Student: *André Bourbonnais (andbou95@gmail.com)*
Supervisor: *Rachel Steward (rachel.steward@biol.lu.se)*
Department: *SPACE / Runemark Lab*
Project Title: Metagenome-Assembled Genome of the Obligate Symbiont *Candidatus Stammerula tephritidis* During Host Plant Shift in the Ecological Divergence of its host, *Tephritis conura*.

Abstract

Insect herbivores exhibit remarkable diversity, with most species specializing on specific host plants. The intricate relationships between these herbivores and their host plants are strong drivers of insect diversification and speciation, particularly during expansions in diet breadth. Although the influence of microbial symbionts on these evolutionary processes is acknowledged, it remains underexplored. Symbionts can have significant positive and negative impacts on their hosts, including supplying essential nutrients, defending against parasitoids, causing hybrid infertility, and facilitating diet expansion by providing key biosynthetic functions that ease the challenges of exploiting novel niches. However, it is still unclear whether symbionts act as leading or adaptive factors in host divergence when adapting to new ecological niches.

The tephritid fruit fly, *Tephritis conura*, is undergoing ecological divergence as it adapts to a novel thistle, *Cirsium oleraceum*, resulting in two host races: the ancestral *C. heterophyllum* specialist and the novel *C. oleraceum* specialist. While the ecological divergence of *T. conura* is well-documented, the role of its maternally transmitted symbionts, *Candidatus Stammerula tephritidis*, remains largely unexplored. In this study, whole-metagenomic sequencing was facilitated by using whole-genome sequencing data from *T. conura* to assemble the metagenome-assembled genome (MAG) of the unculturable symbiont. This novel MAG, meeting the high-quality criteria of MIMAG guidelines, was successfully assembled, annotated, and compared across different populations in a preliminary analysis of the symbionts relative abundance.

This resulting 1.5 megabase pair MAG, consisting of 11 contigs and 1.262 protein-coding genes, provides a robust foundation for exploring the role of symbionts in ecological divergence. This research lays the groundwork for future studies aimed at unraveling the complexities of symbiont-medated evolution in insect-plant interactions.

Introduction

Insects represent more than 50% of all animal species, making their immense diversity one of the captivating aspects of biology (Adler & Foottit, 2009), particularly, when it comes to understanding the intricacies of how such diversity arises and is maintained. Approximately 90% of herbivore insect species are specialists, utilizing only a few plant families as mating locations and food sources throughout their life cycles (Murphy & Loewy, 2015). This strong dependence of herbivorous insects on their host plant creates numerous distinct ecological niches, often leading to adaption to the challenges of their host plants, and often diet specialization, and ultimately diversification of insect lineages associated with different host plants (Ehrlich & Raven, 1964; Janz et al., 2006). This process of diversification has been tested at the macroevolutionary scale, showing increased rates of diversification after switches to new host plants (Edger et al., 2015; Wheat et al., 2007), and at the microevolutionary scale, showing rapid divergence and isolation of lineages after switching onto new hosts plants (Bush, 1969; Feder et al., 2003; Steward et al., 2024). Besides the selective pressures from the host plant (bottom-up forces), the top-down forces of predators and parasitoids are increasingly recognized as drivers of diversification in insect herbivores (Vidal & Murphy, 2018). How the microbial communities colonizing insect herbivores, their host plant and their predators, contribute to processes of divergence and speciation remains a poorly understood piece of this multi-trophic puzzle.

Insect microbiomes can mediate plant-insect interactions and their evolutionary consequences. The hologenome concept of evolution views hosts and their microbiome as unified entities (holobionts) where their combined genetic makeup generates a synergistic genome (hologenome; Rosenberg & Zilber-Rosenberg, 2016). For example, Rudman et al. (2019) found that manipulation of microbiome composition can drive rapid evolution in *Drosophila melanogaster*, as the different compositions led to divergence in allele frequencies across the genome. However, due to the complexity of the hologenome, and the sensitivity of the microbiome to generational and environmental variation, in most cases it is difficult to differentiate the effects of microbiomes from other ecological factors that might affect host genome evolution, and to determine which core microbes have the strongest effects (Douglas & Werren, 2016; Rudman et al., 2019; Shapira, 2016).

On the other hand, vertically transmitted microbes are likely to show much stronger signatures of coevolution with their insect hosts and offer an opportunity to explore the reciprocal consequences of microbial and host genome evolution during host shifts. Many herbivorous insects have acquired and co-evolved with bacterial and fungal symbionts as these microbes can offer primary and secondary metabolites that help insect species utilize new host plants (Brucker & Bordenstein, 2012; Douglas, 2014; Sudakaran et al., 2017). For example, Wu et al. (2006) assembled and analyzed genomes of two vertically transmitted symbionts *Baumannia cicadellinicola* and *Sulcia muelleri* of the xylem-feeding sharpshooter, *Homalodisca coagulata*, finding that these bacteria complement their hosts genome by producing critical primary metabolites. *B. cicadellinicola* supplies *H. coagulata* with vitamins and co-factors and *S. muelleri* contributed with essential amino acid biosynthesis, which due to the poor nutritional profile of xylem, are vital for growth, development and diet breadth of *H. coagulata* and are critical for *H. coagulata* to exploit their specialized host plant niche (McCutcheon & Moran, 2007; Wu et al., 2006). These nutrients are vital for the growth, development and diet breadth of the host, but they represent only two of the microbes in the microbiome of *H. coagulate* (D. Wu et al., 2006). On a macroevolutionary scale, both *B. cicadellinicola* and *S. muelleri* have been shown to have coevolved with sharpshooters due to the poor nutritional profile of xylem (Takiya et al., 2006). Similarly, Salem et al. (2020) discovered that the pectinolytic phenotype of the tortoise beetles (Cassidinae) is derived from their obligate symbiont *Candidatus Stammera capleta* (*Stammera*). While all *Stammera* strains produce an enzyme to degrade homogalacturonan, only one strain also produces an enzyme to break down rhamnogalacturonan I. This variation in enzyme production is reflected in the beetle's phylogeny and diet breadth. Beetles harboring the *Stammera* strain with both enzymes have evolved to specialize across three distinct eudicot clades (lamiids, campanulids, and fabids), compared to those associated with *Stammera* strain that only produces a single pectinolytic enzyme, which are limited only to lamiids. Besides nutrition, symbionts have been found to provide a range of other benefits, including defense against parasitoids (Oliver et al., 2003), mitigate of toxic compounds (Adams et al., 2013), synthesis of defensive compounds (Kellner, 2001) and alterered reproduction (Werren et al., 2008). Despite these clear mechanistic effects and evidence for codivergence between symbionts and their hosts (i.e., phylosymbiosis), it is still unclear whether the symbionts are mediating or responding to host plant shift and adaptation of their insect hosts. Investigating genomic change in a vertically transmitted symbiont will help to reveal how rapidly a symbiont evolves during initial steps of host plant shifts and host insect diversification.

The univoltine tephritid fruit fly, Tephritis conura and its heritable obligate symbiont, Candidatus Stammerula tephritidis (Stammerula), serve as an ideal model system to study the genomic changes in a symbiont during a novel host plant shift. Stammerula is held in sheltered extracellular compartments called crypts on the fly midgut, which separates the symbiont from the lumen, and

vertical transmission occurs during oviposition by smearing the eggs with the symbionts (Petri, 1909; Stammer, 1929). Despite, the higher chance of contact with the outer environment during oviposition and in the larval stage, Stammerula exhibits a significant phylosymbiosis with its host species, suggesting a coevolutionary relationship between the tephritid flies and Stammerula (Mazzon et al., 2008, 2010). While this coevolutionary relationship has been established at the species level, it is unclear whether the symbiont is a leader or follower during host shifts and incipient speciation.

Tephritis conura is a thistle specialist, mating on, laying its eggs in and feeding within the buds of its thistle host plant. In continental Europe, T. conura has recently, in evolutionary terms, diverged into two distinct host races: the ancestral Cirsium heterophyllum (CH) specialist and the novel C. oleraceum (CO) specialist (Seitz & Komma, 1984; Steward et al., 2024). Extensive research using allozyme markers, haplotypes, morphology, fitness performance, and whole-genome analysis has been conducted to show that this shift has been accompanied by both strong reproductive isolation and genomic divergence (Diegisser et al., 2007, 2008; Diegisser, Johannesen, et al., 2006; Diegisser, Seitz, et al., 2006; Nilsson et al., 2022; Seitz & Komma, 1984; Steward et al., 2024). Interestingly, in Scotland T. conura is oligaphagous (hereon referred to as a "generalist"), utilizing both C. heterophyllum and a third plant, Cirsium palustre (CP), as hosts. Yet, the potential influence of Stammerula in either the divergence of specialists or the shift to generalists remain unexplored, as neither the genome nor the functional role of Stammerula is known, leaving a significant gap in understanding how this symbiont might influence or be influenced by the novel host plant adaptation.

To bridge this gap, the genome of *Stammerula* is required but due to the obligate symbiont aspect, *Stammerula* is unculturable *in situ*. By employing a whole-metagenome sequencing (WMS) approach, I mined existing whole-genomes sequencing (WGS) data from entire adult *T. conura* flies to assemble and annotate the metagenome-assembled genome (MAG) of *Stammerula*. WMS defines the process of sequencing both host and microbiome, to generate a metagenome representing the microbiota of a host or environment. At its core, the reads are assembled into contigs which are then clustered (binned) based on unsupervised metric, supervised metrics or a combination of them both. These clusters (bins) represent putative genomes that in practice represent one microbe. These bins can then be refined further by discarding and adding contigs by analysing their coverage, assembly statistics, gene functions, taxonomy, and the number of single-copy genes to determine completeness and contamination. If the bin meets a high enough standard, they can be defined as a metagenome-assembled MAG (Bharti & Grimm, 2021; Meyer et al., 2022).

In this study, I successfully assembled, identified, and annotated the highly complete 1.5 megabase pair (Mbp) MAG of *Stammerula* spanning 11 contigs and containing 1262 protein-coding genes. I further compared the relative abundance of *Stammerula* between *T. conura* host races and between specialists and generalist populations for a preliminary understanding of how this symbiont may contribute or respond to the diet breadth evolution of its host. Finally, I contrast these results with the MAGs of more widely studied insect symbionts, including *Wolbachia pipientis*, to better place the *Stammerula* results in context. Having access to the highly complete *Stammerula* MAG sets a foundation for future studies into the role of *Stammerula* in the host plant shift of *T. conura* to help clarify how symbionts contribute to or are affected by their host's adaptation to new ecological niches.

Methods

Tephritis conura collection and sequencing

I used previously reported WGS samples of *T. conura.* For sampling and sequencing details, see Steward et al. (2024). Briefly, adult male *Tephritis conura* flies were collected from four CH specialist populations and four CH specialist populations throughout Scandinavia, including populations where host plants grow allopatrically and sympatrically on both sides of the Baltic Sea (Figure 1; Table 1). DNA was extracted from whole bodies of 96 flies and sequenced on Illumina HiSeqX lanes to acquire 2x150 bp paired-end reads. Additionally, long read sequences were generated for one male fly from the allopatric CH population in the west (CHST) using PacBio sequencing. Eleven flies were identified as recent hybrids between the host races (Steward et al., 2024) and were excluded from the analyses, for a total of 84 samples across these eight populations (Table 1). Whole bodies from an additional 31 flies feeding on CH and a third plant, CP, in Scotland were sequenced using Illumina HiSeqX and included in the metagenome analysis (Table 1). Data were accessed from a storage project on Uppmax (Crex).

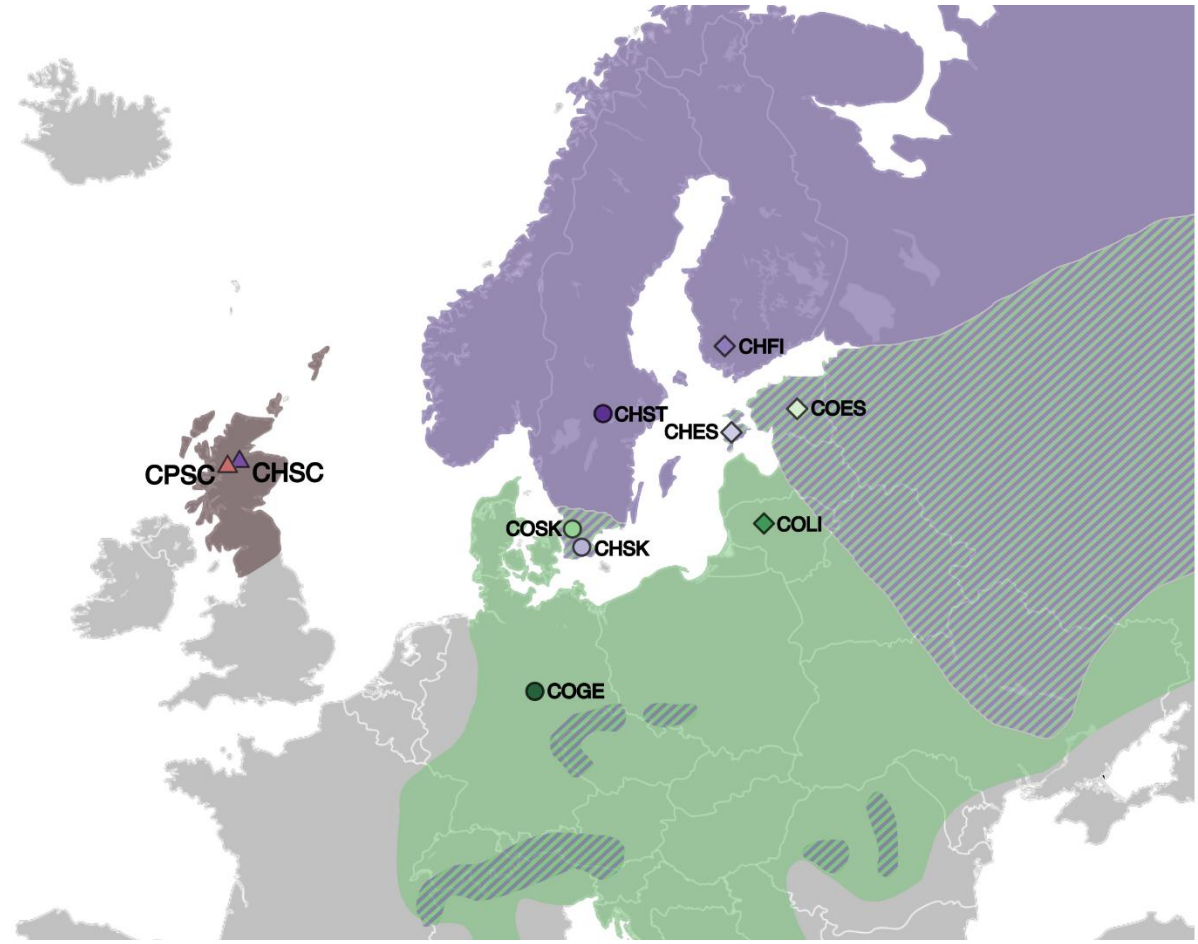

Figure 1. Geographical overview of sampled *Tephritis conura* populations. Triangles represent the Scottish generalists; circles represent the West region and diamonds represents the East region. The shaded regions represent sympatry, whereas full colour indicates allopatry (besides Scotland).

Table 1. Metadata describing each population and number of short read sequences after trimming, decontamination and merging. Populations include flies feeding on Cirsium heterophyllum (CH), C. oleraceum (CO), or C. palustre (CP) host plants in regions (Continental or in Scotland) where these hosts are allopatric (Allo.) or sympatric (Symp.).

Population	Plant	Range	Region	Samples	Raw Sequences	Decontaminated Sequences	Perc after decon.
CHST	*CH*	Allo.	Cont., West	9	561 576 652	508 927 504	0.342
CHSK	*CH*	Symp.	Cont., West	12	783 621 980	713 830 118	0.706
COSK	*CO*	Symp.	Cont., West	12	873 256 416	796 911 528	0.474
COGE	*CO*	Allo.	Cont., West	11	776 954 474	700 340 950	0.502
CHFI	*CH*	Allo.	Cont., East	9	589 558 902	535 449 122	0.512
CHES	*CH*	Symp.	Cont., East	10	797 488 918	725 338 808	0.439
COES	*CO*	Symp.	Cont., East	10	941 060 648	853 153 432	0.266
COLI	*CO*	Allo.	Cont., East	11	788 987 742	714 803 726	0.538
CHSC	*CH*	Symp.	Scotland	16	1 401 613 286	1 293 525 398	0.252
CPSC	*CP*	Symp.	Scotland	15	1 167 605 398	1 089 779 992	0.146

Quality control of the FASTQ files.

To ensure high-quality sequencing data, raw Illumina reads were trimmed and deduplicated using fastp v. 0.23.4 (Chen, 2023) by removing adapter sequences, discarding reads with average Phread score below 20, and excluding reads shorter than 36 base pairs (bp). Host decontamination was then performed by aligning both the short and long reads against the assembled *T. conura* genome (Steward et al., 2024). However, prior to alignment we masked microbial content previously identified with Kraken2 and discarded all contigs that were shorter than 50 kilobase pairs (Kbp), which are more prone to assembly errors. The short reads were aligned to the reference genome with BWA-MEM v. 0.7.17 (Li, 2013) and only read pairs where both the forward and reverse did not align were kept. The resulting BAM files containing the unmapped read pairs were then converted back to FASTQ format using the `bamtofastq` module from BEDTools v. 2.31.1 (Quinlan & Hall, 2010) to obtain host decontaminated short reads. Unlike the Illumina short reads, no trimming was done on the long PacBio reads, which were instead aligned directly using Minimap2 v. 2.26 (Li, 2018, 2021) with the `map-hifi`. Unmapped reads were then converted back to FASTQ as described above.

Co-assembled metagenomes

Three separate metagenomes were co-assembled using reads from different population samples: one representing all 10 population samples (all-pop), and two for the single populations CHST and COGE. This was accomplished by first merging all the forward and reverse reads from the relevant samples to create single forward and reverse FASTQ files for each selected co-assembly. The metagenomes were then assembled using SPAdes v. 3.15.5 (Prjibelski et al., 2020) with a k-mer size set of 21, 33, and 55 but utilizing different modules to accommodate the specific read data. The metaSPAdes v. 3.15.5 (Nurk et al., 2017) module was employed for the COGE co-assembly as it only consisted of short reads, while hybridSPAdes v. 3.15.5 (Antipov et al., 2016) was used to handle both the long and short reads for the all-pop and CHST co-assemblies.

Achieving high-quality MAGs is a non-trivial task as it relies on a combination of supervised and unsupervised methods to cluster contigs into putative genomes where each chosen program has strengths and weaknesses. To manually confirm and inspect the resulting MAGs, the program Anvio (Eren et al., 2015; Kiefl et al., 2023) integrates numerous command-line tools that helps validate the resulting MAGs through interactive visualization of the clustering and coverage data. Following the Anvio workflow v. 8 (Eren, 2016), the metagenomes were first reformatted to standardize the contig headers and to exclude contigs smaller than 2500 bp. Contig databases were then created for each metagenome using `anvi-gen-contigs-database`, which calculates contig statistics and identifies open

reading frames using Prodigal v. 2.6.3 (Hyatt et al., 2010). The identification of single-copy genes (SCGs) and functional annotation was performed using `anvi-run-hmms` and `anvi-run-ncbi-cogs` using the COG20 database. Finally, corresponding profile databases for each metagenome were generated by aligning the short reads from each population against the three reformatted metagenomes to acquire coverage statistics.

Metagenome-assembled genomes

Using metaWRAP v. 1.3.2 (Uritskiy et al., 2018), the three metagenomes were initially binned using three different tools: CONCOCT (Alneberg et al., 2014), MaxBin2 (Y.-W. Wu et al., 2016), and MetaBAT2 (Kang et al., 2019). The resulting bins were then refined using metaWRAP's `bin_refinment` module, applying completeness and contamination thresholds of 50% and 10%, respectively, to generate a consensus set of high-quality bins. This consensus is based on the contigs consistently clustering together across the bin results from CONCOCT, MaxBin2, and MetaBAT2. Additionally, because metaWRAP relies on an earlier version of CheckM (Parks et al., 2015; Uritskiy et al., 2018) to assess bin completeness and contamination, I also evaluated the refined bins using Anvio v. 8, CheckM2 v. 1.0.1 (Chklovski et al., 2023), and BUSCO v. 5.5.0 (Manni et al., 2021) with the bacteria_odb10 database to ascertain the quality of the refined bins. All these tools assess bin quality based on SCGs, but they differ in their underlying databases and algorithms. Furthermore, to check the quality of the assembled MAG, that was classified as *Wolbachia pipientis*, I aligned it with Minimap2 v. 2.26 (Li, 2018, 2021) to an existing reference genome of Wolbachia pipientis (GCF_014107475.1) and generated a dot plot with the R script dotPlotly (Poorten, 2017/2024).

Finally, GTDB-Tk v. 2.4.0 (Chaumeil et al., 2022) was used to classify the taxonomy of each bin, and bins that had a completeness higher than 50% were annotated using Bakta v. 1.9.3 (Schwengers et al., 2021). Together, these metrics were used to generate a list of putative high- and medium-quality MAGs accordingly to the guidelines of MIMAG (Bowers et al., 2017). The corresponding set of refined MAGs was then imported as a collection to the respective contig databases for manual assessment by using ´anvi-interactive´ and ´anvi-refine´.

Finding the *Stammerula* MAG

To identify the MAGs representing *Stammerula,* I used the published 16S rRNA sequences of *Stammerula* (EF469613.1, EF469615.1, EF469618.1) reported by Mazzon et al. (2008). These sequences were used as queries in a local BLAST search against a custom database generated from all refined bins using BLAST+ v. 2.15.0 (Camacho et al., 2009) with default parameters. Additionally, I included from NCBI GenBank three *Candidatus Erwinia dacicola* 16S sequences (AJ586620.2, FM958431.1, GQ478378.1) and two *Wolbachia* spp. 16S sequences from GCF_018454455.1 (located on NZ_JADCND010000050.1 and NZ_JADCNE010000036.1) in the query file for comparison and as an outgroup. Hits were assessed based on percentage identity, number of mismatches, e-value, length of query and length on matching sequence.

Exploration of Read Recruitment Data for *Stammerula* and *Wolbachia* MAGs

To evaluate the relative abundance of *Stammerula* and *Wolbachia* across different populations, I used the percent recruitment data derived from the contig and profile database of Anvio. Percent recruitment refers to the percentage of all mapped reads in a given sample that aligned with contigs within a specific MAG. For this study, the percent recruitment for the all-pop *Wolbachia* MAG (AWp7) and the CHST *Stammerula* MAG (CHSt1) across all merged samples were imported into R Studio (R Core Team, 2020). The corresponding metadata for each population were categorized by host plant

(HP), allopatry or sympatry (range), generalist or specialist (type), number of samples merged (n_samples), and region (geographical region).

Simple linear models were fitted to determine the differences in read recruitment between host plant within specialist populations and between specialist and generalist populations. Due, to the low sample size caused by the merger of reads from each population, the degrees of freedom were limited which is why only one predictor was used for the models. The dataset was accordingly subset into two groups: one containing both generalists and specialists, and the other containing only specialists.

The linear models were constructed as following:

$$CHSt1 \sim HP, data = specialists$$

$$CHSt1 \sim type, data = generalists\ and\ specialists$$

Similarly, to explore the factors affecting the abundance of AWp7, the same linear modelling approach was applied using data from the all-pop metagenome. Model fit was assessed using the adjusted R^2, and significance of the model predictors was asses using the effect size (coefficient estimate) and the p-value (alpha = 0.05).

Results

Co-assembled metagenomes

Out of the 8.6 billion raw short reads, only 17 million (0.2%) passed both the trimming and host decontamination processes, with an average length of 150 bp. A similar result was observed with the 4.7 million initial long read contigs where after applying the 50 Kbp threshold and host decontamination, only 1.58% (65982) of these contigs were retained. Nonetheless, the co-assembly of the all-pop metagenome resulted in a 29.5 Mbp metagenome composed of 1,891 contigs, with an N50 of 247, an L50 of 29.38 Kbp, and a longest contig of 331.4 Kbp. This complexity contrasts to the co-assembled metagenomes of the individual populations, CHST and COGE, which had fewer contigs (345 and 332, respectively), and total sizes of 5.2 Mbp and 2.8 Mbp. Although the CHST assembly had a slightly higher N50 value of 40 compared to COGE's N50 of 32, it also produced a more contiguous assembly, with the longest contig reaching 770.5 Kbp, whereas COGE's longest contig measured only 126.6 Kbp (Table 2).

Table 2. Co-assembled metagenome assembly statistics for metagenomes derived using read from all populations (all-pop), Allopatric *C. heterophyllum* specialist population (CHST), and sympatric *C. oleraceum* specialist population (COGE); Assembly statistics is after discarding contigs less than 2.5 Kbp.

Assembly	Contigs	Size	N50	L50	Max_contig	GC
all-pop	1 891	29 551 643	247	29 380	331 444	0.54
CHST	345	5 172 028	40	23 845	770 501	0.46
COGE	332	2 815 915	32	16 577	126 625	0.43

Putative MAGs

Following the MIMAG guidelines for high- and medium-quality MAGs, I defined a high-quality MAG as one with a completeness score above 90%, contamination less than 5%, and the presence of 23S, 16S, and 5S rRNA genes along with at least 18 tRNAs. Medium-quality MAGs were defined as having a completeness score above 50% and contamination below 10%, without the rRNA or tRNA requirements. The average nucleotide identity is reported if given.

From the all-pop metagenome, I identified two high-quality and five medium-quality MAGs, with CheckM2 completeness ranging from 63% to 100% and contamination ranging from 0% to 1.88% (Table X). The requirement for rRNA and tRNA genes downgraded three MAGs to medium-quality, despite their high completeness, as only two of the refined bins had CheckM2 completeness scores below 90% (70.75% and 63.0%) (Supplementary Table 2). Of the seven MAGs, six belonged to the Gammaproteobacteria clade, including four from the family Enterobacteriaceae, one from Pseudomonodaceae, and one from Xanthomonadaceae. Three MAGs, were taxonomically classified to the species level by GTDB-Tk as *Serratia ureilytica* (Average nucleotide identity): 98.66%, GCF_013375155.1), *Erwinia aphidicola* (ANI: 98.42%, GCF_024169515.1), and *Stenotrophomonas maltophilia* (ANI: 97.85%, GCF_013387485.1), with the remaining three classified to the genus level as *Pseudomonas*, *Erwinia*, and *Citrobacter* (JGM124 strain). The final MAG, referred to as AWp7, was classified within the Alphaproteobacteria clade as *Wolbachia pipientis* with an ANI of 98.01% (GCF_021378375.1) (Table 4). I compared AWp7 to an existing assembly of *W. pipientis* and found that it aligned with mid to high identity (>0.4) with little evidence of misassembly within contigs (Figure 2).

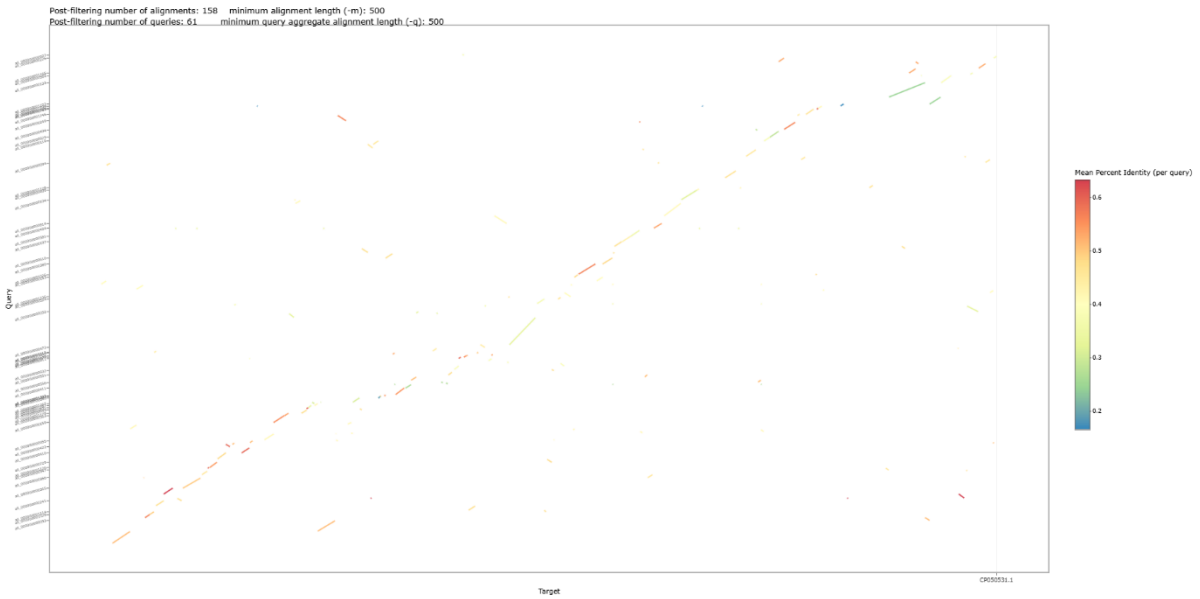

Figure 2. Dotplot visualization of the alignment of the *Wolbachia pipientis* metagenome-assembled genome (AWp7) from the co-assembled metagenome of all samples (all-pop) against the *W. pipientis* reference genome GCF_014107475.1 from NCBI GenBank.

In contrast to the all-pop metagenome, the CHST metagenome yielded two MAGs (CHSt1 and CHWp2), while only one MAG (COSt1) was recovered from the COGE metagenome. CHSt1 and COSt1 closely resembled bin 6 from the all-pop metagenome (ASt6) in terms of genome size, GC content, and taxonomic classification to the *Erwinia* genus. The completeness of these bins, as determined by CheckM2, was consistently high, approaching ~99% with zero contamination across all three bin sets. However, the number of contigs varied strongly, with CHSt1 containing 11 contigs, ASt6 containing 32 contigs, and COSt1 containing 91 contigs. CHWp2 was classified as *W. pipientis* with an ANI of 97.58% (GCF_021378375.1) and displayed a slightly higher CheckM2 completeness of 97.63% but also exhibited a higher contamination level of 1.96% compared to AWp7. Additionally, CHWp2 was more fragmented, consisting of 99 contigs compared to the 65 contigs observed in AWp7 from the all-pop metagenome.

Characterizing the *Stammerula MAG*

I identified CHSt1 to be the highest quality *Stammerula* MAG. This 1.5 Mbp MAG achieved the high-quality criteria with a low fragmented assembly comprising 11 contigs, GC content of 43%, and presence of all required rRNA and tRNA genes. Bakta identified 1262 protein-coding genes, including 43 tRNA genes and 12 rRNA genes (one 23S, six 16S, and five 5S rRNA genes) (Table 3). CheckM2, Anvio and BUSCO reported zero contamination but differed slightly in the reported completeness. CheckM2 assessed the completeness to 99.03%, Anvio reported 97.18% and BUSCO determined 96.0% completeness, with 0% duplication. Of the 124 BUSCOs, 119 were found complete, one was fragmented, and four were missing (Table 3).

Table 3. *Stammerula* MAG assembly, annotation and quality statistics

Population		**all-pop**	**CHST**	**COGE**
Bin ID		ASt6	CHSt1	COSt1
Contigs		34	11	91
N50		8	1	11
L50		64 899	770 501	38 359
Contig_max		142776	770 501	126 625
GC		0.43	0.43	0.44
Size		1 413 378	1 506 883	1 749 575
CDS_Dens.		0.744	0.733	0.755
Tot.CDS		1 142	1 262	1 589
23S rRNA		3	1	0
16S rRNA		4	6	0
5S rRNA		4	5	1
tRNA		44	43	37
ori		0	0	4
Bin		6	1	1
***Stammerula* 16S hits**		4	6	N/A
Species		*S. tephritidis*	*S. tephritidis*	*S. tephritidis*
CheckM2	**Completeness (%)**	98.81	99.03	99.07
	Contamination (%)	0.0	0.0	0.22
BUSCO (bacteria_odb10, n = 124)	**Completeness (C)**	119 (96.0%)	119 (96.0%)	119 (96.0%)
	Complete and single-copy (S)	119 (96.0%)	119 (96.0%)	119 (96.0%)
	Complete and duplicated (D)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	Fragmented (F)	1 (0.8%)	1 (0.8%)	1 (0.8%)
	Missing (M)	4 (3.2%)	4 (3.2%)	4 (3.2%)
Anvio	**Completeness (%)**	97.18	97.18	97.18
	Redundancy (%)	0.0	0.0	0.0

GTDB-Tk classified the MAG to the genus *Erwinia* but the local 16S BLAST search revealed that contigs 03, 04, 06, 168, and 240 had significant *Stammerula* 16S rRNA hits. Notably, contig 03 contained two 16S rRNA regions in both forward and reverse directions. All hits shared identical scores, e-values, and number of mismatches for each separate query (Supplementary Table 2). The identity percentage for the *Wolbachia* queries was 76.47% and 76.90%, with 276 and 281 mismatches. In contrast, *Erwinia dacicola* queries showed greater identity ranging from 94.95% to 95.23% and 62 to 64 mismatches.

Interestingly, the highest percentage identity for bin 1 was with the *Stammerula* 16S rRNA sequence from the *T. hyoscyami* (EF469613.1) isolate, showing only one mismatch and a percentage identity of 99.92%. The second highest percentage identity of 99.85% and two mismatches was with the *Stammerula* 16S rRNA sequence derived from the symbiont previously found in *T. conura* (EF469618.1; Mazzon et al., 2008). The *Stammerula* 16S rRNA sequence from *T. cometa* (EF469615.1) showed a percentage identity of 99.70% due to four mismatches (Supplementary Table 2). Similar BLAST result patterns were found for ASt6 but no hits were found for COSt1. Importantly, these remarkable percentage identities with the *Stammerula* 16S rRNA sequences strongly suggests that ASt6, CHSt1, and COSt1 are the MAGs of *Stammerula* from each metagenome, with CHSt1 being the optimal MAG due to its high contiguity and comparable completeness.

Characterizing the *W. pipientis* MAG

Both the CHST and all-pop metagenomes yielded *W. pipientis* MAGs with high bin quality, while no similar bins were identified from the COGE metagenomes. Although Anvio rated the same bin quality for both MAGs, discrepancies were found in the assessments by CheckM2 and BUSCO. CHWp2 exhibited higher completeness, with scores of 97.3% by CheckM2 and 83.9% by BUSCO, compared to 94.3% and 80.6% for AWp7 (Table 4). Moreover, CHWp2 contained four more complete and one more fragmented BUSCO gene than AWp7. However, this increased completeness came with a higher contamination, as CheckM2 reported 1.92% for CHWp2 compared to the reported 0.66% contamination for AWp7 (Table 4). GTDB-Tk classified both MAGs as *W. pipientis* with a high ANI of 98.01% and 97.58% for AWp7 and CHWp2, respectively. Importantly, AWp7 had fewer contigs (65 in AWp7 versus 99 in CHWp2), a lower N50 (13 vs. 19), and a higher L50 (24.9 Kbp vs. 17.4 Kbp) as reported in Table 4. Additionally, AWp7 qualified as a high-quality MAG, while CHWp2 was downgraded to medium-quality due to the absence of a 5S rRNA gene. Given the similar bin quality and ANI, I favored AWp7 as the better representative due to its lower fragmentation and higher MAG quality.

Table 4. *W. pipientis* MAG assembly, annotation, alignment quality with reference genome (ANI and alignment fraction, AF) and quality statistics.

Population	**all-pop**	**CHST**	**COGE**
Bin ID	AWp7	CHWp2	N/A
Contigs	65	99	N/A
N50	13	19	N/A
L50	24864	17351	N/A
Contig_max	80080	70370	N/A
GC	0.35	0.35	N/A
Size	1102802	1109950	N/A
CDS_Dens.	0.849	0.859	N/A
Tot.CDS	1098	1075	N/A
ANI (%)	98.01	97.58	N/A
Alignment Fraction	0.865	0.862	N/A
23S rRNA	1	1	N/A
16S rRNA	1	1	N/A
5S rRNA	1	0	N/A
tRNA	32	35	N/A

ori		1	1	N/A
Bin		7	2	N/A
Species		*W. pipientis*	*W. pipientis*	N/A
CheckM2	**Completeness (%)**	94.3	97.63	N/A
	Contamination (%)	0.66	1.96	N/A
BUSCO (bacteria_odb10, n = 124)	**Completeness (C)**	100 (80.6%)	104 (83.9%)	N/A
	Complete and single-copy (S)	100 (80.6%)	104 (83.9%)	N/A
	Complete and duplicated (D)	0 (0.0%)	0 (0.0%)	N/A
	Fragmented (F)	5 (4.0%)	4 (3.2%)	N/A
	Missing (M)	19 (15.4%)	16 (12.9%)	N/A
Anvio	**Completeness (%)**	90.14	90.14	N/A
	Redundancy (%)	0.0	0.0	N/A

***Stammerula* abundance across populations**

Including generalists, CHSt1 recruited an average of 31.48% of all the reads that mapped to the CHST metagenome. However, the majority of the aligned reads for each population did not map to either of the two MAGs in the CHST metagenome, as an average of 54.96% aligned to unbinned contigs. The linear model fitted for the specialist dataset revealed that CHSt1 abundance was significantly lower in CO populations with an effect of -13.432 percentage points (SE = 4.604, p = 0.0267, adj.R^2 = 0.518) compared to the CH populations (Figure 3).

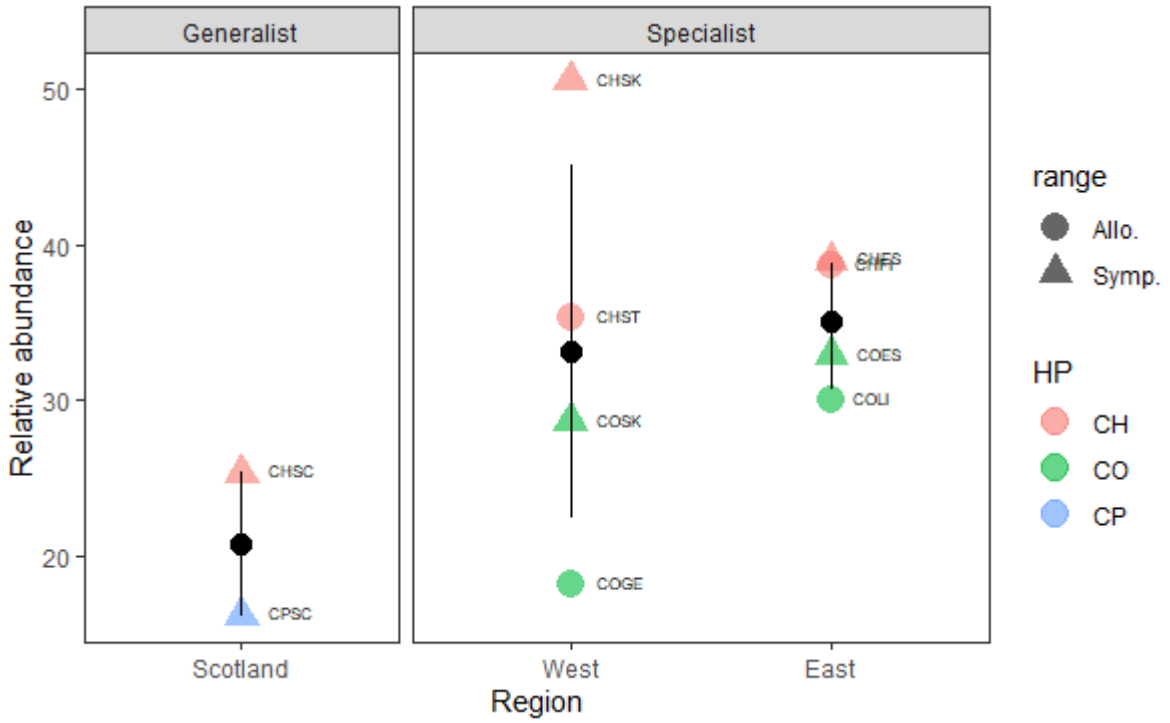

Figure 3. Relative abundance of the *Stammerula* MAG (CHSt1) derived from Anvio's read recruitment summary data from the merged reads of whole adult flies emerging from *Cirsium heterophyllum* (CH), *C. Oleraceum* (CO), and *C. Palustre* (CP) thistle buds, mapping to the CHST co-assembled metagenome. The error bar represents the 95% confidence interval and regional mean is represented by a black dot (for exact population locations, see Figure 1.

Similarly, CHSt1 was less abundant in generalists than in specialists, although this effect was marginally nonsignificant (estimate: -13.342, SE = 7.166, p = 0.0997, adj. R^2 = 0.215). Because the generalists eat both CH and CP, they are assumed to be more closely related to the continental CH specialists. However, when CO populations were removed, the negative effect of being a generalist increased to a significant -20.053 percentage points (SE = 5.719, p = 0.0247, adj.R^2: = 0.693). The results suggests that *Stammerula* derived from an allopatric CH specialist significantly differs in abundance with the generalist populations and CO specialists (Figure 3).

***Wolbachia pipientis* abundance across populations.**

Interestingly, of all the reads that aligned to the all-pop metagenome, an average of 61.0% were aligned to unbinned contigs. The bin that was determined as *Stammerula* (ASt6) had a mean of 25.33% whereas AWp7 had a median of 1.49% due to the skewness of the recruited reads from the generalist populations (Figure 3). An average of 39.8% of the generalist reads from Scotland aligned to AWp7 while the specialist populations ranged from 0.36% to 4.31%. As expected, the AWp7 model fitted to determine the difference between generalists and specialists explained 0.96% of the variation in the data showing that generalists have a significant higher abundance of AWp7 (estimate: 38.19, SE = 2.741, p < 0.0001, adj.R^2 = 0.956, Supplementary Table 3). As indicated by Figure 3, no significant difference between CH and CO specialist populations could be found (estimate: -0.3820, SE = 0.985, p = 0.712, R^2 = -0.138.

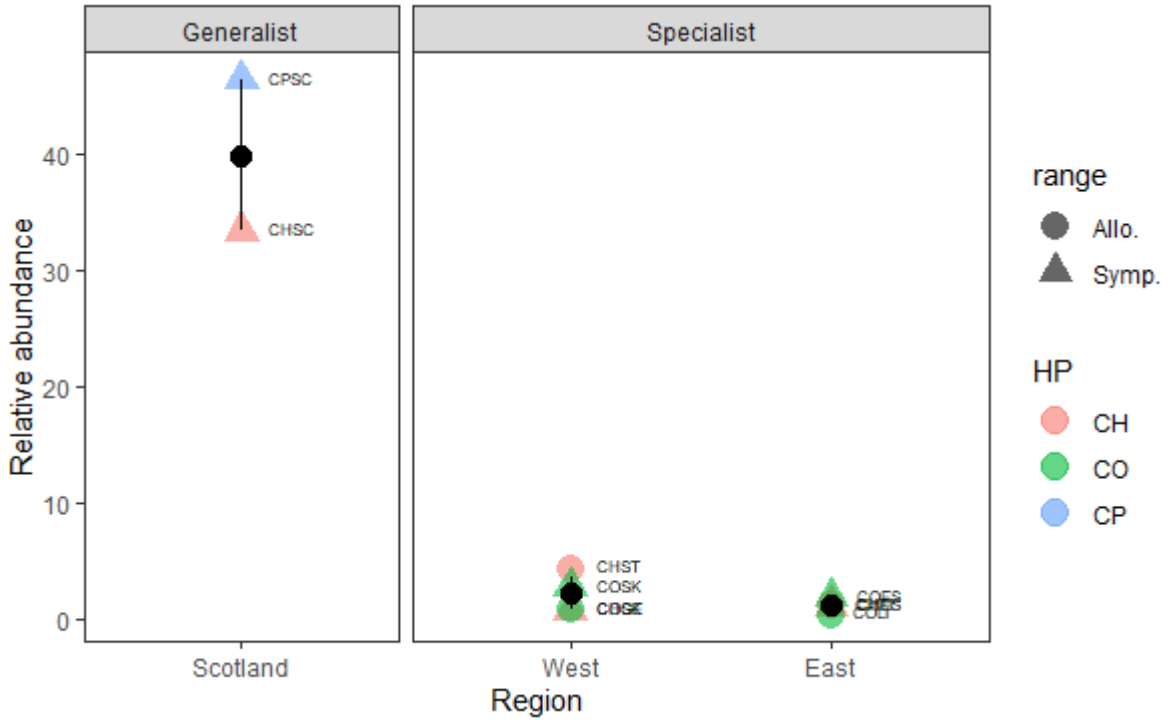

Figure 4. Relative abundance of the *Wolbachia pipientis* MAG (AWp7) derived from Anvio's read recruitment summary data from the merged reads of whole adult flies emerging from *Cirsium heterophyllum* (CH), *C. Oleraceum* (CO), and *C. Palustre* (CP) thistle buds, mapping to the metagenome using all samples (all-pop). The error bar represents the 95% confidence interval and regional mean is represented by a black dot (for exact population locations, see Figure 1).

Discussion

Tephritis conura and its symbiont *Candidatus Stammerula tephritidis* present an exceptional model for investigating the intricacies of ecological speciation driven by colonization of and adaptation to a novel host plant. The extensive body of research on *T. conura* provides a robust foundation for exploring the microbial factors that may influence insect divergence or how symbionts themselves are impacted by host shifts. The foundational work by Stammer (1929), who first identified that *T. conura* harbored a symbiont, and the subsequent studies by Girolami (1973), which detailed the extracellular crypts on the hindgut in *T. conura* that protects *Stammerula* from the environment laid the groundwork for investigating the coevolution of microbe and host. Recently, Mazzon et al. (2008, 2010) reared flies in controlled environments and managed to isolate and sequence the 16S rRNA region of *Stammerula,* showing strong phylosymbiosis of *Stammerula* and its tephritid fly hosts (Mazzon et al., 2008, 2010). Recent advancements in sequence technology and metagenomics provide an opportunity to build on these precedents, to analyze, annotate, and assemble the genome of this unculturable but clearly important symbiont.

The present study aimed to expand the existing knowledge and provide a solid foundation for future research into the evolutionary role of *Stammerula* in the ecological speciation of *T. conura* flies. To this end, I successfully assembled and annotated a low-fragmention, high-quality MAG in accordance with MIMAG guidelines, which I was able to confirm as a draft genome of Stammerula using existing 16S rRNA sequences (Mazzon et al., 2008). This is the first draft genome of this bacterial genus, marking a critical step toward a comprehensive understanding of this vertically transmitted symbiont.

One of the most noteworthy outcomes of this study was the assembly of the *Stammerula* MAG, CHSt1, which emerged as a high-quality MAG in accordance with MIMAG guidelines. The high completeness and low contamination were confirmed across CheckM2, BUSCO and Anvio, underscoring the feasibility of assembling high-quality genomes from metagenomic data. CHSt1 was derived from a hybrid assembly using both long and short reads, which contributed to its robustness. Despite the greater fragmentation observed in the short read assembled COSt1 MAG, key genomic features such as the number of annotated genes, GC content, genome size, and completeness levels were remarkably similar between COSt1 and CHSt1. The all-pop MAG, ASt6, which was also assembled using a hybrid approach, incorporated data from all samples, spanning geographic distances, ecological specializations, and sympatric and allopatric populations. The complexity in such a broad dataset inevitably influenced the success of the assembly. Nevertheless, the primary distinction between ASt6 and CHSt1 laid in the contig count, with CHSt1 comprising 11 contigs compared to the 34 contigs of ASt6, yet yielding comparable assembly statistics, MAG quality, and gene annotation. Not to mention *Stammerula* 16S rRNA gene hits which was absent in COSt1.

Even with an average passing rate of ~95% through the trimming process for all short reads, the host decontamination excluded 99.8% of the short reads. For individual fly samples, this left too few reads for a metagenome assembly and forced the merge of short read samples within in each population to increase the microbial coverage for the co-assembly. While this introduces additional complexity, the results demonstrate that it was still possible to bin contigs into a complete, low contamination MAG, as shown by COSt1. My results demonstrate the power of using existing mid- to low coverage short read WMS data from eukaryotic hosts to assemble MAGs of abundant, vertically transmitted symbionts. The addition of PacBio hifi long reads in a co-assembled metagenome considerably reduced the fragmentation and contamination of the Stammerula MAG, but had little effect on completeness.

The abundance analysis revealed subtle but significant patterns of potential divergence among *Stammerula* strains associated with different populations. Specifically, the CO specialists exhibited a notable reduction in CHSt1 abundance, with a decrease of 13.43 percentage points compared to CH specialists. Although CHST1 abundance did not differ significantly when comparing all specialist

populations with generalist populations, excluding CO populations from the model revealed CHSt1 was relatively less abundant in generalists. The observed results can be interpreted through several lenses, particularly when considering the underlying factors that could contribute to differences in read recruitment. Firstly, the limited sample size in this study presents a significant challenge in managing the influence of potential outliers, such as the CHSK and COGE populations. As CHSK exerts considerable influence when comparing generalist versus CH specialist populations. Secondly, the differences in *Stammerula* abundance might directly reflect the true bacterial abundance within the fly gut. Lastly, another plausible explanation lies in the genomic divergence between *Stammerula* derived from the different populations. The reduced abundance of CHSt1 in CO specialists could reflect the differences in genomic composition, which would affect the mapping rate.

In this study, I was also able to assemble a high-quality MAG of *W. pipientis*. Importantly, when aligned to existing genome assemblies for this microbe, the MAG was found to be very complete. This lends confidence to our Stammerula assembly. However, my analysis of relative abundance dawned more questions than it clarified. Why relative abundance was so much higher in generalists than specialists is unclear. We confirmed that this was not an artefact of mapping reads to AWp7, as similar abundance trends were found when reads were mapped to the CHWp2 MAG. It is possible that these generalist flies harbor much larger populations of Wolbachia. Alternatively, generalists may have additional Wolbachia strains that I was unable to assemble and whose sequence similarity may have led to contigs derived from these sequences being collapsed into a single MAG.

In conclusion, the *Stammerula* MAG presented here is a first step to understanding the role of this symbiont in host plant adaptation and speciation of *T. conura* flies. An important next step will be to re-assemble and curate the MAG into one contiguous genome with scaffolds and masking repetitive reads. Ongoing Nanopore long read sequencing of CH specialists, CO specialists and Scottish generalists will support this goal. Secondly, it is necessary to evaluate the functional profile of *Stammerula* to understand its physiological role for *T. conura*. Does it harbor defensive or immunological benefits or contribute complementary nutritional compounds or pathways? Understanding this will help answer if *Stammerula* improves or inhibits fly fitness. Finally, a complete genome will support population genomic and transcriptomic comparisons between CH and CO specialists and between specialists and generalists, allowing me to find the underlying cause of the abundance differences found in this study and to test whether *Stammerula* strains are diverging at the same time as their *T. conura* hosts.

Acknowledgements

The sincerest gratitude to Rachel Steward for her guidance and continuous support throughout my research project. Her expertise and mentorship helped shape the success of this project. Additionally, I want to extend my appreciation to Runemark Lab for providing resources, assistance and an environment to grow.

References

Adams, A. S., Aylward, F. O., Adams, S. M., Erbilgin, N., Aukema, B. H., Currie, C. R., Suen, G., & Raffa, K. F. (2013). Mountain Pine Beetles Colonizing Historical and Naïve Host Trees Are Associated with a Bacterial Community Highly Enriched in Genes Contributing to Terpene Metabolism. *Applied and Environmental Microbiology*, *79*(11), 3468–3475. https://doi.org/10.1128/AEM.00068-13

Adler, P. H., & Foottit, R. G. (2009). Introduction. In *Insect Biodiversity* (pp. 1–6). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781444308211.ch1

Alneberg, J., Bjarnason, B. S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., Lahti, L., Loman, N. J., Andersson, A. F., & Quince, C. (2014). Binning metagenomic contigs by coverage and composition. *Nature Methods*, *11*(11), 1144–1146. https://doi.org/10.1038/nmeth.3103

Antipov, D., Korobeynikov, A., McLean, J. S., & Pevzner, P. A. (2016). hybridSPAdes: An algorithm for hybrid assembly of short and long reads. *Bioinformatics*, *32*(7), 1009–1015. https://doi.org/10.1093/bioinformatics/btv688

Bharti, R., & Grimm, D. G. (2021). Current challenges and best-practice protocols for microbiome analysis. *Briefings in Bioinformatics*, *22*(1), 178–193. https://doi.org/10.1093/bib/bbz155

Bowers, R. M., Kyrpides, N. C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T. B. K., Schulz, F., Jarett, J., Rivers, A. R., Eloe-Fadrosh, E. A., Tringe, S. G., Ivanova, N. N., Copeland, A., Clum, A., Becraft, E. D., Malmstrom, R. R., Birren, B., Podar, M., Bork, P., … Woyke, T. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature Biotechnology*, *35*(8), 725–731. https://doi.org/10.1038/nbt.3893

Brucker, R. M., & Bordenstein, S. R. (2012). Speciation by symbiosis. *Trends in Ecology & Evolution*, *27*(8), 443–451. https://doi.org/10.1016/j.tree.2012.03.011

Bush, G. L. (1969). SYMPATRIC HOST RACE FORMATION AND SPECIATION IN FRUGIVOROUS FLIES OF THE GENUS RHAGOLETIS (DIPTERA, TEPHRITIDAE). *Evolution*, *23*(2), 237–251. https://doi.org/10.1111/j.1558-5646.1969.tb03508.x

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*(1), 421. https://doi.org/10.1186/1471-2105-10-421

Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P., & Parks, D. H. (2022). GTDB-Tk v2: Memory friendly classification with the genome taxonomy database. *Bioinformatics*, *38*(23), 5315–5316. https://doi.org/10.1093/bioinformatics/btac672

Chen, S. (2023). Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *iMeta*, *2*(2), e107. https://doi.org/10.1002/imt2.107

Chklovski, A., Parks, D. H., Woodcroft, B. J., & Tyson, G. W. (2023). CheckM2: A rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nature Methods*, *20*(8), 1203–1212. https://doi.org/10.1038/s41592-023-01940-w

Diegisser, T., Johannesen, J., & Seitz, A. (2006). The role of geographic setting on the diversification process among Tephritis conura (Tephritidae) host races. *Heredity*, *96*(5), Article 5. https://doi.org/10.1038/sj.hdy.6800821

Diegisser, T., Johannesen, J., & Seitz, A. (2008). Performance of host-races of the fruit fly, Tephritis conura on a derived host plant, the cabbage thistle Cirsium oleraceum: Implications for the original host shift. *Journal of Insect Science*, *8*(1), 66. https://doi.org/10.1673/031.008.6601

Diegisser, T., Seitz, A., & Johannesen, J. (2006). Phylogeographic patterns of host-race evolution in Tephritis conura (Diptera: Tephritidae). *Molecular Ecology*, *15*(3), 681–694. https://doi.org/10.1111/j.1365-294X.2006.02792.x

Diegisser, T., Seitz, A., & Johannesen, J. (2007). Morphological adaptation in host races of Tephritis conura. *Entomologia Experimentalis et Applicata*, *122*(2), 155–164. https://doi.org/10.1111/j.1570-7458.2006.00501.x

Douglas, A. E. (2014). Symbiosis as a General Principle in Eukaryotic Evolution. *Cold Spring Harbor Perspectives in Biology*, *6*(2), a016113. https://doi.org/10.1101/cshperspect.a016113

Douglas, A. E., & Werren, J. H. (2016). Holes in the Hologenome: Why Host-Microbe Symbioses Are Not Holobionts. *mBio*, *7*(2), 10.1128/mbio.02099-15. https://doi.org/10.1128/mbio.02099-15

Edger, P. P., Heidel-Fischer, H. M., Bekaert, M., Rota, J., Glöckner, G., Platts, A. E., Heckel, D. G., Der, J. P., Wafula, E. K., Tang, M., Hofberger, J. A., Smithson, A., Hall, J. C., Blanchette, M., Bureau, T. E., Wright, S. I., dePamphilis, C. W., Eric Schranz, M., Barker, M. S., … Wheat, C. W. (2015). The butterfly plant arms-race escalated by gene and genome duplications. *Proceedings of the National Academy of Sciences*, *112*(27), 8362–8366. https://doi.org/10.1073/pnas.1503926112

Ehrlich, P. R., & Raven, P. H. (1964). Butterflies and Plants: A Study in Coevolution. *Evolution*, *18*(4), 586–608. https://doi.org/10.2307/2406212

Eren, A. M. (2016, June 22). *Anvi'o User Tutorial for Metagenomic Workflow*. Meren Lab. https://merenlab.org/2016/06/22/anvio-tutorial-v2/

Eren, A. M., Esen, Ö. C., Quince, C., Vineis, J. H., Morrison, H. G., Sogin, M. L., & Delmont, T. O. (2015). Anvi'o: An advanced analysis and visualization platform for 'omics data. *PeerJ*, *3*, e1319. https://doi.org/10.7717/peerj.1319

Feder, J. L., Roethele, J. B., Filchak, K., Niedbalski, J., & Romero-Severson, J. (2003). Evidence for Inversion Polymorphism Related to Sympatric Host Race Formation in the Apple Maggot Fly, Rhagoletis pomonella. *Genetics*, *163*(3), 939–953. https://doi.org/10.1093/genetics/163.3.939

Girolami, V. (1973). *Reperti morfo-istologici sulle batteriosimbiosi del Dacus oleae Gmelin e di altri ditteri tripetidi, in natura e negli allevamenti su substrati artificiali*. https://www.semanticscholar.org/paper/Reperti-morfo-istologici-sulle-batteriosimbiosi-del-Girolami/21ffccbf8246719a42d90bbfccfdbff7fa3575c4

Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, *11*(1), 119. https://doi.org/10.1186/1471-2105-11-119

Janz, N., Nylin, S., & Wahlberg, N. (2006). Diversity begets diversity: Host expansions and the diversification of plant-feeding insects. *BMC Evolutionary Biology*, *6*(1), 4. https://doi.org/10.1186/1471-2148-6-4

Kang, D. D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., & Wang, Z. (2019). MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, *7*, e7359. https://doi.org/10.7717/peerj.7359

Kellner, R. L. L. (2001). Suppression of pederin biosynthesis through antibiotic elimination of endosymbionts in *Paederus sabaeus*. *Journal of Insect Physiology*, *47*(4), 475–483. https://doi.org/10.1016/S0022-1910(00)00140-2

Kiefl, E., Esen, O. C., Miller, S. E., Kroll, K. L., Willis, A. D., Rappé, M. S., Pan, T., & Eren, A. M. (2023). Structure-informed microbial population genetics elucidate selective pressures that shape protein evolution. *Science Advances*, *9*(8), eabq4632. https://doi.org/10.1126/sciadv.abq4632

Li, H. (2013). *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM* (arXiv:1303.3997). arXiv. https://doi.org/10.48550/arXiv.1303.3997

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li, H. (2021). New strategies to improve minimap2 alignment accuracy. *Bioinformatics*, *37*(23), 4572–4574. https://doi.org/10.1093/bioinformatics/btab705

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2021). BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution*, *38*(10), 4647–4654. https://doi.org/10.1093/molbev/msab199

Mazzon, L., Martinez-Sañudo, I., Simonato, M., Squartini, A., Savio, C., & Girolami, V. (2010). Phylogenetic relationships between flies of the Tephritinae subfamily (Diptera, Tephritidae) and their symbiotic bacteria. *Molecular Phylogenetics and Evolution*, *56*(1), 312–326. https://doi.org/10.1016/j.ympev.2010.02.016

Mazzon, L., Piscedda, A., Simonato, M., Martinez-Sañudo, I., Squartini, A., & Girolami, V. (2008). Presence of specific symbiotic bacteria in flies of the subfamily Tephritinae (Diptera Tephritidae) and their phylogenetic relationships: Proposal of 'Candidatus Stammerula tephritidis.' *International Journal of Systematic and Evolutionary Microbiology*, *58*(6), 1277–1287. https://doi.org/10.1099/ijs.0.65287-0

McCutcheon, J. P., & Moran, N. A. (2007). Parallel genomic evolution and metabolic interdependence in an ancient symbiosis. *Proceedings of the National Academy of Sciences*, *104*(49), 19392–19397. https://doi.org/10.1073/pnas.0708855104

Meyer, F., Fritz, A., Deng, Z.-L., Koslicki, D., Lesker, T. R., Gurevich, A., Robertson, G., Alser, M., Antipov, D., Beghini, F., Bertrand, D., Brito, J. J., Brown, C. T., Buchmann, J., Buluç, A., Chen, B., Chikhi, R., Clausen, P. T. L. C., Cristian, A., … McHardy, A. C. (2022). Critical Assessment of Metagenome Interpretation: The second round of challenges. *Nature Methods*, *19*(4), Article 4. https://doi.org/10.1038/s41592-022-01431-4

Murphy, S. M., & Loewy, K. J. (2015). Trade-offs in host choice of an herbivorous insect based on parasitism and larval performance. *Oecologia*, *179*(3), 741–751. https://doi.org/10.1007/s00442-015-3373-8

Nilsson, K. J., Ortega, J., Friberg, M., & Runemark, A. (2022). Non-parallel morphological divergence following colonization of a new host plant. *Evolutionary Ecology*, *36*(5), 859–877. https://doi.org/10.1007/s10682-022-10189-2

Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). metaSPAdes: A new versatile metagenomic assembler. *Genome Research*, *27*(5), 824–834. https://doi.org/10.1101/gr.213959.116

Oliver, K. M., Russell, J. A., Moran, N. A., & Hunter, M. S. (2003). Facultative bacterial symbionts in aphids confer resistance to parasitic wasps. *Proceedings of the National Academy of Sciences*, *100*(4), 1803–1807. https://doi.org/10.1073/pnas.0335320100

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, *25*(7), 1043–1055. https://doi.org/10.1101/gr.186072.114

Poorten, T. (2024). *Tpoorten/dotPlotly* [HTML]. https://github.com/tpoorten/dotPlotly (Original work published 2017)

Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A., & Korobeynikov, A. (2020). Using SPAdes De Novo Assembler. *Current Protocols in Bioinformatics*, *70*(1), e102. https://doi.org/10.1002/cpbi.102

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rosenberg, E., & Zilber-Rosenberg, I. (2016). Microbes Drive Evolution of Animals and Plants: The Hologenome Concept. *mBio*, *7*(2), 10.1128/mbio.01395-15. https://doi.org/10.1128/mbio.01395-15

Rudman, S. M., Greenblum, S., Hughes, R. C., Rajpurohit, S., Kiratli, O., Lowder, D. B., Lemmon, S. G., Petrov, D. A., Chaston, J. M., & Schmidt, P. (2019). Microbiome composition shapes rapid genomic adaptation of Drosophila melanogaster. *Proceedings of the National Academy of Sciences*, *116*(40), 20025–20032. https://doi.org/10.1073/pnas.1907787116

Salem, H., Kirsch, R., Pauchet, Y., Berasategui, A., Fukumori, K., Moriyama, M., Cripps, M., Windsor, D., Fukatsu, T., & Gerardo, N. M. (2020). Symbiont Digestive Range Reflects Host Plant Breadth in Herbivorous Beetles. *Current Biology*, *30*(15), 2875-2886.e4. https://doi.org/10.1016/j.cub.2020.05.043

Schwengers, O., Jelonek, L., Dieckmann, M. A., Beyvers, S., Blom, J., & Goesmann, A. (2021). Bakta: Rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. *Microbial Genomics*, *7*(11), 000685. https://doi.org/10.1099/mgen.0.000685

Seitz, A., & Komma, M. (1984). Genetic Polymorphism and its Ecological Background in Tephritid Populations (Diptera: Tephritidae). In K. Wöhrmann & V. Loeschcke (Eds.), *Population Biology and Evolution* (pp. 143–158). Springer. https://doi.org/10.1007/978-3-642-69646-6_12

Shapira, M. (2016). Gut Microbiotas and Host Evolution: Scaling Up Symbiosis. *Trends in Ecology & Evolution*, *31*(7), 539–549. https://doi.org/10.1016/j.tree.2016.03.006

Stammer, H.-J. (1929). Die bakteriensymbiose der trypetiden (Diptera). *Zeitschrift für Morphologie und Ökologie der Tiere*, *15*(3), 481–523. https://doi.org/10.1007/BF00410561

Steward, R. A., Nilsson, K. J., Ortega Giménez, J., Nolen, Z. J., Yan, C., Huang, Y., Ayala López, J., & Runemark, A. (2024). *The genomic landscape of adaptation to a new host plant*. https://doi.org/10.1101/2023.04.17.537225

Sudakaran, S., Kost, C., & Kaltenpoth, M. (2017). Symbiont Acquisition and Replacement as a Source of Ecological Innovation. *Trends in Microbiology*, *25*(5), 375–390. https://doi.org/10.1016/j.tim.2017.02.014

Takiya, D. M., Tran, P. L., Dietrich, C. H., & Moran, N. A. (2006). Co-cladogenesis spanning three phyla: Leafhoppers (Insecta: Hemiptera: Cicadellidae) and their dual bacterial symbionts. *Molecular Ecology*, *15*(13), 4175–4191. https://doi.org/10.1111/j.1365-294X.2006.03071.x

Uritskiy, G. V., DiRuggiero, J., & Taylor, J. (2018). MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome*, *6*(1), 158. https://doi.org/10.1186/s40168-018-0541-1

Vidal, M. C., & Murphy, S. M. (2018). Bottom-up vs. top-down effects on terrestrial insect herbivores: A meta-analysis. *Ecology Letters*, *21*(1), 138–150. https://doi.org/10.1111/ele.12874

Werren, J. H., Baldo, L., & Clark, M. E. (2008). Wolbachia: Master manipulators of invertebrate biology. *Nature Reviews Microbiology*, *6*(10), 741–751. https://doi.org/10.1038/nrmicro1969

Wheat, C. W., Vogel, H., Wittstock, U., Braby, M. F., Underwood, D., & Mitchell-Olds, T. (2007). The genetic basis of a plant–insect coevolutionary key innovation. *Proceedings of the National Academy of Sciences*, *104*(51), 20427–20431. https://doi.org/10.1073/pnas.0706229104

Wu, D., Daugherty, S. C., Aken, S. E. V., Pai, G. H., Watkins, K. L., Khouri, H., Tallon, L. J., Zaborsky, J. M., Dunbar, H. E., Tran, P. L., Moran, N. A., & Eisen, J. A. (2006). Metabolic Complementarity and Genomics of the Dual Bacterial Symbiosis of Sharpshooters. *PLOS Biology*, *4*(6), e188. https://doi.org/10.1371/journal.pbio.0040188

Wu, Y.-W., Simmons, B. A., & Singer, S. W. (2016). MaxBin 2.0: An automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, *32*(4), 605–607. https://doi.org/10.1093/bioinformatics/btv638

Supplementary Table 1. Summarized statistics for all bins with a CheckM2 completeness > 50%.

all-pop	bin.1	bin.2	bin.3	bin.4	bin.5	bin.6	bin.7
CheckM2 Completeness (%)	95.54	100	70.75	63	94.19	98.81	94.3
CheckM2 Contamination (%)	1.54	0.19	0.76	1.88	0.04	0	0.66

Contigs	230	175	322	105	66	34	65
N50	51	40	75	21	12	8	13
L50	27884	30981	14216	30279	91053	64899	24864
Contig_max	102846	102912	78659	101563	220337	142776	80080
GC	0.6	0.57	0.61	0.67	0.47	0.43	0.35
Size	4427102	3895730	3547305	2175021	3358317	1413378	1102802
CDS_Dens.	0.875	0.885	0.856	0.877	0.835	0.744	0.849
Tot.CDS	4359	3609	4069	2234	3156	1142	1098
Avg. Gene .Len.	296.77	319.03	249.99	285.47	296.89	307.22	285.14
23S rRNA	0	0	0	0	0	3	1
15S rRNA	0	0	0	0	0	4	1
5S rRNA	0	0	0	0	1	4	1
tRNA	53	15	31	32	38	44	32
Ori	2	1	2	1	1	0	1
ANI	98.66	98.42	N/A	97.85	N/A	N/A	98.01
AF	0.884	0.937	N/A	0.919	N/A	N/A	0.865
Genus	*Serratia*	*Erwinia*	*Pseudomonas*	*Stenotrophomonas*	*JGM124*	*Erwinia*	*Wolbachia*
Species	*S. ureilytica*	*E. aphidicola*	N/A	*S. maltophilia*	N/A	N/A	*W. pipientis*

CHST	**bin.1**	**bin.2**
CheckM2 Completeness (%)	99.03	97.63
CheckM2 Contamination (%)	0	1.96
Contigs	11	99
N50	1	19
L50	77001	17351
Contig_max	77001	70370
GC	0.43	0.35

Size	1506883	1109950
CDS_Dens.	0.733	0.859
Tot.CDS	1262	1075
Avg. Gene .Len.	292.46	296.44
23S rRNA	1	1
15S rRNA	6	1
5S rRNA	5	0
tRNA	43	35
ori	0	1
ANI	N/A	97.58
AF	N/A	0.862
Genus	*Erwinia*	*Wolbachia*
Species	N/A	*W. pipientis*

COGE	**bin.1**
CheckM2 Completeness (%)	99.07
CheckM2 Contamination (%)	0.22
Contigs	91
N50	11
L50	38359
Contig_max	126625
GC	0.44
Size	1749575
CDS_Dens.	0.755
Tot.CDS	1589
Avg. Gene .Len.	277.65
23S rRNA	0

15S rRNA	0
5S rRNA	1
tRNA	37
ori	4
ANI	N/A
AF	N/A
Closest_ref	N/A
Genus	Erwinia
Species	

Table 5 Supplementary Table 2. BLAST results for 16S rRNA sequences from Stammerula isolated from Tephritis fly hosts, *Erwinia dacicola* isolated from *Bactrocera oleae*, and *Wolbachia pipientis* isolated from *Ceratatis capitata*.

Query ID	Query length	Sequence ID (MAG contig)	Seq. length	Num. ident.	Perc. Ident.	Length	Mismatch	evalue	score	Seq. start	Seq. End	Host
EF469613.1	1315	metawrap_bin.1_CHST_000000000240	4465	1314	99.924	1315	1	0.0	1312	1422	108	*T. hyoscyami*
EF469613.1	1315	metawrap_bin.1_CHST_000000000168	7668	1314	99.924	1315	1	0.0	1312	6247	7561	*T. hyoscyami*
EF469613.1	1315	metawrap_bin.1_CHST_000000000006	72	13	99.	13	1	0.0	13	37	36	*T. h*

			653	14	924	15			12	722	408	*yoscyami*
EF469613.1	1315	metawrap_bin.1_CHST_000000000004	142915	13 14	99.924	13 15	1	0.0	13 12	14 22	108	*T. hyoscyami*
EF469613.1	1315	metawrap_bin.1_CHST_000000000003	14536	13 14	99.924	13 15	1	0.0	13 12	220 6	892	*T. hyoscyami*
EF469613.1	1315	metawrap_bin.1_CHST_000000000003	14536	13 14	99.924	13 15	1	0.0	13 12	14 015	14 529	*T. hyoscya*

													mi
EF469618.1	1315	metawrap_bin.1_CHST_000000000240	4465	1313	99.848	1315	2	0.0	1309	1422	108		*T. conura*
EF469618.1	1315	metawrap_bin.1_CHST_000000000168	7668	1313	99.848	1315	2	0.0	1309	6247	7561		*T. conura*
EF469618.1	1315	metawrap_bin.1_CHST_000000000006	7265	1313	99.848	1315	2	0.0	1309	3722	3640	8	*T. conura*
EF469618.1	1315	metawrap_bin.1_CHST_000000000004	14291	1313	99.848	1315	2	0.0	1309	1422	108		*T. conura*
EF469618.1	1315	metawrap_bin.1_CHST_000000000003	1454	1313	99.8	1315	2	0.0	1309	2206	892		*T. con*

			36		48								*ura*
EF469618.1	1315	metawrap_bin.1_CHST_000000000003	145436	1133	99.848	135	2	0.0	1309	14401	14532	9	*T. conura*
EF469615.1	1315	metawrap_bin.1_CHST_000000000240	4465	1311	99.696	135	4	0.0	1303	1422	108		*T. cometa*
EF469615.1	1315	metawrap_bin.1_CHST_000000000168	7668	1311	99.696	135	4	0.0	1303	6247	7561		*T. cometa*
EF469615.1	1315	metawrap_bin.1_CHST_000000000006	72653	1311	99.696	135	4	0.0	1303	3722	3640	8	*T. cometa*
EF469615.1	1315	metawrap_bin.1_CHST_000000000004	1429	1311	99.6	135	4	0.0	1303	1422	108		*T. com*

			15		96							*eta*
EF469615.1	1315	metawrap_bin.1_CHST_000000000003	145436	113116	99.6996	1315	4	0.00	13033	2206	892	*T. cometa*
EF469615.1	1315	metawrap_bin.1_CHST_000000000003	145436	113116	99.6996	1315	4	0.00	13033	14015	14529	*T. cometa*
FM958431.1	1464	metawrap_bin.1_CHST_000000000240	44655	13988	95.232	14682	64	0.00	12555	15699	104	*B. olea e*
FM958431.1	1464	metawrap_bin.1_CHST_000000000168	76668	13988	95.232	14682	64	0.00	12555	61006	7565	*B. olea e*
FM958431.1	1464	metawrap_bin.1_CHST_000000000006	72653	13988	95.232	14682	64	0.00	12555	37869	36404	*B. olea e*

FM958431.1	1464	metawrap_bin.1_CHST_000000000004	14291 5	1398	95.232	1468	64	0.0	12555	15569	104	*B. olea e*
FM958431.1	1464	metawrap_bin.1_CHST_000000000003	14543 6	1398	95.232	1468	64	0.0	12555	23553	888	*B. olea e*
FM958431.1	1464	metawrap_bin.1_CHST_000000000003	14543 6	1398	95.232	1468	64	0.0	12555	14386 8	14533 3	*B. olea e*
AJ586620.2	1482	metawrap_bin.1_CHST_000000000240	4465	1401	95.177 7	1472	63	0.0	12555	15577	108	*B. olea e*
AJ586620.2	1482	metawrap_bin.1_CHST_000000000168	7668	1401	95.177 7	1472	63	0.0	12555	60092	75611	*B. olea e*
AJ586620.2	1482	metawrap_bin.1_CHST_000000000006	726	14	95.11	14	63	0.0	12	3784	364	*B. olea e*

			53	01	77	72			55	77	08	*ae*
AJ586620.2	1482	metawrap_bin.1_CHST_000000000004	1429115	1401	95.177	1472	63	0.0	12555	15577	108	*B. olea e*
AJ586620.2	1482	metawrap_bin.1_CHST_000000000003	14543 6	1401	95.177	1472	63	0.0	12555	23612	892	*B. olea e*
AJ586620.2	1482	metawrap_bin.1_CHST_000000000003	14543 6	1401	95.177	1472	63	0.0	12555	14386 0	14532 9	*B. olea e*
GQ478378.1	1379	metawrap_bin.1_CHST_000000000240	4465	13 15	94.946	1385	62	0.0	1171	15 35	153	*B. olea e*
GQ478378.1	1379	metawrap_bin.1_CHST_000000000168	7668	13 15	94.946	1385	62	0.0	1171	6134	7516	*B. olea e*

GQ478378.1	1379	metawrap_bin.1_CHST_000000000006	726553	13155	94.9466	13855	62	0.0	11171	378355	364533	*B. olea e*
GQ478378.1	1379	metawrap_bin.1_CHST_000000000004	14291515	13155	94.9466	13855	62	0.0	11171	153535	15353	*B. olea e*
GQ478378.1	1379	metawrap_bin.1_CHST_000000000003	14543636	13155	94.9466	13855	62	0.0	11171	231919	9377	*B. olea e*
GQ478378.1	1379	metawrap_bin.1_CHST_000000000003	14543636	13155	94.9466	13855	62	0.0	11171	143902	145284	*B. olea e*
NZ_JADCND010000050.1:1117-2621	1505	metawrap_bin.1_CHST_000000000240	44652	1202	76.9033	15633	2766	0.0	4337	15933	554	*C. caputiata*

NZ_JADCND010000050.1:1117-2621	1505	metawrap_bin.1_CHST_000000000168	76.68	120.2	76.903	156.3	276	0.0	437	6076	7615	*C. capitata*
NZ_JADCND010000050.1:1117-2621	1505	metawrap_bin.1_CHST_000000000006	72.653	120.2	76.903	156.3	276	0.0	437	37893	36354	*C. capitata*
NZ_JADCND010000050.1:1117-2621	1505	metawrap_bin.1_CHST_000000000004	142.915	120.2	76.903	156.3	276	0.0	437	1593	54	*C. capitata*
NZ_JADCND010000050.1:1117-2621	1505	metawrap_bin.1_CHST_000000000003	145.436	120.2	76.903	156.3	276	0.0	437	2377	8388	*C. capitata*
NZ_JADCND010000050.1:1117-2621	1505	metawrap_bin.1_CHST_000000000003	145	12	76.9	15	276	0.0	437	143	145	*C. ca*

			436	02	03	63				844	383	*pt it ata*
NZ_JADCNE010000036.1:101726-103230	1505	metawrap_bin.1_CHST_000000000240	4465	1196	76.471	1564	281	0.0	416	1593	54	*C. c apt it ata*
NZ_JADCNE010000036.1:101726-103230	1505	metawrap_bin.1_CHST_000000000168	7668	1196	76.471	1564	281	0.0	416	6076	7615	*C. c apt it ata*
NZ_JADCNE010000036.1:101726-103230	1505	metawrap_bin.1_CHST_000000000006	7265	1196	76.471	1564	281	0.0	416	3789	3634	*C. c apt it ata*
NZ_JADCNE010000036.1:101726-103230	1505	metawrap_bin.1_CHST_000000000004	14291	1196	76.471	1564	281	0.0	416	1593	54	*C. c apt it*

												ata
NZ_JADCNE010000036.1:101726-103230	1505	metawrap_bin.1_CHST_000000000003	145436	1196	76.471	1564	281	0.0	416	2377	838	*C. capitata*
NZ_JADCNE010000036.1:101726-103230	1505	metawrap_bin.1_CHST_000000000003	145436	1196	76.471	1564	281	0.0	416	14438	14538	*C. capitata*

Supplementary Table 3. Summary statistics from fitted linear models against predicting relative abundance for Stammerula MAG (CHSt1) from the CHST metagenome and W. pipientis MAG (AWp7) from the all-pop metagenome; Standard Error (SE), Residual Standard Error (RSE), Degrees of Freedom (DF), Adjusted R-squared (Adj.R^2)

Model		Coefficient	Estimate	SE	t-value	p-value
CHSt1.lm.1		Intercept	40.868	3.256	12.552	1.56E-05
		HPCO	-13.432	4.604	-2.917	0.0267
	RSE	7.025 on 5 DF				
	Adj.R^2	0.4386				
	F-statistic	8.511 on 1 and 6 DF, p = 0.02672				
	Formula	CHSt1 ~ HP, data = specialists				
CHSt1.lm.2		Intercept	34.151	3.205	10.656	5.27E-06
		typeGeneralists	-13.342	7.166	-1.86	0.0997
	RSE	9.065 on 8 DF				
	Adj.R^2	0.2151				
	F-statistic	3.466 on 1 and 8 DF, p = 0.09967				
	Formula	CHSt1 ~ type, data = generalist and specialist				
CHSt1.lm.2		Intercept	34.151	3.205	10.656	5.27E-06
		typeGeneralists	-13.342	7.166	-1.86	0.0997
	RSE	9.065 on 8 DF				
	Adj.R^2	0.2151				
	F-statistic	3.466 on 1 and 8 DF, p = 0.09967				
	Formula	CHSt1 ~ type, data = generalist and specialist				
CHSt1.lm.3		Intercept	40.868	3.302	12.377	2.45E-04
		typeGeneralists	-20.058	5.719	-3.507	0.024734
	RSE	6.604 on 4 DF				
	Adj.R^2	0.6933				
	F-statistic	12.3 on 1 and 4 DF, p = 0.02473				
	Formula	CHSt1 ~ type, data = generalists and specialists without CO populations				
AWp7.lm.1		Intercept	1.639	1.226	1.337	0.218
		typeGeneralist	38.192	2.741	13.933	6.82E-07
	RSE	3.467 on 8 DF				
	Adj.R^2	0.9555				
	F-statistic	194.1 on 1 and 8 DF, p = 6.819E-07				
	Formula	AWp7 ~ type, data = generalists and specialists				
AWp7.lm.2		Intercept	1.8301	0.6965	2.628	0.0392
		HPCO	-0.3820	0.9850	-0.388	0.7115
	RSE	1.393 on 6 DF				
	Adj.R^2	-0.1381				
	F-statistic	0.1504 on 1 and 6 DF, p = 0.7115				
	Formula	AWp7 ~ HP, data = specialists				