNA synthetase,UDP-N-acetylmuramoylalanine-D-glutamate ligase	scaffold707: 1-38814	38813	scaffold707:35826-36154	Mariner Mos1 transposase (Ant origin)	Partial (few gene region)	*Vollenhovia emeryi (NW_011966954.1,NW_011966496), Wasmannia auropunctata (scf7180000735528), Brugia pahangi (BPAG_contig0000608, BPAG_scaffold0000225)*
DNA methylase, Ankyrin repeat domain protein, regulatory protein RepA,site-specific recombinase, cytochrome b-like	scaffold741: 1-47265	47264	scaffold741:54020-54482, scaffold741:52587-52910	IS110 family transposase, Integrase	Partial (few gene region)	*Vollenhovia emeryi (LOC105557561,NW_011966954.1,NW_011967060.1,NW_011967015.1),Wasmannia auropunctata (LOC105460331 , scf7180000733651), Drosophila ananassae (WD_0580 gene)*