# Supplementary Information for: "Multiple *Wolbachia* subpopulations co-occur in single *Culex pipiens* mosquito organs"

# Supplementary Notes

**Supplementary Note 1. Sample collection, preparation, and sequencing.** Briefly, we collected Culex mosquito specimens in Languedoc (France) in May 2017 using a carbon dioxide mosquito trap (Camping l'Europe de Vic La Gardiole, Entente Interdépartementale de Démoustication EID, Méditerranée). We transported them alive to the laboratory directly afterward and anesthetized females by incubating them for 4 minutes (min) at -20°C. We washed them with ethanol 96% during 1mn in order to remove eventual surface contaminants and PBS 1X to avoid DNA precipitation due to ethanol. We transferred them onto a sterile microscope slide with one drop of sterile PBS 1X and dissected ovaries and midgut from each mosquito using sterilized tweezers. Finally, we stored four ovaries (O03, O07, O11, O12) and their corresponding midgut samples (M03, M07, M11, M12, together with two additional "orphan" samples M01, M09) from individual mosquitoes at -80°C to preserve them until further processing. We extracted the DNA from each sample using the MoBio PowerFecal DNA Isolation Kit (QIAGEN Inc., Germantown, MD, USA). We sonicated 3.8-5.7 ng of genomic DNA using an E200 Covaris instrument (Covaris, Woburn, MA, USA). We used the NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs, Ipswich, MA, USA) to end-repair and 3'-adenylate the resulting fragments and added NEXTflex PCR-free barcode adapters (Bioo Scientific, Austin, TX, USA). We purified the ligation products using Ampure XP (Beckman Coulter, Brea, CA, USA) and amplified the DNA fragments by PCR (>200 bp; 2 PCR reactions, 14 cycles) using Illumina adapter-specific primers and NEBNext Ultra II Q5 Master Mix (NEB). We used an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA) to analyze library profiles and performed a qPCR analysis using the KAPA Library Quantification Kit for Illumina Libraries (KapaBiosystems, Wilmington, MA, USA). We sequenced the library using a HiSeq4000 Illumina sequencer (Illumina, San Diego, CA, USA) at the Genoscope in Evry, France, generating 151 bp paired-end reads. Finally, we used a cluster intensity and chastity filter as described in [1] in order to filter the raw sequencing.

**Supplementary Note 2. Estimation of host contamination.** We estimated the host contamination in the ovary and midgut quality filtered short reads using the software phyloFlash [2] v3.4. We used the "phyloFlash.pl" script with the "--readlength 150" and "--almosteverything" flags that i) extracted the small-subunit rRNA (SSU rRNA) sequences from our data by mapping our reads to the SILVA SSU Ref database [3], ii) assembled the extracted reads into SSU rRNA using SPAdes [4] assembler, iii) estimated the proportion of assembled sequences by re-mapping them to the full-length SSU sequences, iv) provided a taxonomic summary of the reads from the initial mapping, the full-length SSU rRNA sequences and the unassembled sequences. We plotted the taxonomic assignment of the mapped reads on the SILVA SSU databases using a custom R script (R version 3.6.3 [5]) to visualize the estimation of the host contamination in our samples.

**Supplementary Note 3 –*Wolbachia* SNVs at the intra-sample level.** We mapped the metagenome from each sample to the corresponding *Wolbachia* MAG to access intra-sample variability. We characterized a mean of 3,743 raw SNVs within our *Wolbachia* MAGs (intra-

sample variability) with a mean coverage of 367X (Supplementary Table 6). After our filtration steps (entropy > 0.2 and departure from consensus > 0.2), we obtained a final number of 293 SNVs on average with a mean coverage value of 533X in our samples (Supplementary Table 6). In order to quantify the extent of variability at variable nucleotide positions in the metagenome compared to the reference sequence, we used the metric 'departure from consensus', which reports the ratio of nucleotides at a given position that do not match to the consensus nucleotide. The departure from consensus values varied from 0.2 to 0.72 indicating between 20 to 72% of base variation at one SNV position in comparison to the consensus sequence within each *Wolbachia* MAG. (Figure 1, Supplementary Table 6).

We then focused our analysis on SNVs occurring within gene calls in order to study the potential influence of genomic variation on protein-coding genes in our MAGs. A SNV located in the 1st or 2nd position of the codon will be more susceptible to result in a non-synonymous mutation (resulting in changes in the amino-acid sequence) whereas a SNV located in the 3rd position will be more likely to result in a synonymous mutation (no changes in the amino-acid sequence). We reported 160, 194, 92, 347 and 59 SNVs occurring in gene call within O11 (Figure 1 - panel A, Supplementary Table 6, Supplementary Table 7, M11, O03, O07 and O12 (Figure 1 - panel B, Supplementary Table 6, Supplementary Table 7), respectively. Among those gene-call SNVs, we counted a mean number of 41, 35 and 94 SNVs at the 1$^{st}$, the 2$^{nd}$ and the 3$^{rd}$ base position, respectively (Figure 1, Supplementary Table 6, Supplementary Table 7.

From these SNVs, we reported an average number of 140 SCVs (after similar filtration metrics including entropy and departure from consensus) in coding genes with a mean coverage of 483X in our MAGs (Supplementary Table 9, see Supplementary Table 7 for the raw and filtered intra-sample SCVs in details). From them, we counted a mean number of 67 SAAVs (after filtration) with a mean coverage of 463X (Supplementary Table 9, see Supplementary Tables 8 for the raw and filtered intra-sample SAAVs in details).

**Supplementary Note 4 – *Wolbachia* SNVs at the inter-sample level.** We mapped the metagenomes from all samples to each *Wolbachia* MAGs to access inter-sample variability. We reported 22,602 raw SNVs occurring in our five *Wolbachia* MAGs on average with a mean coverage of 367X (Supplementary Table 10). After filtration, we counted 1,977 inter-sample SNVs on average with a mean coverage value of 492X (Supplementary Table 10). We counted a mean of 1,621 SNVs in each MAG shared by at least one another MAG at the same position with a mean coverage of 681X (Supplementary Table 10). In details, we reported a mean of 226 previously identified intra-sample SNVs from each MAG (157 out of 281 for M11, 155/186 for O11, 133/179 for O03, 533/701 for O07 and 92/119 for O12) that are shared by at least two metagenomes (Supplementary Table 6, Supplementary Table 10). Among them, 5 SNVs on average fell in *w*SCG (Supplementary Table 6, Supplementary Table 10).

**Supplementary Note 5- *Wolbachia* SNVs at the intra-sample level in wSCGs**

On average, we reported 8 filtered SNVs among a mean of 3 *w*SCGs in our different *Wolbachia* MAGs (Supplementary Table 9).

In *Wolbachia* MAG M11, we reported a total of 6 filtered SNVs including 3 SNVs falling in two *w*SCGs (gene caller ids 301 and 771) that were not annotated to any function using COG20 and KEGG databases. In *w*SCG 301, 1 SNV falls at the 2nd position in codon and 2

SNVs at the 3rd position in codon. In wSCG 771, the 3 SNVs fall at the 1st position in codon. On average, all these SNVs have a mean coverage of 164X and a departure from consensus of 0.23 (Supplementary Table 9 sheet 2).

In *Wolbachia* MAG O11, we reported a total of 5 filtered SNVs including 4 SNVs falling in a *w*SCG (gene caller ids 755) that was annotated to an « Ankyrin repeat » function by the COG20 database and 1 SNV falling in a *w*SCG (gene caller id 1152) without any function. In *w*SCG 755, a number of 1,1 and 2 SNVs falls the 1st, 2nd and 3rd position in codon, respectively. In *w*SCG 1152, the unique SNV fall at the 3rd position in codon. On average, all these SNVs have a mean coverage of 510X and a departure from consensus of 0.26 (Supplementary Table 9 sheet 2).

In *Wolbachia* MAG O03, we reported a total of 10 filtered SNVs in 3 *w*SCGs (gene caller ids 1, 589 and 931). Among them, 4 SNVs fall in a *w*SCG annotated to an « Ankyrin repeat » function (COG20), 5 in an « ATP-dependant 26S proteasome regulatory subunit » (COG20) and 1 in a *w*SCG without function. In *w*SCG 1, a number of 2, 1 and 1 SNVs fall at the 1st, 2nd and 3rd codon position, respectively. In *w*SCG 589, 1, 3, and 1 SNV falls at the 1st, 2nd and 3rd position in codon,respectively. In *w*SCG 931, the unique SNV falls at the 1st position in codon. On average, all these SNVs have a mean coverage of 253X and a departure from consensus of 0.24. Among them, 4, 4, and 2 SNVs occur at the 1st, 2nd and 3rd codon position, respectively (Supplementary Table 9).

In *Wolbachia* MAG O07, we reported a total of 9 filtered SNVs in four *w*SCGs (gene caller ids 230, 350, 615 and 996). Among them, 6 SNV fall in three *w*SCGs without function, 4 SNVs in a *w*SCG annotated to an « ATP-dependant 26S proteasome regulatory subunit » (COG20) and 1 SNV in a *w*SCG annotated to a « Chromosome segregation ATPase Smc » (COG20). In *w*SCG 230, the unique SNV falls at the 3rd position in codon. In *w*SCG 350, 3 SNVs fall at the 1st position in codon and 1 SNV at the 2nd position. In *w*SCG 615, the unique SNV falls at the 1st position in codon. And in *w*SCG 996, 2 SNVs occur at the 1st position in codon and 1 SNV at the 2nd position in codon. On average, all these SNVs have a mean coverage of 146X and a departure from consensus of 0.25 (Supplementary Table 9).

In *Wolbachia* MAG O12, we reported a total of 9 filtered SNVs in 4 *w*SCGs (gene caller ids 423, 611 and 640). Among them, 2 SNVs fall in 1 *w*SCGs with no annotated function, 3 SNVs in a *w*SCG annotated to an « Ankyrin repeat » function (COG20) and 4 SNVs in a *w*SCG annotated to an « ATP-dependant 26S proteasome regulatory subunit » (COG20). In *w*SCG 423, 1 SNV falls at the 1st position in codon and 1 SNV at the 3rd position in codon. In *w*SCG 611, 1 SNV fall at the 1st position in codon and 2 SNVs at the 2nd position. In *w*SCG 640, 1, 2, and 1 SNV falls at the 1st , 2nd position and 3rd position. On average, all these SNVs have a mean coverage of 204X and a departure from consensus of 0.26 (Supplementary Table 9).

# Supplementary Tables

**Supplementary Table 1** – ***Wolbachia*** **MAGs estimates.** Number of clusters used for binning (using the "--clusters" flag of the "anvi-cluster-contigs" anvi'o program as detailed in Material and Methods section), raw percent completion (before refinement), raw percent redundancy (before refinement), percent completion (after refinement), percent redundancy

(after refinement), number of contigs (after refinement), length (after refinement), GC content (after refinement).

**Supplementary Table 2** – **Quality filtering and assembly statistics for each sample (M01, M03, M07, M09, M11, M12, O11, O03, O07, O12).** Total number of raw pairs of reads, number of filtered pairs of reads, percent of filtered (passed) reads, number of contigs > 1kb, L50, N50 and percent of high-quality recruited reads by sample.

**Supplementary Table 3** – **Taxonomic profiling.** Taxonomic summaries based on the mapping hits of the quality filtered reads to the SILVA SSU rRNA database performed during the phyloFlash analyses for each sample (M01, M03, M07, M09, M11, M12, O03, O07, O11, O12). The table reports all the identified taxa (from the Domain to the Species rank) and their number of reads.

**Supplementary Table 4** – **Gene cluster statistics.** Gene clusters identified during the pangenomics analyses performed on all metagenomes (M11, O11, O03, O07 and O12) and the three selected *Wolbachia* reference genomes *w*PipPel, *w*PipMol, *w*PipJHB. Table report for each identified gene cluster: ids (unique_id, gene_cluster_id), bin and genome names (bin_name, genome_name), gene caller id (gene_callers_id), number of genomes where gene cluster occurs (num_genomes_gene_cluster_has_hits), number of genes in gene cluster (num_genes_in_gene_cluster), maximum number of paralogs (max_num_paralogs), identification as Single-Copy Core genes (SCG), functional, geometric and combined homogeneity indices (functional_homogeneity_index, geometric_homogeneity_index, combined_homogeneity_index), KEGG assignation information (KEGG_Module_ACC, KEGG_Module, , KEGG_Class_ACC, KEGG_Class), COG20 assignation information (COG20_PATHWAY_ACC, COG20_PATHWAY, COG20_FUNCTION_ACC, COG20_FUNCTION, COG20_CATEGORY_ACC, COG20_CATEGORY), KOfam assignation information (KOfam_ACC, KOfam, KEGG_Module_ACC) and corresponding amino acid sequence (aa_sequence).

**Supplementary Table 5** –**Single-copy Core Gene from *Wolbachia*.** *w*SCG identified among the unique GC computed during the pangenomics analyses performed on all metagenomes and selected *Wolbachia* reference genomes. Table reports for each unique GC: gene cluster id (gene_cluster_id) and *w*SCG info (1=*w*SCG).

**Supplementary Table 6** – **Raw and filtered SNV tables at the intra-sample level.** Table reports for each SNV position: unique identifier of the SNV (entry_id), unique position identifier of the SNV (unique_pos_identifier), nucleotide position of the SNV in the split and the contig (pos, pos_in_contig), name of the split where SNV is occurring (split_name), sample where the SNV has been identified (sample_id), unique gene caller identifier where the SNV is occurring (corresponding_gene_call with "-1" if the position is not in a gene call), coding or non-coding status of the gene (in_noncoding_gene_call, in_coding_gene_call), nucleotide position in codon (base_pos_in_codon), order of the codon in the gene call (codon_order_in_gene with "-1" if the position is not in a gene call, codon_number), gene length (gene_length), number of recruited reads mapping to this position (coverage), outlier status of the SNV based on coverage estimated in split and contig (cov_outlier_in_split, cov_outlier_in_contig), number of mapped reads covering the position corresponding to the different bases (A, C, G, N, T), reference nucleotide in the reference genome (reference), most mapped nucleotide at the position (consensus), the two most represented nucleotides at the position (competing_nts), ratio of nucleotides in a given position that diverge from the

reference nucleotide (departure_from_reference), ratio of nucleotides in a given position that diverge from the consensus nucleotide (departure_from_consensus), ratio of the second most frequent nucleotide to the consensus nucleotide (n2n1ratio) and the entropy value (entropy). All these SNV information are also explained at the following link https://merenlab.org/2015/07/20/analyzing-variability/. For filtered SNVs, the table also reports the presence of the SNV in a *w*SCG (*w*SCG), prophage WO regions assignment (WO_assignmnent, WO_pct_alignment), *cid* genes assignment on the gene where the SNV occurred (*cid*_genes_from_Bonneau_et_al._2018 [6], *cid*_Bonneau_pct_alignment), the MLST and *wsp* genes assignment (MLST_and_*wsp*_genes_from_PubMLST [7], MLST_*wsp*_pct_alignment), the COG20 assignation information (COG20_PATHWAY_ACC, COG20_PATHWAY, COG20_FUNCTION_ACC, COG20_FUNCTION, COG20_CATEGORY_ACC, COG20_CATEGORY), the KOfam assignation information (KOfam_ACC, KOfam), the KEGG assignation information (KEGG_Module_ACC, KEGG_Module, KEGG_Class_ACC, KEGG_Class) and the corresponding amino acid sequence (aa_sequence).

**Supplementary Table 7** – **Variability statistics**. Sheet 1. For each sample, the table reports the count of the intra-sample SNVs, SCVs and SAAVs (after filtration; entropy and departure from consensus >= 0.2) that occurred in genes, coding genes and *w*SCGs. Coverage and departure from consensus are reported for SNVs, SCVs and SAAVs. Codon positions are reported only for SNVs. The pN/pS ratio (non-synonymous to synonymous mutations) is reported for SCVs. Sheet 2. For each SNV, table reports the *Wolbachia* MAG, their number in each split and gene they fall in, their positions in the split and in the codon as well as the function associated to the gene.

**Supplementary Table 8** – **Raw and filtered SCV tables at the intra-sample level for each metagenome**. Table reports for each SCV: the unique identifier of the SCV (entry_id), unique position identifier of the SCV (unique_pos_identifier), name of the split where SCV is occurring (split_name), sample where the SCV has been identified (sample_id), unique gene caller identifier where the SCV is occurring (corresponding_gene_call with "-1" if the position is not in a gene call), order of the codon in the gene call (codon_order_in_gene with "-1" if the position is not in a gene call, codon_number), gene length (gene_length), number of recruited reads mapping to this codon (coverage), number of mapped reads covering the SCV corresponding to the different codons (64 codon combinations), reference codon in the reference genome (reference), most mapped codon (consensus), the two most represented codons (competing_codons), ratio of codon in a given SCV that diverge from the reference codon (departure_from_reference), ratio of codon in a given SCV that diverge from the consensus codon (departure_from_consensus), ratio of the second most frequent codon to the consensus codon (n2n1ratio), the entropy value (entropy) and the synonymity values (pN, pS, nN, nS for consensus, reference and popular consensus). All these SCV information are also explained at the following link https://merenlab.org/2015/07/20/analyzing-variability/. For filtered SCVs, additional variables are reported : prophage WO regions assignment (WO_assignmnent, WO_pct_alignment), *cid* genes assignment (cid_genes_from_Bonneau_et_al._2018 [6], *cid*_Bonneau_pct_alignment), the MLST and *wsp* genes assignment (MLST_and_wsp_genes_from_PubMLST [7] MLST_*wsp*_pct_alignment) and wSCG info (*w*SCG).

**Supplementary Table 9 - Raw and filtered SAAV tables at the intra-sample level.** Table reports for each SAAV: the unique identifier of the SAAV (entry_id), unique position identifier of the SAAV (unique_pos_identifier), name of the split where SAAV is occurring

(split_name), sample where the SAAV has been identified (sample_id), unique gene caller identifier where the SAAV is occurring (corresponding_gene_call with "-1" if the position is not in a gene call), order of the codon in the gene call (codon_order_in_gene with "-1" if the position is not in a gene call, codon_number), gene length (gene_length), number of recruited reads mapping to this amino acid (coverage), number of mapped reads covering the SAAV corresponding to the different amino acids (Ala, Arg, Asn, Asp, Cys, Gln, Glu, Gly, His, Ile, Leu, Lys, Met, Phe, Pro, STP, Ser, Thr, Trp, Tyr, Val), reference amino acid in the reference genome (reference), most mapped amino acid at the SAAV (consensus), the two most represented amino acids (competing_aas), ratio of amino acids in a given SAAV that diverge from the reference amino acids (departure_from_reference), ratio of amino acids in a given SAAV that diverge from the consensus amino acids (departure_from_consensus), ratio of the second most frequent amino acid to the consensus amino acid (n2n1ratio), the entropy value (entropy), the interchangeability of two amino acids represented by BLOSSUM (BLOcks SUbstitution Matrix) matrices (BLOSUM90, BLOSUM90_weighted, BLOSUM62, BLOSUM62_weighted). All these SAAV information are also explained at the following link https://merenlab.org/2015/07/20/analyzing-variability/. For filtered SAAVs, the table also reports: *cid* genes assignment (*cid*_genes_from_Bonneau_et_al._2018 [6], *cid*_Bonneau_pct_alignment), MLST and *wsp* genes assignment (MLST_and_*wsp*_genes_from_PubMLST [7], MLST_*wsp*_pct_alignment) and *w*SCG information (*w*SCG).

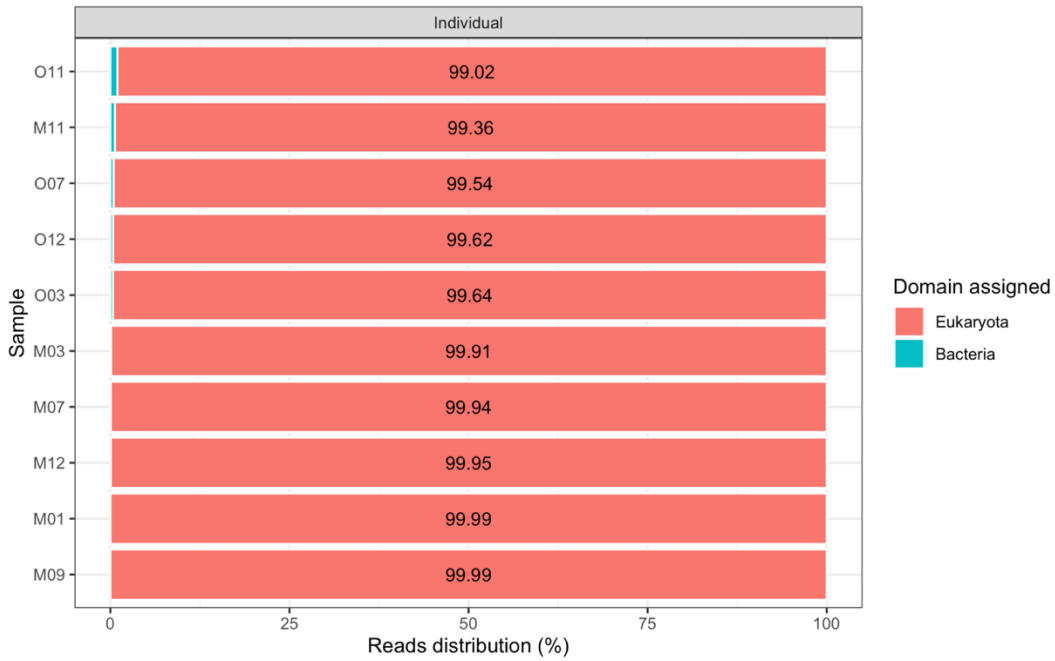
**Supplementary Table 10 – Raw and filtered SNV tables at the inter-sample level**. Table reports the same information as in Supplementary Table 6.

**Supplementary Table 11 – Raw and filtered indel tables for each *Wolbachia* MAGs (M11, O11, O03, O07, O12)**. Table reports the sample where the indel has been identified (sample_id), the split name where the indel has been identified (split_name), its position in split and contig (pos, pos_in_contig), its corresponding gene call (corresponding_gene_call), its status of coding or non-conding gene call (in_noncoding_gene_call, in_coding_gene_call), the indel position in codon (base_pos_in_codon), the order of the codon in the gene call (codon_order_in_gene), the outlier status of the indel based on coverage estimated in split and contig (cov_outlier_in_split, cov_outlier_in_contig), the reference nucleotide in the reference genome (reference), the mutation type of the indel (type), its sequence (sequence), its sequence length (length), its count (count) and its coverage (coverage).

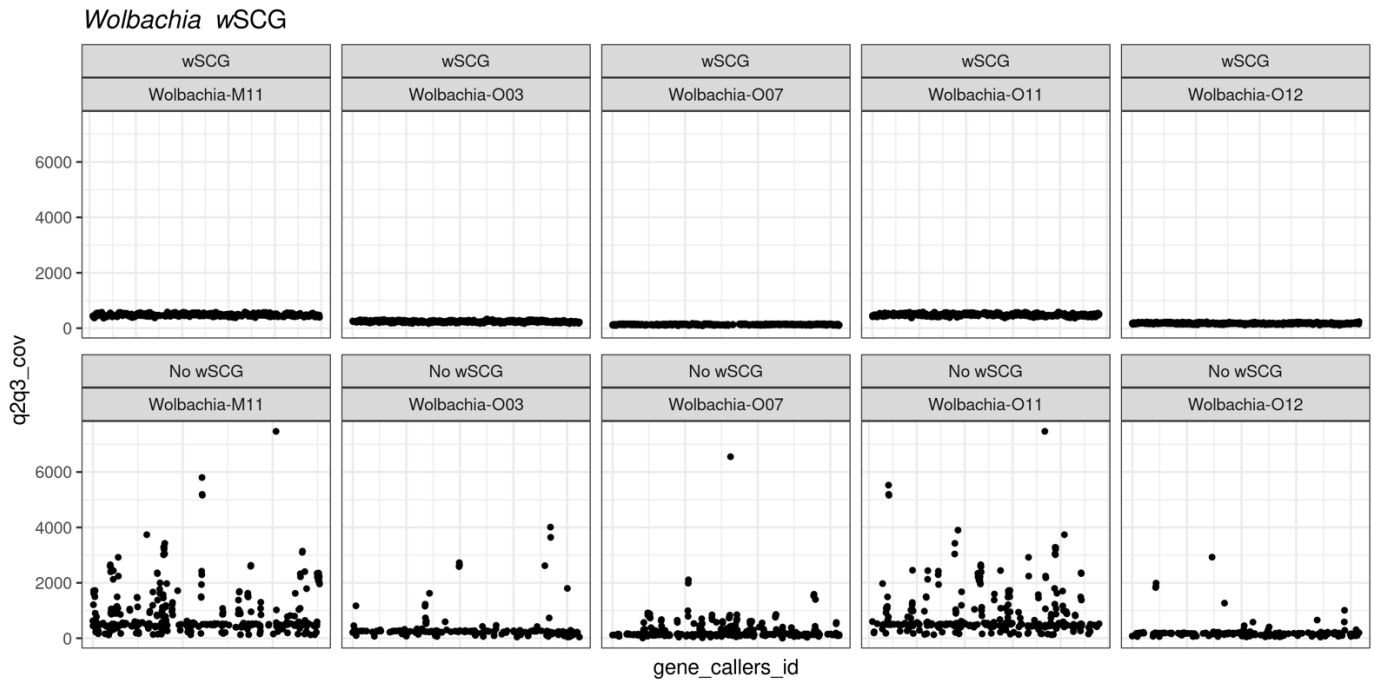
**Supplementary Table 12 –Raw and filtered SNV tables at larger geographic scale (including MAG M11, O11, O03, O07, O12 from this study and Harash, Istanbul and Tunis metagenomes from [6]).** Table of SNVs occurring in each *Wolbachia* MAG after mapping of *Culex pipiens* egg metagenomes (SRR5810516, Pipiens_MGx_Istanbul; SRR5810518, Pipiens_MGx_Tunis; SRR5810517 Pipiens_MGx_Harash) from [6]. Table reports raw and filtered (without outliers, entropy and departure from consensus > 0.2) SNVs for each *Wolbachia* MAG (M11, O11, O03, O07, O12).

**Supplementary Table 13 –Specimen 11 inter-organ *Wolbachia* pangenomic analysis.** Gene cluster table from the pangenomics analyses performed on the M11, O11 *Wolbachia* MAGs and the *w*Pip reference genome. The table reports the same information as in Supplementary Table 4.

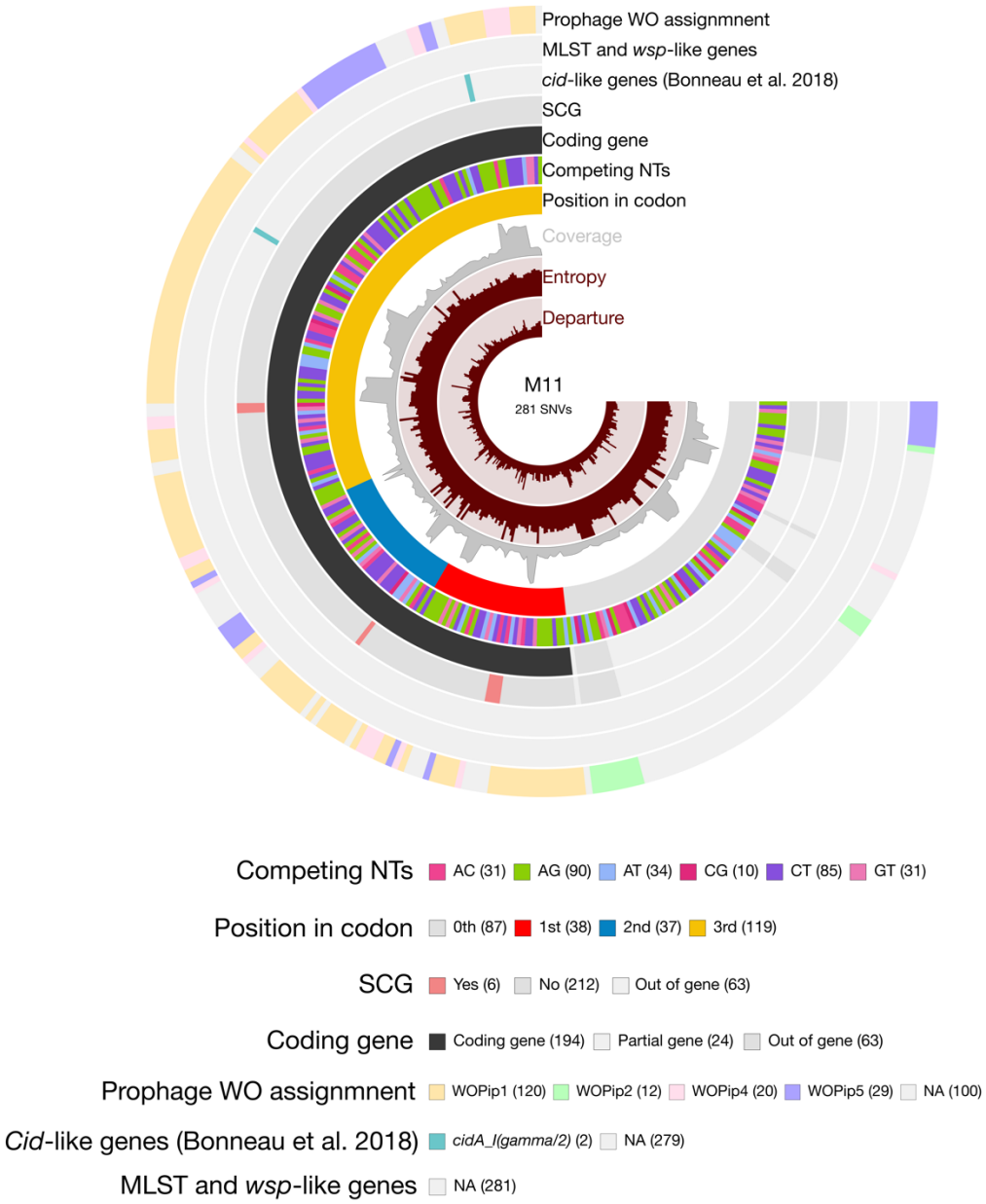
# Supplementary Figures



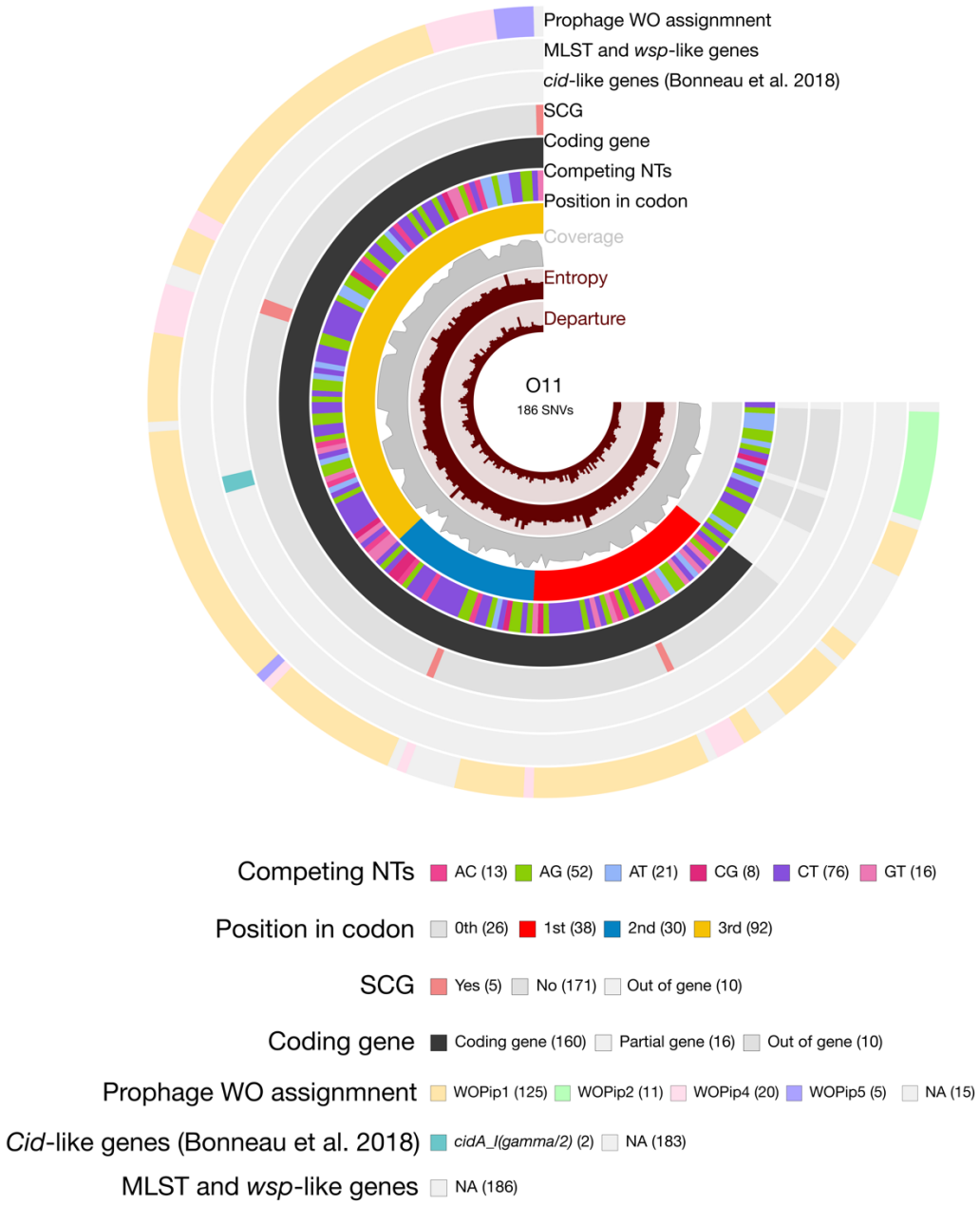

**Supplementary Figure 1 – Estimation of the host contamination in our metagenomes** based on phyloFlash (precisely from the taxonomic assignment of the mapped reads onto the SILVA SSU databases) showing a particularly high eukaryotic contamination rate in midguts samples that may explain the impossibility to reconstruct bacterial genomes in these samples. Red bar represents the percent of eukaryotic reads detected in samples and the blue bar represent the bacterial ones.



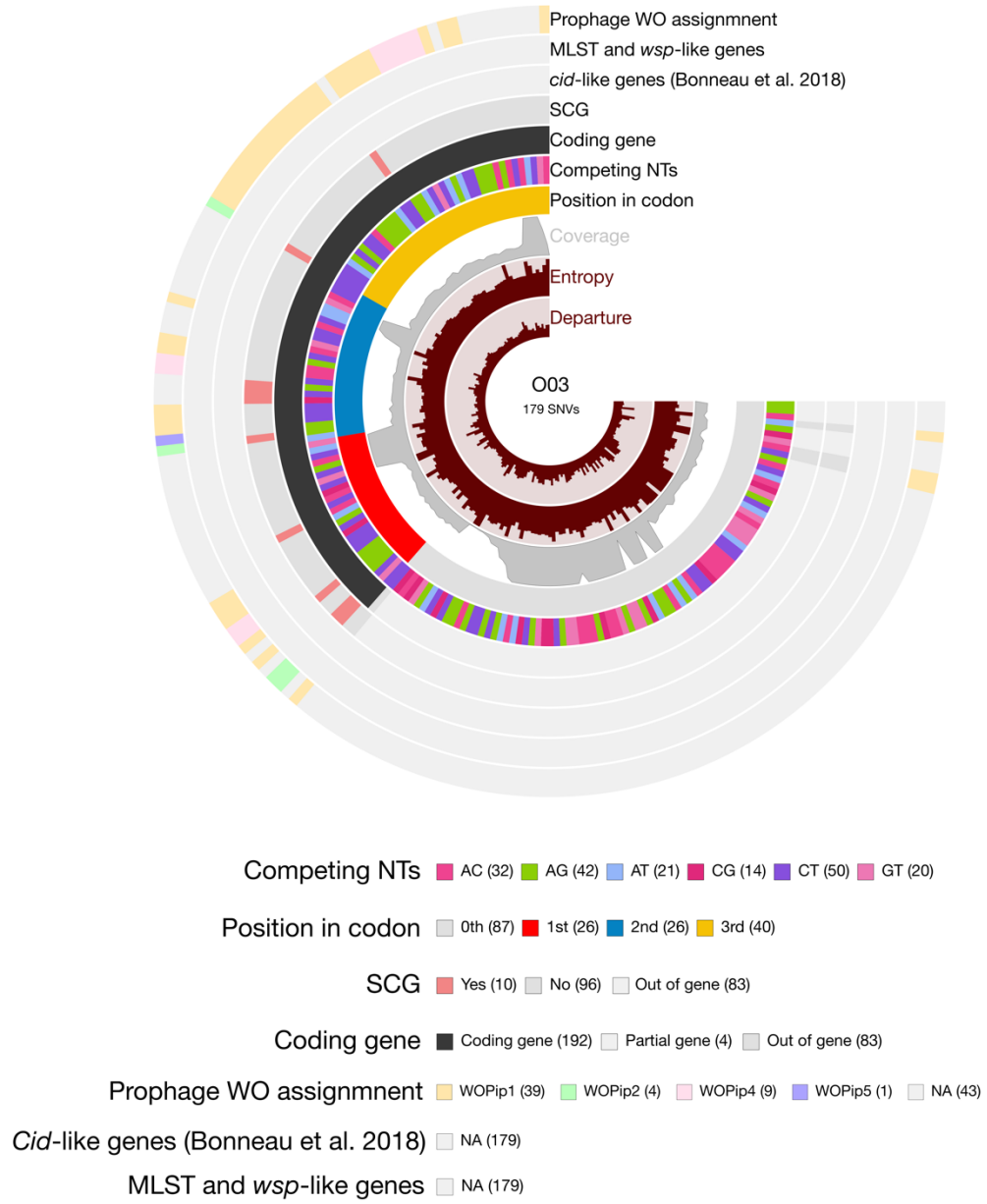

**Supplementary Figure 2 – Coverage of gene clusters identified or not as *w*SCG for each *Wolbachia* MAG**. Metagenomic reads from each sample have been mapped to corresponding *Wolbachia* MAG and coverage (q2_q3_cov) for each gene cluster is represented on y-axis.



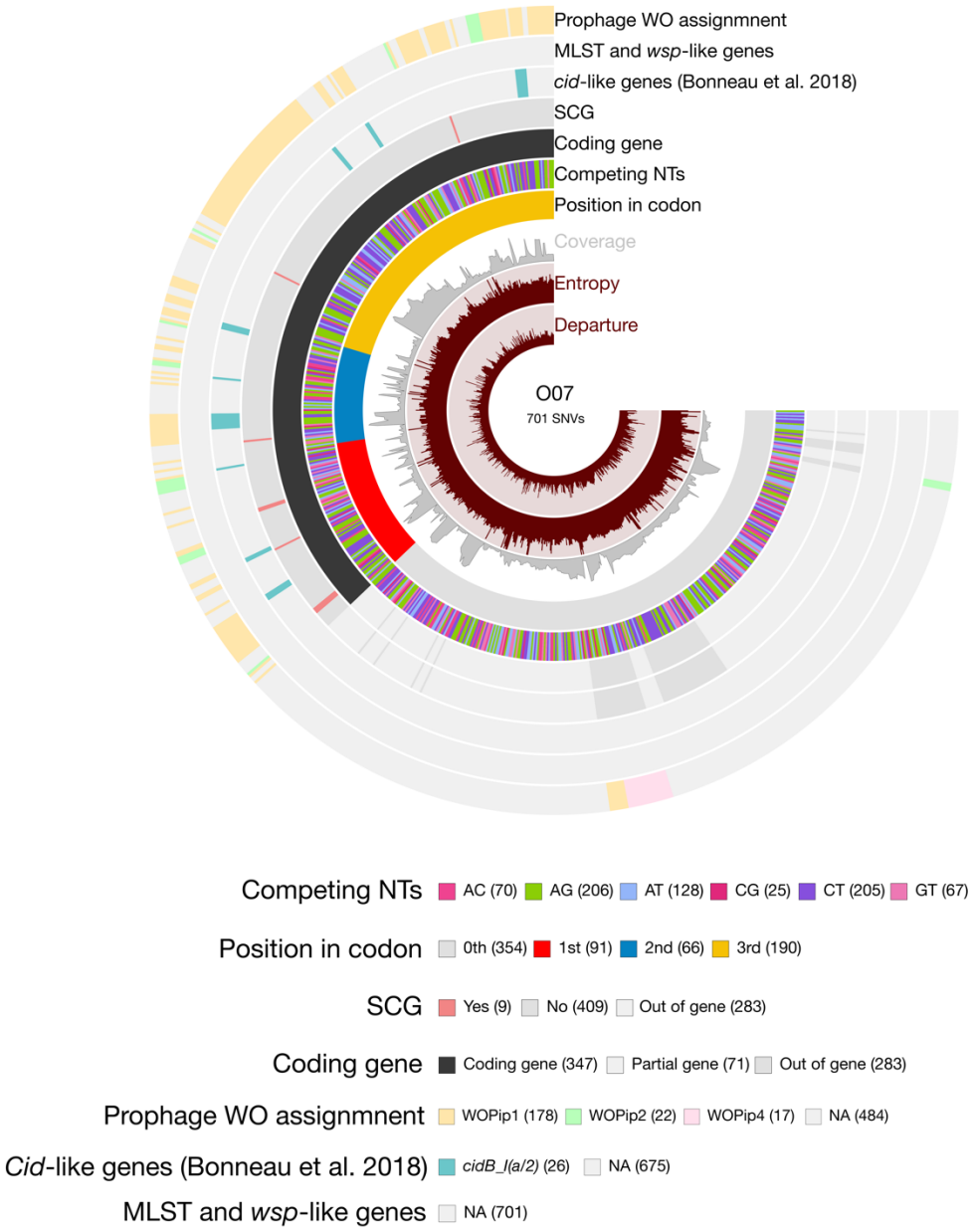

**Supplementary Figure 3** – **Representation of the filtered SNVs (entropy and departure from consensus >= 0.2 and without coverage outliers) in the *Wolbachia* MAG M11**. From inner to outer and for each SNV (represented by a leaf), the layers refer to i) the departure from the consensus (max value set to 1), ii) the Shannon entropy (max value set to 1), iii) the coverage (max value set to 2500X), iv) the position in codon (from 1st to 3rd if the SNV was detected in a coding gene and 0th if it not), v) the competing nucleotides (the two most nucleotides occurring in the SNV), vi) the presence of SNV in a coding gene call, non-coding gene or "out of gene", vii) the presence in *w*SCGs shared by all MAGs and reference genomes, viii) the presence of the SNV in a gene annotated to a *cid* gene (from Bonneau et al. 2018 [6]), ix) the presence of the SNV in a gene annotated to a MLST or a *wsp* gene (from PubMLST [7]), and x) the annotation of prophage WO regions.



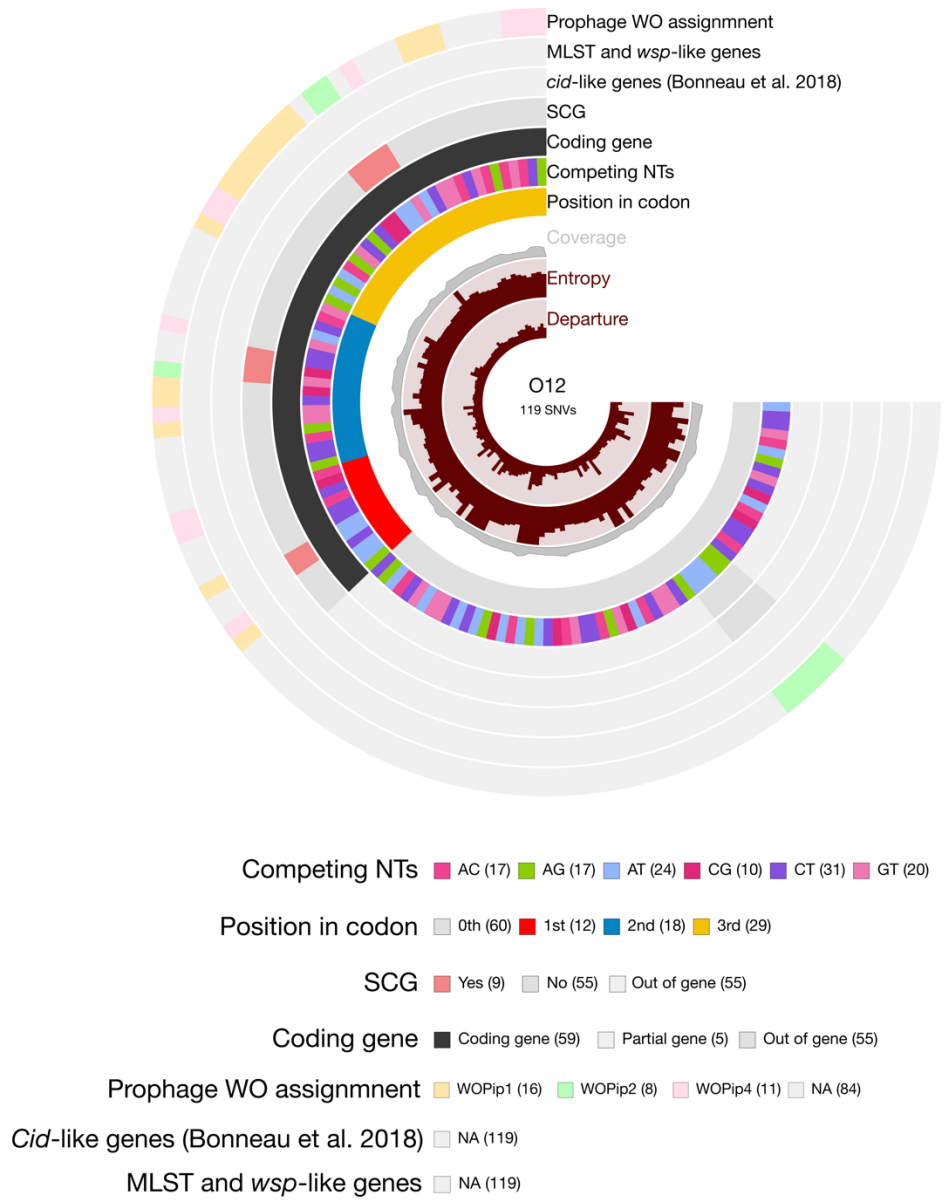

**Supplementary Figure 4 – Representation of the filtered SNVs (entropy and departure from consensus >= 0.2 and without coverage outliers) in the *Wolbachia* MAG O11**. From inner to outer and for each SNV (represented by a leaf), the layers correspond to i) the departure from the consensus (max value set to 1), ii) the Shannon entropy (max value set to 1), iii) the coverage (max value set to 2500X), iv) the position in codon (from 1st to 3rd if the SNV was detected in a coding gene and 0th if it not), v) the competing nucleotides (the two most nucleotides occurring in the SNV), vi) the presence of SNV in a coding gene call, non-coding gene or "out of gene", vii) the presence in *w*SCGs shared by all MAGs and reference genomes, viii) the presence of the SNV in a gene annotated to a *cid* gene (from Bonneau et al. 2018 [6]), ix) the presence of the SNV in a gene annotated to a MLST or a *wsp* gene (from PubMLST [7]), and x) the annotation of prophage WO regions.



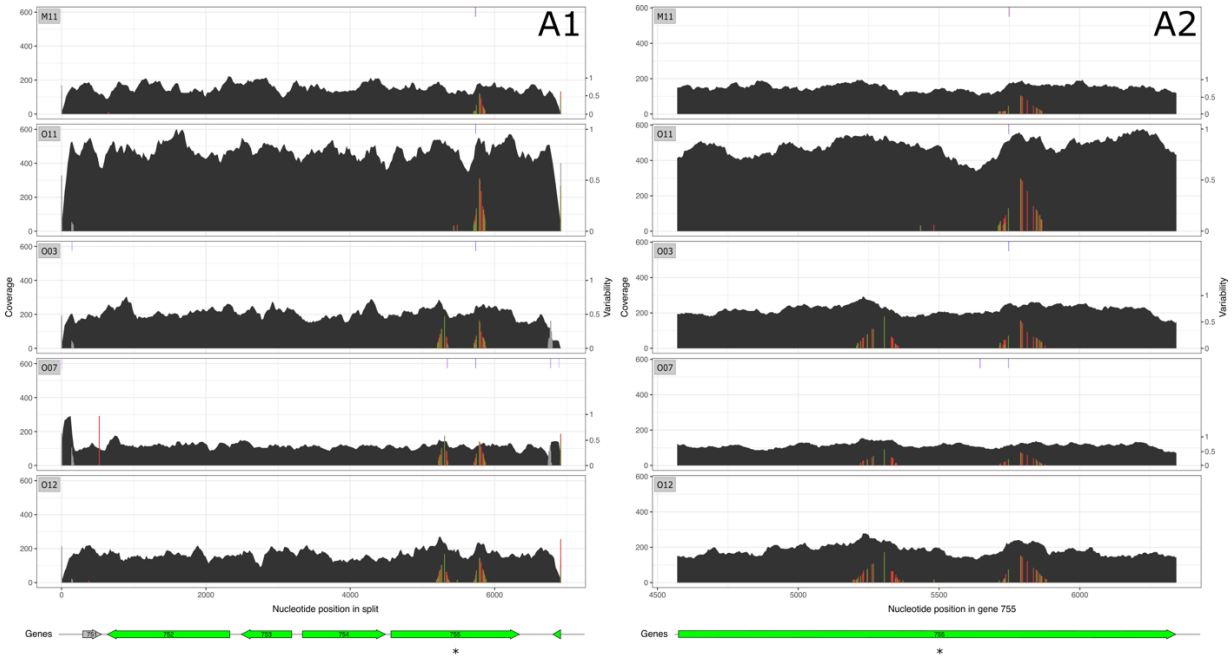

**Supplementary Figure 5 – Representation of the filtered SNVs (entropy and departure from consensus >= 0.2 and without coverage outliers) in the *Wolbachia* MAG O03**. From inner to outer and for each SNV (represented by a leaf), the layers correspond to i) the departure from the consensus (max value set to 1), ii) the Shannon entropy (max value set to 1), iii) the coverage (max value set to 2500X), iv) the position in codon (from 1st to 3rd if the SNV was detected in a coding gene and 0th if it not), v) the competing nucleotides (the two most nucleotides occurring in the SNV), vi) the presence of SNV in a coding gene call, non-coding gene or "out of gene", vii) the presence in *w*SCGs shared by all the MAGs and reference genomes, viii) the presence of the SNV in a gene annotated to a *cid* gene (from Bonneau et al. 2018 [6]), ix) the presence of the SNV in a gene annotated to a MLST or a *wsp* gene (from PubMLST [7]), and x) the annotation of prophage WO regions.



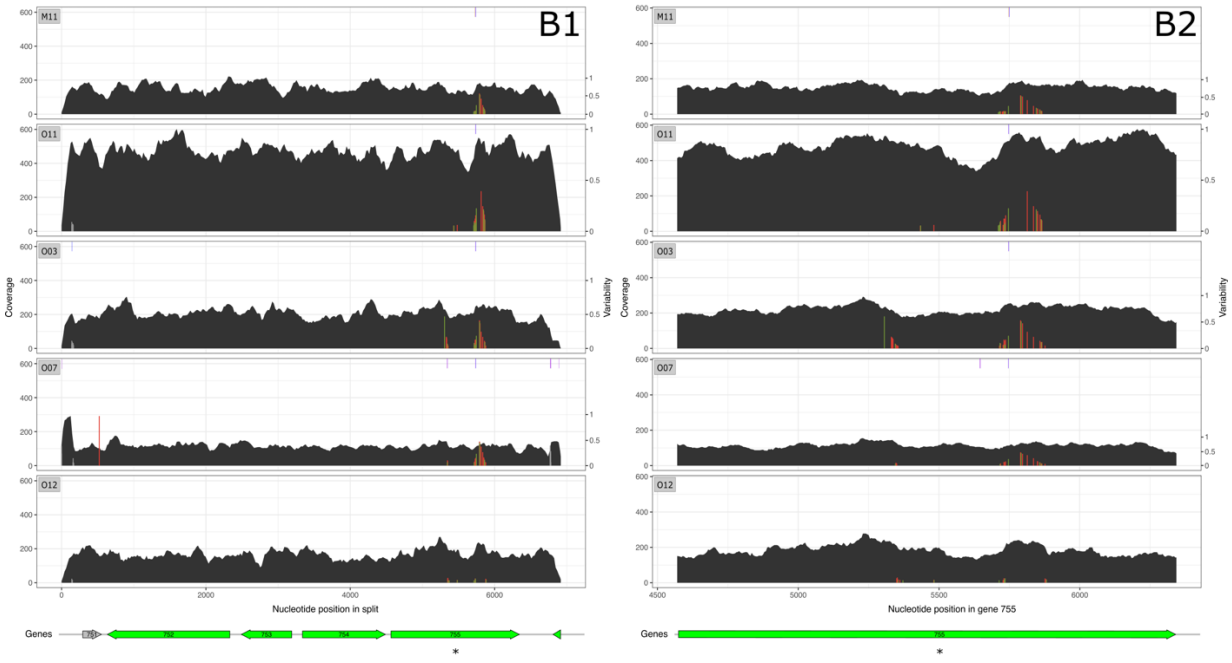

**Supplementary Figure 6 – Representation of the filtered SNVs (entropy and departure from consensus >= 0.2 and without coverage outliers) in the *Wolbachia* MAG O07**. From inner to outer and for each SNV (represented by a leaf), the layers correspond to i) the departure from the consensus (max value set to 1), ii) the Shannon entropy (max value set to 1), iii) the coverage (max value set to 2500X), iv) the position in codon (from 1st to 3rd if the SNV was detected in a coding gene and 0th if it not), v) the competing nucleotides (the two most nucleotides occurring in the SNV), vi) the presence of SNV in a coding gene call, non-coding gene or "out of gene", vii) the presence in *w*SCGs shared by all the MAGs and reference genomes, viii) the presence of the SNV in a gene annotated to a *cid* gene (from Bonneau et al. 2018 [6]), ix) the presence of the SNV in a gene annotated to a MLST or a *wsp* gene (from PubMLST [7]), and x) the annotation of prophage WO regions.



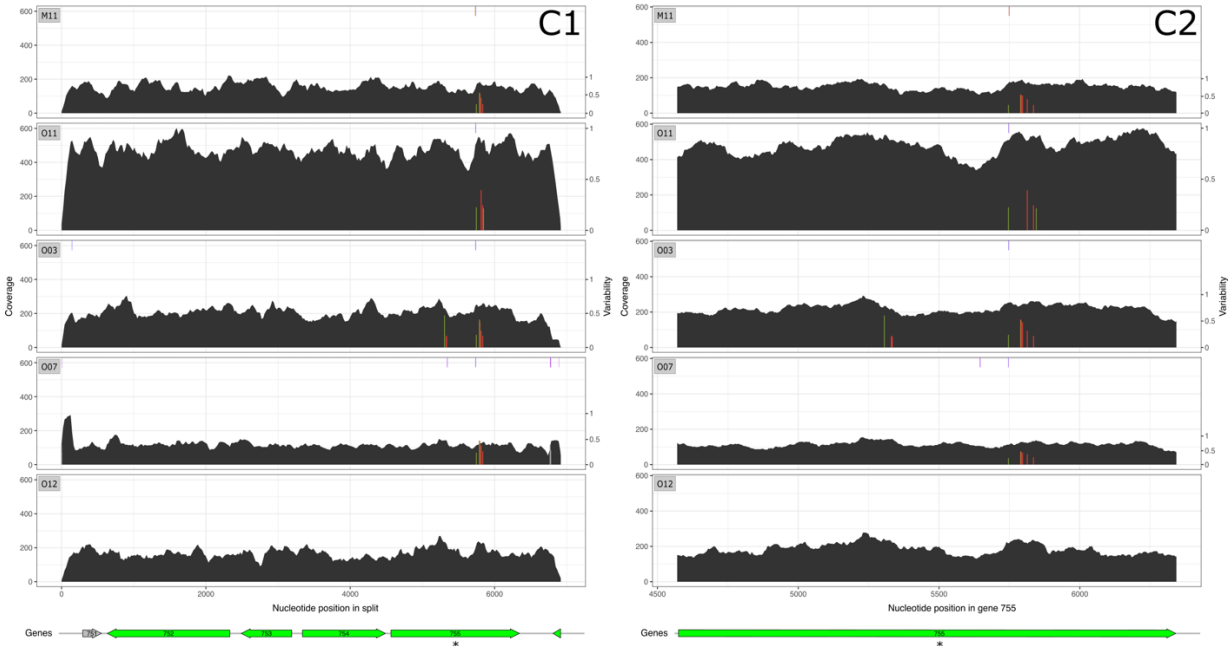

**Supplementary Figure 7 – Representation of the filtered SNVs (entropy and departure from consensus >= 0.2 and without coverage outliers) in the *Wolbachia* MAG O12**. From inner to outer and for each SNV (represented by a leaf), the layers correspond to i) the departure from the consensus (max value set to 1), ii) the Shannon entropy (max value set to 1), iii) the coverage (max value set to 2500X), iv) the position in codon (from 1st to 3rd if the SNV was detected in a coding gene and 0th if it not), v) the competing nucleotides (the two most nucleotides occurring in the SNV), vi) the presence of SNV in a coding gene call, non-coding gene or "out of gene", vii) the presence in *w*SCGs shared by all the MAGs and reference genomes, viii) the presence of the SNV in a gene annotated to a *cid* gene (from Bonneau et al. 2018 [6]), ix) the presence of the SNV in a gene annotated to a MLST or a *wsp* gene (from PubMLST [7]), and x) the annotation of prophage WO regions.


A1
M11
O11
O03
O07
O12
Coverage
Variability
Nucleotide position in split
Genes
*


A2
M11
O11
O03
O07
O12
Coverage
Variability
Nucleotide position in gene 755
Genes
*

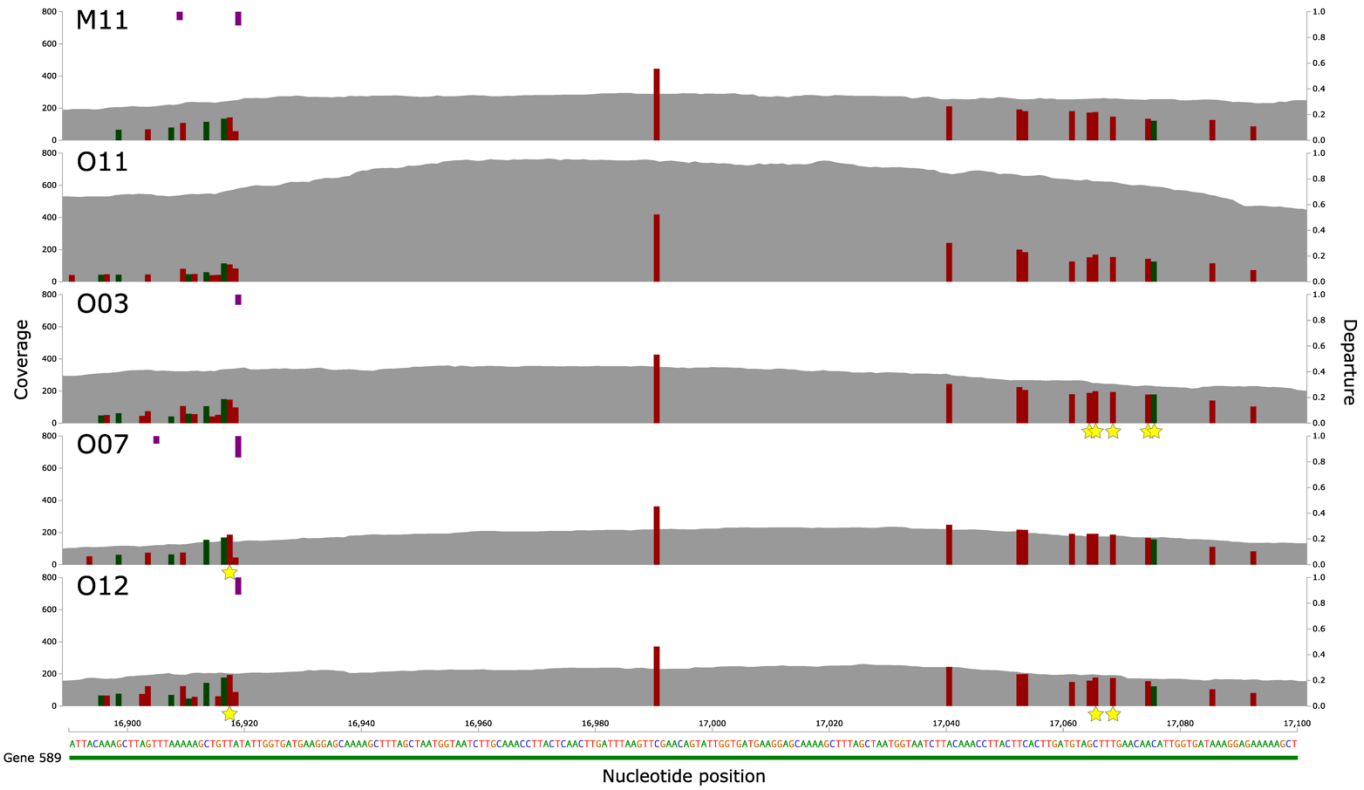

B1
M11
O11
O03
O07
O12
Coverage
Variability
Nucleotide position in split
Genes
*


B2
M11
O11
O03
O07
O12
Coverage
Variability
Nucleotide position in gene 755
Genes
*

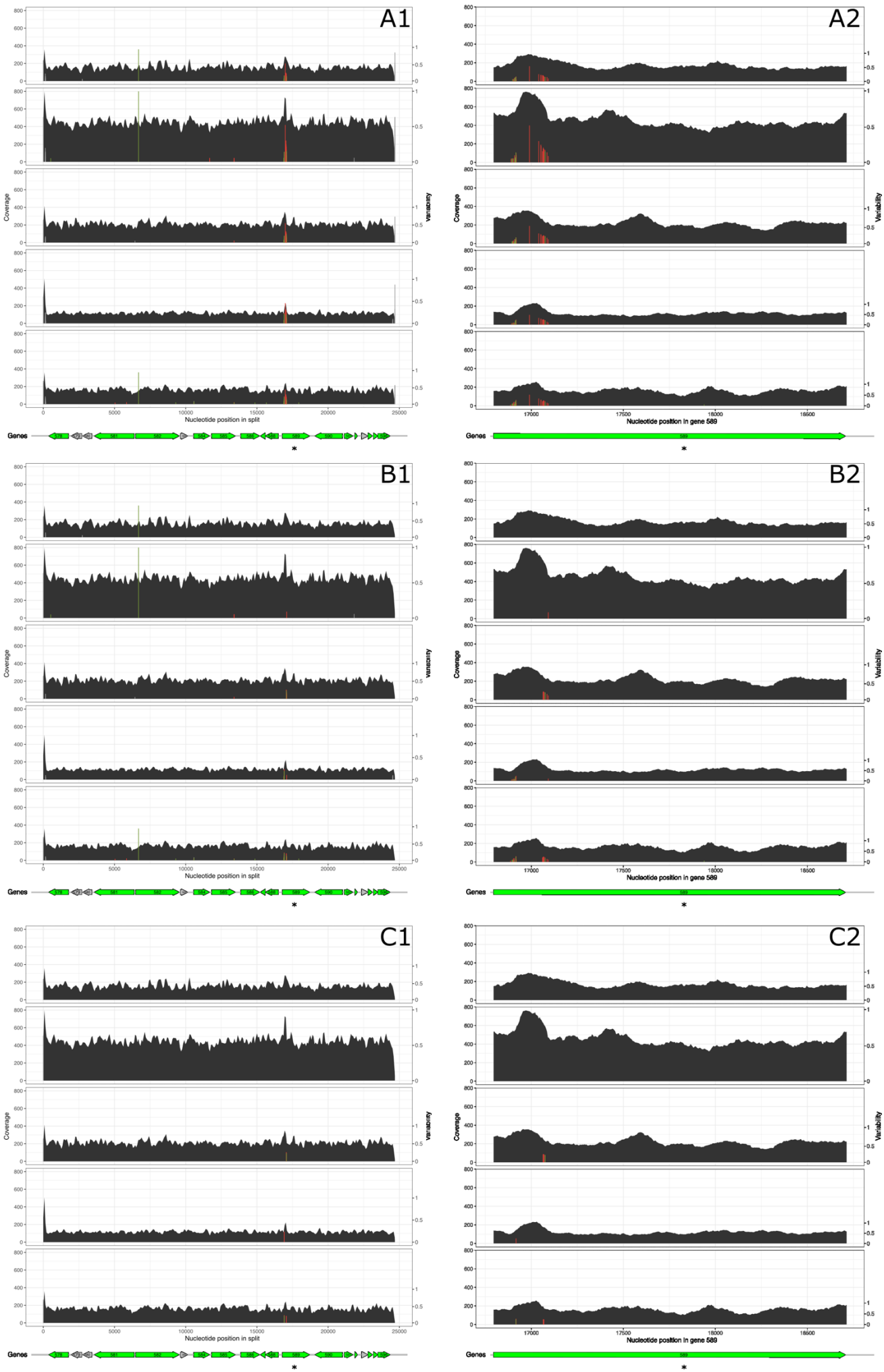

C1
M11
O11
O03
O07
O12
Coverage
Variability
Nucleotide position in split
Genes
*


C2
M11
O11
O03
O07
O12
Coverage
Variability
Nucleotide position in gene 755
Genes
*

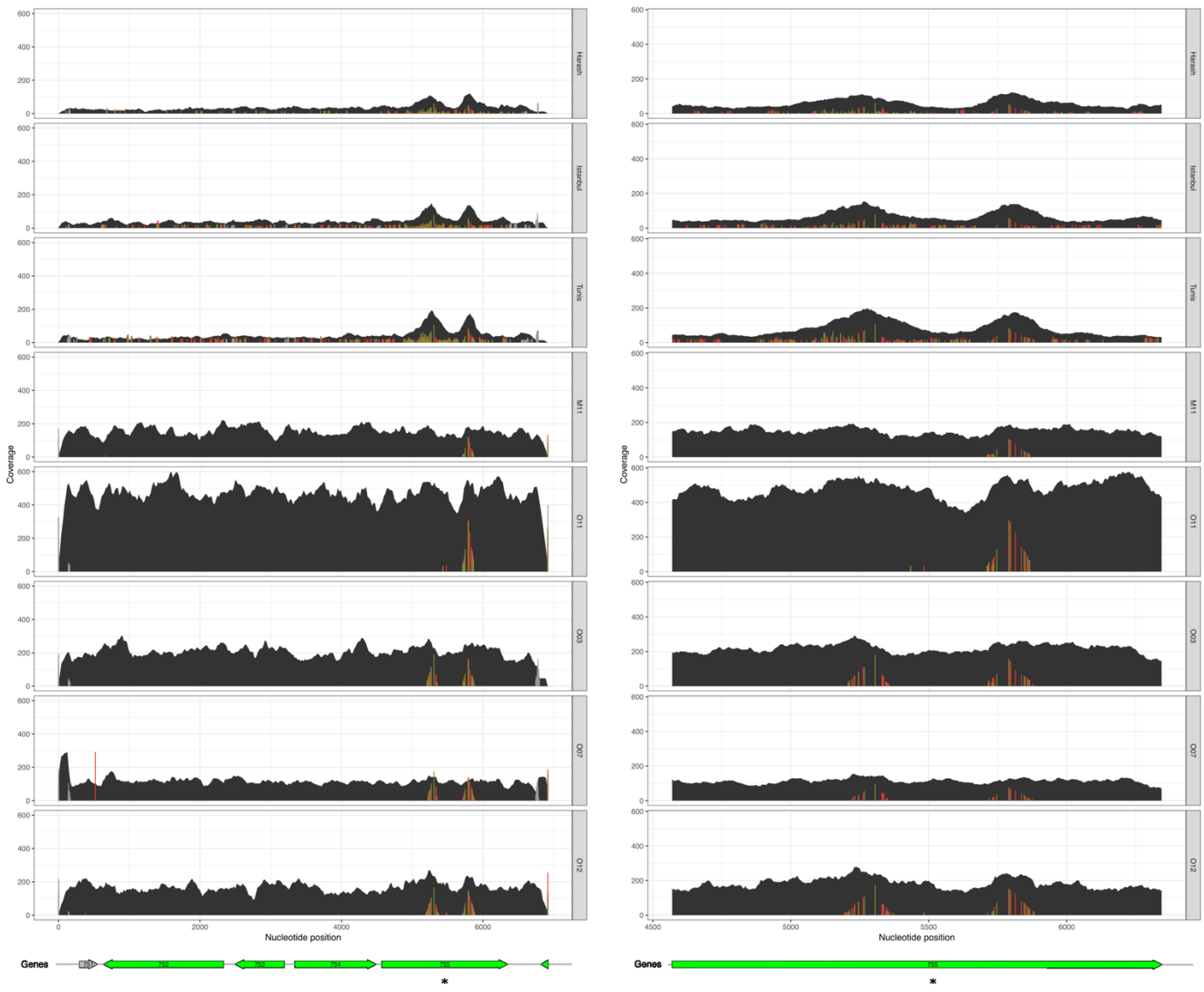
**Supplementary Figure 8** – **SNVs following recruitments of reads from five metagenomes (M11, O11, O03, O07, O12) onto *Wolbachia* MAG O11 for split 98617** A1. Raw SNVs occurring in the full split 98617 containing gene 755 defined as *w*SCG annotated to an "Ankyrin repeat function" by COG20 (Supplementary Table 6). A2. Zoom of the raw SNVs occurring in the gene 755. B1. SNVs without coverage outliers occurring in the full split 98617. B2. Zoom of SNVs without coverage outliers occurring in gene 755. C1. Filtered SNVs (without coverage outliers and including filtration on entropy and departure from consensus) occurring in the full split 98617. C2. Filtered SNVs occurring in gene 755. In each panel, each layer represents the coverage in each metagenome and SNVs are represented by color bars starting from the bottom across coverage values (green bars correspond to nucleotide difference in the 3rd position of the codon, red bars to nucleotide difference in the 1st or the 2nd one and the grey bars to nucleotide difference occurring outside the genes). The left y-axis corresponds to coverage values and the right y-axis corresponds to the variability values (departure from reference). On the bottom, COG-annotated genes are colored in green and non-annotated genes in grey. A star (*) below the gene indicates a gene previously defined as *w*SCG.



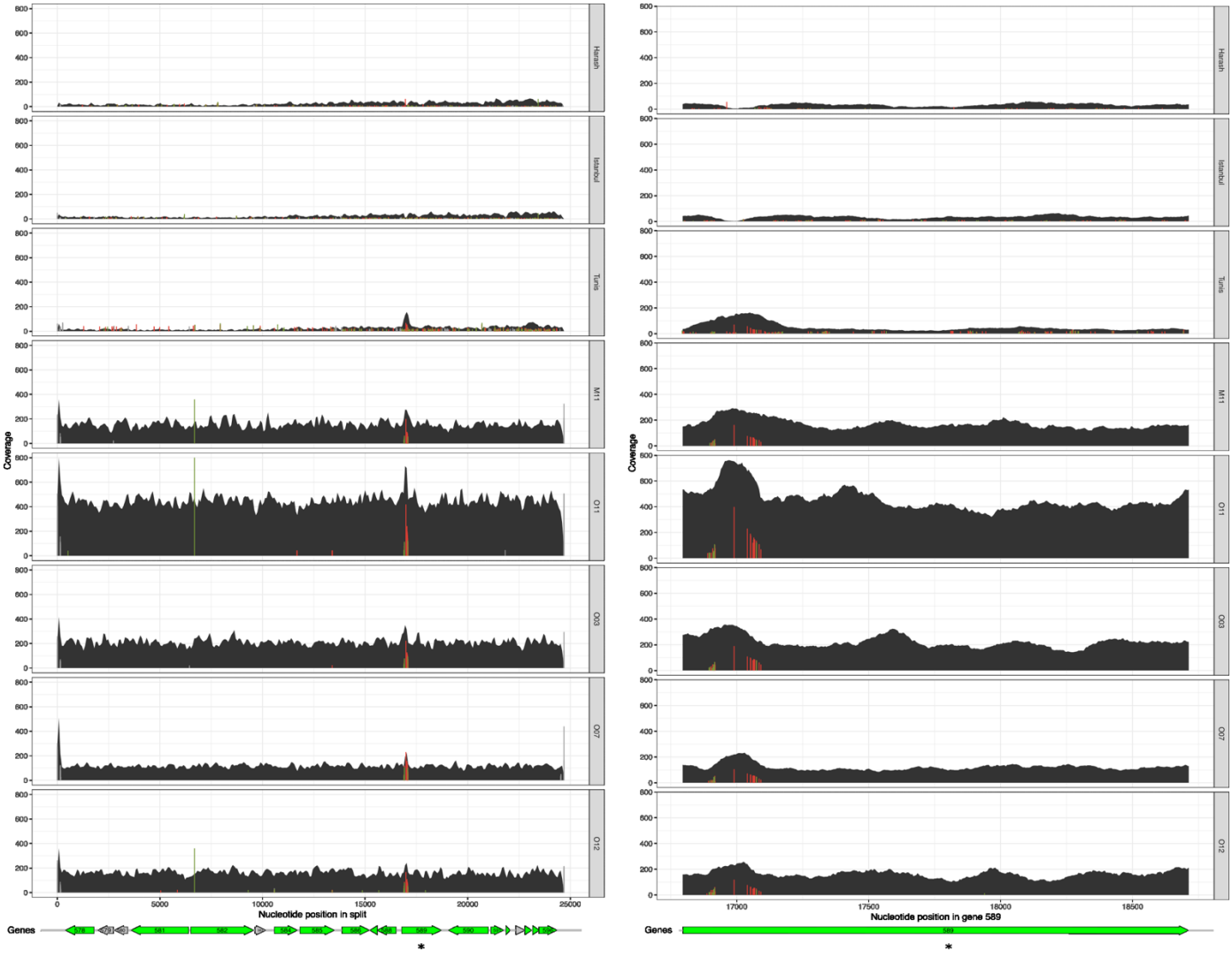

**Supplementary Figure 9 – SNVs following recruitment of reads from five metagenomes (M11, O11, O03, O07, O12) onto *Wolbachia* MAG O03 for a portion of gene 589 (annotated to an "ATP-dependent 26S proteasome regulatory subunit") from split 128791**. The x-axis corresponds to the nucleotide position in the portion of gene 589 (from position 16,890 to 17,100). The reference nucleotide bases for each nucleotide position are described just above by letters (A, C, T, G). The left-y axis corresponds to the coverage value for each metagenome. SNVs are represented by color bars starting from the bottom colored in red when SNV falls at 1st or 2nd position in codon and colored in green when a SNV falls at the 3rd position in codon. The right-y axis corresponds to the variability value of the SNVs (departure from reference). Also, indels are represented by purple bars starting from the top. To note, all the SNVs (without any filtration) are represented in this figure but a yellow star indicates SNVs that remain after the filtration steps (removing of SNVs identified as coverage outliers and SNVs with entropy or departure from consensus < 0.2).


A1
A2
B1
B2
C1
C2
Coverage
Variability
Nucleotide position in split
Nucleotide position in gene 589
Genes
589
*

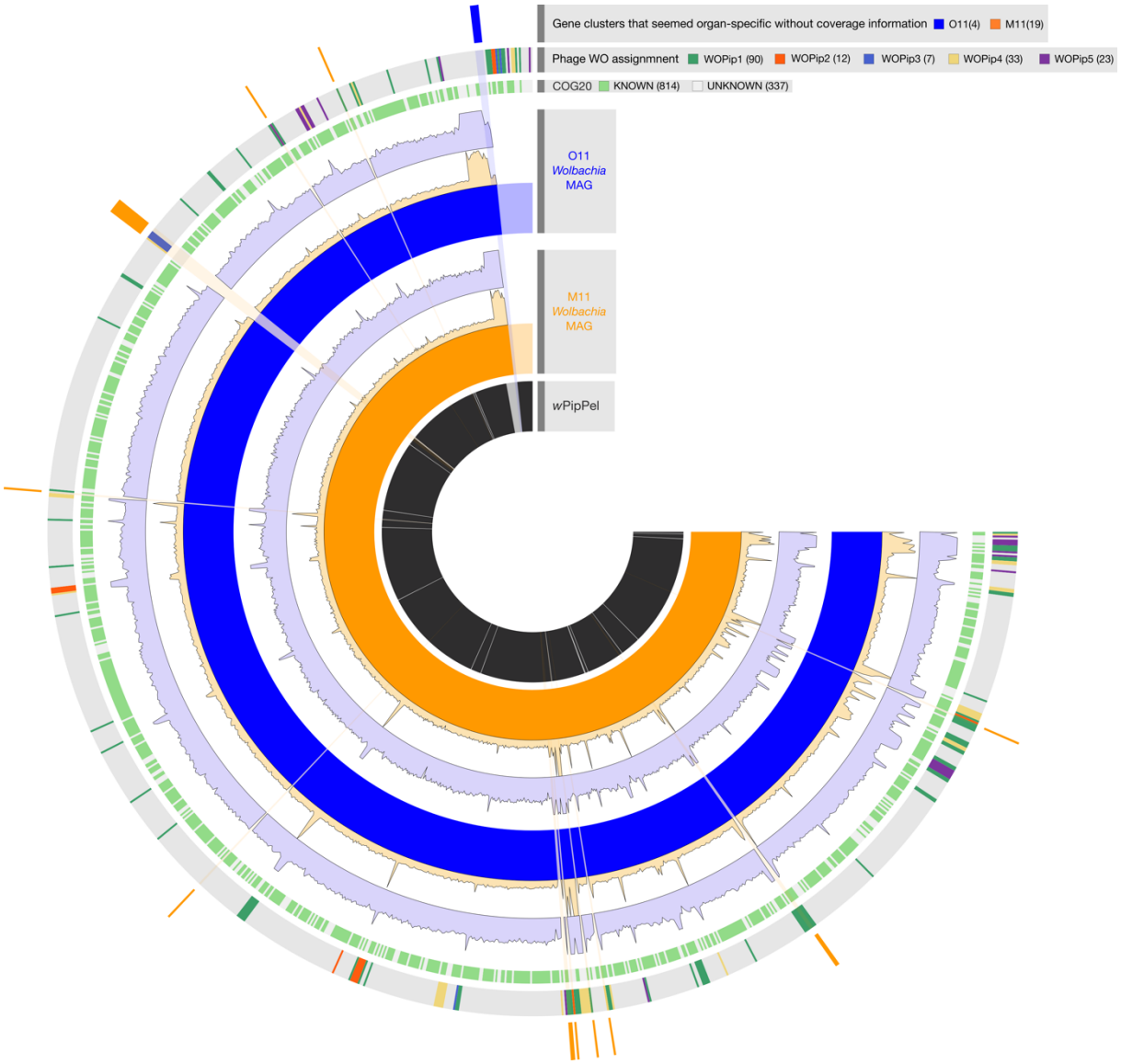
**Supplementary Figure 10 - Read recruitments for split 128791 from *Wolbachia* O03 MAG across five metagenomes (M11, O11, O03, O07, O12) and resulting SNVs**: A1. Raw SNVs occurring in the full split 128791 containing the gene 589 defined as *w*SCGs annotated to an "ATP-dependent 26S proteasome regulatory subunit" by COG20 (Supplementary Table 6). A2. Zoom of the raw SNVs occurring in gene 589. B1. SNVs without coverage outliers occurring in the full split 128791. B2. Zoom of SNVs without coverage outliers occurring in gene 589. C1. Filtered SNVs (without coverage outliers and including filtration on entropy and departure from consensus) occurring in the full split 128791. C2. Filtered SNVs occurring in gene 589. In each panel, each layer represents the coverage in each metagenome and SNVs are represented by color bars starting from the bottom across coverage values (green bars correspond to nucleotide difference in the 3rd position of the codon, red bars to nucleotide difference in the 1st or the 2nd one and the grey bars to nucleotide difference occurring outside the genes). The left y-axis corresponds to coverage values and the right y-axis corresponds to the variability values (departure from reference). On the bottom, COG-annotated genes are colored in green and non-annotated genes in grey. A star (*) below the gene indicates the gene previously defined as *w*SCG.



**Supplementary Figure 11 - Read recruitments onto *Wolbachia* MAG O11 from ovary metagenomes collected at the same location (M11, O11, O03, O07, O12) and from egg *metagenomes* collected at geographical distant locations (Harash, Istanbul and Tunis from [6])** in split 98617 containing gene 755 defined as *w*SCG annotated to an "Ankyrin repeat function" by COG20 (Supplementary Table 6). In each panel, each layer represents the coverage in each metagenome and SNVs are represented by color bars starting from the bottom across coverage values (green bars correspond to nucleotide difference in the 3rd position of the codon, red bars to nucleotide difference in the 1st or the 2nd one and the grey bars to nucleotide difference occurring outside the genes). The left y-axis corresponds to coverage values and the right y-axis corresponds to the variability values (departure from reference). On the bottom, COG-annotated genes are colored in green and non-annotated genes in grey. A star (*) below the gene indicates a gene previously defined as *w*SCG. Here, SNVs are not filtered because the analyzed metagenomes have very different coverage values. To note, anvi'o identified SNV outliers as nucleotide positions that occur in positions that are way too high or way too low considering the coverage values of all nucleotides in a given contig for a given MAG. Thus, a SNV can be identified as outlier in a MAG and not in another one at the same position because each MAG have a different mean coverage for a given split. Here, the coverage of the splits varied considerably from one MAG to another with a lower coverage in the egg metagenomes. We therefore herein did not to apply any filter to keep a maximum of shared SNVs.



**Supplementary Figure 12 - Read recruitments onto *Wolbachia* MAG O03 from ovary metagenomes collected at the same location (M11, O11, O03, O07, O12) and from egg metagenomes collected at geographical distant locations (Harash, Istanbul and Tunis from [6])** in split 128791 containing gene 589 defined as *w*SCG annotated to an "ATP-dependent 26S proteasome regulatory subunit" by COG20 (Supplementary Table 6). In each panel, each layer represents the coverage in each metagenome and SNVs are represented by color bars starting from the bottom across coverage values (green bars correspond to nucleotide difference in the 3$^{rd}$ position of the codon, red bars to nucleotide difference in the 1$^{st}$ or the 2$^{nd}$ one and the grey bars to nucleotide difference occurring outside the genes). The left y-axis corresponds to coverage values and the right y-axis corresponds to the variability values (departure from reference). On the bottom, COG-annotated genes are colored in green and non-annotated genes in grey. A star (*) below the gene indicates a gene previously defined as *w*SCG. Here also, SNVs are not filtered because the analyzed metagenomes have very different coverage values. To note, anvi'o identified SNV outliers as nucleotide positions that occur in positions that are way too high or way too low considering the coverage values of all nucleotides in a given contig for a given MAG. Thus, a SNV can be identified as outlier in a MAG and not in another one at the same position because each MAG have a different mean coverage for a given split. Here, the coverage of the splits varied considerably from one MAG to another with a lower coverage in the egg metagenomes. For this reason, we herein did not to apply any filter to keep a maximum of shared SNVs.




**Supplementary Figure 13 –Metapangenome of *Wolbachia* across midgut M11 and ovary O11 samples** with gene clusters organized based on *Wolbachia* M11 MAG synteny because of its higher length compared to O11's one. The black layer corresponds to the gene clusters from the *Wolbachia* reference genome *w*PipPel. Each MAG is represented by three layers: the gene clusters (colored in orange for M11 and blue for O11), the coverage of raw reads from M11 sample (colored in orange) and the coverage of raw reads from O11 sample (colored in blue). The maximum value shown in these coverage layers is set at 800X to improve their visualization. The three last layers represent COG20 annotation, the "phage-like" gene clusters identified from [8] in the *w*PipPel reference genome, and the gene clusters that seem to be present only in O11 or M11 *Wolbachia* MAG.

## References


1. Alberti A, Poulain J, Engelen S, Labadie K, Romac S, Ferrera I, et al. Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. *Sci Data.* 2017; **4**: 170093.
2. Gruber-Vodicka HR, Seah BKB, Pruesse E. phyloFlash: rapid small-subunit rRNA profiling and targeted assembly from metagenomes. *mSystems.* 2020; **5**: e00920-20.
3. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 2013; **41**: D590–D596.
4. Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. Using SPAdes de novo assembler. *Curr Protoc Bioinforma.* 2020; **70**: e102.
5. R Core Team. R: The R Project for Statistical Computing. R version 3.6.3. https://www.r-project.org/. Accessed 7 Jun 2021.
6. Bonneau M, Atyame C, Beji M, Justy F, Cohen-Gonsaud M, Sicard M, et al. *Culex pipiens* crossing type diversity is governed by an amplified and polymorphic operon of *Wolbachia. Nat Commun.* 2018; **9**: 319.
7. Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res.* 2018; **3**: 124.
8. Klasson L, Walker T, Sebaihia M, Sanders MJ, Quail MA, Lord A, et al. Genome evolution of *Wolbachia* strain *w*Pip from the *Culex pipiens* group. *Mol Biol Evol.* 2008; **25**: 1877–1887.