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

Supplementary Notes

Supplementary Note 1 –Intra-MAGs variability. We characterized a mean of 3,742 raw SNVs within our *Wolbachia* MAGs (intra-MAG variability) with a mean coverage of 367X (Supplementary Table 6). After our filtration steps (entropy > 0.2 and departure from reference > 0.2), we obtained a final number of 293 SNVs on average with a mean coverage value of 533X in our samples (Supplementary Table 7). 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 7).

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, Table 2, Supplementary Table 7), M11, O03, O07 and O12 (Figure 1 - panel B, Table 2, Supplementary Table 7), respectively. Among those gene-call SNVs, we counted a mean number of 41, 35 and 94 SNVs at the 1st, the 2nd and the 3rd base position, respectively (Figure 1, Table 2, Supplementary Table 7).

From these SNVs, we reported an average number of 140 SCVs (after similar filtration metrics including entropy and departure from reference) in coding genes with a mean coverage of 483X in our MAGs (Supplementary Table 12, see Supplementary Table 8 and Supplementary Table 9 for the raw and filtered intra-MAGs SCVs in details, respectively). From them, we counted a mean number of 67 SAAVs (after filtration) with a mean coverage of 463X (Supplementary Table 12, see Supplementary Tables 10 and Supplementary Table 11 for the raw and filtered intra-MAGs SAAVs in details, respectively).

Supplementary Note 2 – Inter-MAGs variability. We reported 23,999 raw SNVs on average in our 5 *Wolbachia* MAGs with a mean coverage of 421X (Supplementary Table 14). After filtration, we counted 2,487 inter-MAG SNVs on average in our 5 *Wolbachia* MAGs with a mean coverage value of 625X (Supplementary Table 15). 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 16). In details, we reported a mean of 226 previously identified intra-MAG SNVs from each MAG (157/281, 155/186, 135/179, 591/701 and 92/119) that are shared by at least two metagenomes (Supplementary Table 13). Among them, 32 SNVs on average were identified as SSG-metagenomes, 5 in SSG-references, and 5 in both SSG-metagenomes and SSG-references (Supplementary Table 13).

Supplementary Note 3 – SNVs in *cid* genes (outside of SSG). We reported 26 SNVs identified in multiple genes that were annotated to multiple *cidB* variants in O07 (from Bonneau et al. 2018[38]) but the latest didn't correspond to SSG (Supplementary Table 13, sheet 4 and Supplementary Figure 4). Among them, 6, 11 and 9 were found at the 1st, 2nd and 3rd positions of multiple codons (Supplementary Table 13, sheet 4 and Supplementary Figure 4).

Supplementary Tables

Supplementary Table 1 – 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 2 – 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 3 – *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 4 – GC (Gene Cluster) identified during the pangenomics analyses performed on i) all metagenomes (Sheet 1), ii) each metagenome (M11, O11, O03, O07 and O12) and selected *Wolbachia* reference genomes *w*PipPel, *w*PipMol, *w*PipJHB (from Sheet 2 to 6). Tables report for each identified GC: its ids (unique_id, gene_cluster_id), the bin and genome names it has been identified from (bin_name, genome_name), the gene caller id of its gene (gene_callers_id), the number of genomes where its occurs (num_genomes_gene_cluster_has_hits), the number of genes occurring in it (num_genes_in_gene_cluster), the maximum number of paralogs (max_num_paralogs), the Single-Copy Core genes (SCG) identified by the pangenomics analysis, the functional, geometric and combined homogeneity indices (functional_homogeneity_index, geometric_homogeneity_index, combined_homogeneity_index), the KEGG assignation information (KEGG_Module_ACC, KEGG_Module, , KEGG_Class_ACC, KEGG_Class), 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, KEGG_Module_ACC) and the corresponding amino acid sequence (aa_sequence).

Supplementary Table 5 – SSG (Shared Single-Copy core Gene) identified among the unique GC computed during the pangenomics analyses performed on all metagenomes (Sheet 1), each metagenome and selected *Wolbachia* reference genomes (Sheet 2 to 5). Tables reports for each unique GC: its SSG-metagenomes status (SSG_MAGs_'sample-id' in Sheet 1) and the metagenomes or references they are shared with (SSG_'sample-id'_shared_in_metagenomes, Sheet 1) or its SSG-references status (SSG_references) and the references they are shared with (SSG_shared_in_references in Sheets 2-6).

Supplementary Table 6 – Raw SNV table (intra-MAG variability table) for each metagenome (Sheet 1; M11, Sheet 2; O11, Sheet 3; O03, Sheet 4; O07, Sheet 5; O12). Table reports for each SNV position: the 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/.

Supplementary Table 7 – Filtered SNV table (intra-MAG variability table) for each metagenome (Sheet 1; M11, Sheet 2; O11, Sheet 3; O03, Sheet 4; O07, Sheet 5; O12). Table reports the same information as in Supplementary Table 6 but for filtered SNVs.

Supplementary Table 8 – Raw SCV table (intra-MAG variability table) for each metagenome (Sheet 1; M11, Sheet 2; O11, Sheet 3; O03, Sheet 4; O07, Sheet 5; O12). 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 consensus and pN, pS, nN, nS reference, pN, pS, nN, nS popular consensus). All these SCV information are also explained at the following link https://merenlab.org/2015/07/20/analyzing-variability/.

Supplementary Table 9 - Filtered SCV table (inter-MAG variability table) identified from the filtered SNVs in each metagenome (Sheet 1; M11, Sheet 2; O11, Sheet 3; O03, Sheet 4; O07, Sheet 5; O12). Table reports the same information as in Supplementary Table 8 but for filtered SCVs. In addition, it reports the *cid* genes assignment

(cid_genes_from_Bonneau_et_al._2018[1], Bonneau_pct_alignment), the MLST and *wsp* genes assignment (MLST_and_wsp_genes_from_PubMLST[2], MLST_wsp_pct_alignment), the metagenome and the *Wolbachia* reference genome level (SSG_MAGs_'sample-id', SSG_references) and the metagenomes or the references where the SSG are shared (SSG_shared_in_metagenomes, SSG_shared_in_references).

Supplementary Table 10 - Raw SAAV table (inter-MAG variability table) identified from the filtered SNVs in each metagenome (Sheet 1; M11, Sheet 2; O11, Sheet 3; O03, Sheet 4; O07, Sheet 5; O12). 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/.

Supplementary Table 11 - Filtered SAAV table (inter-MAG variability table) identified from the filtered SNVs in each metagenome (Sheet 1; M11, Sheet 2; O11, Sheet 3; O03, Sheet 4; O07, Sheet 5; O12). Table reports the same information as in Supplementary Table 10 but for filtered SAAVs.

Supplementary Table 12 - For each sample, the table reports the count of the intra-MAGs SNVs, SCVs and SAAVs (after filtration; entropy and departure from consensus >= 0.2) that occurred in SSG-metagenomes, SSG-metagenomes annotated to *cid* genes, SSG-references, and both SSG-metagenomes and references. Coverage and departure from consensus are reported for SNVs, SCVs and SAAVs. Codon positions are reported only for SNVs. The pN/pS (non-synonymity polymorphism on synonymity polymorphism) is reported for SCVs.

Supplementary Table 13 – SSG and shared SNVs information from SNV table in each metagenome (Sheet 1; M11, Sheet 2; O11, Sheet 3; O03, Sheet 4; O07, Sheet 5; O12). Table reports the same information as in Supplementary Table 6. In addition, it reports the presence of the SNV in a SSG (as defined in the Material and Methods section) at the metagenome and the *Wolbachia* reference genome level (SSG_MAGs_'sample-id', SSG_references), the metagenomes or the references where the SSG are shared (SSG_shared_in_metagenomes, SSG_shared_in_references), the number of the SNV occurrence across the metagenomes and their corresponding sample id (Number_of_metagenomes_where_SNV_are_shared, SNV_shared_in_metagenomes), the *cid* genes assignment on the gene where the SNV occurred (cid_genes_from_Bonneau_et_al._2018[1], Bonneau_pct_alignment), the MLST and

wsp genes assignment (MLST_and_wsp_genes_from_PubMLST[2], 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 14 – Raw SNVs (inter-MAG) table for each metagenome (Sheet 1; M11, Sheet 2; O11, Sheet 3; O03, Sheet 4; O07, Sheet 5; O12). Table reports the same information as in Supplementary Table 6.

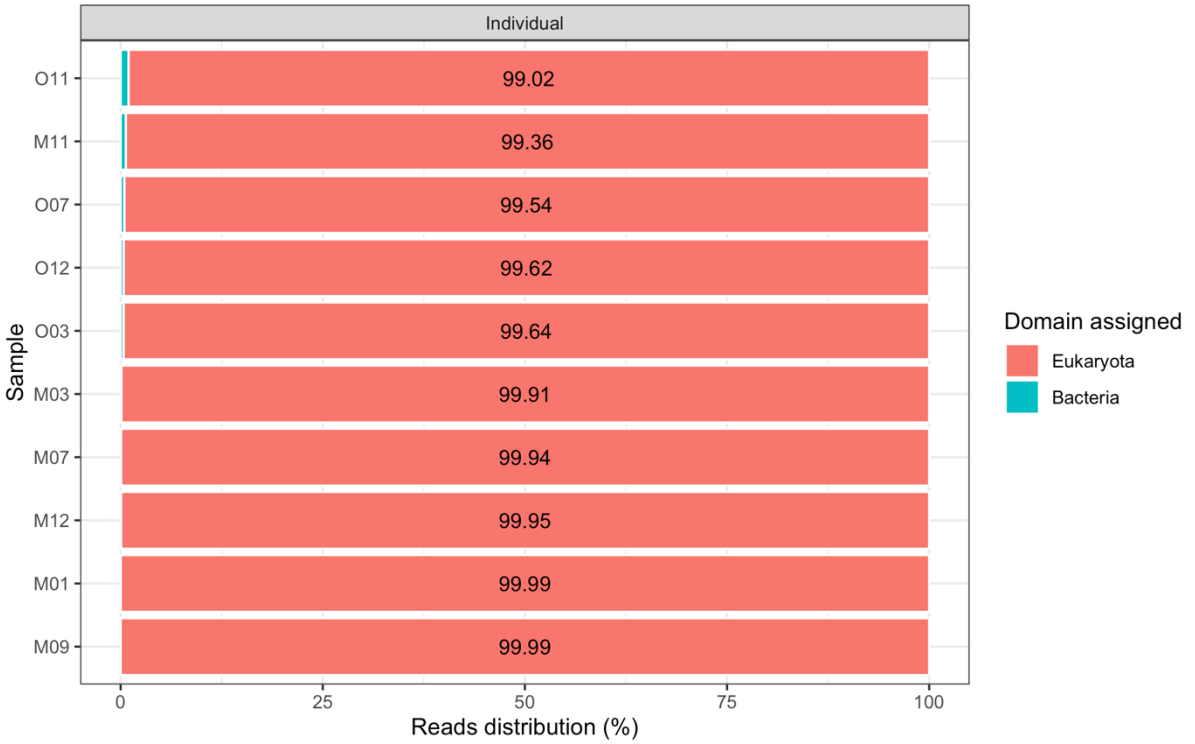
Supplementary Table 15 – Filtered SNVs (inter-MAG) table for each metagenome (Sheet 1; M11, Sheet 2; O11, Sheet 3; O03, Sheet 4; O07, Sheet 5; O12). Table reports the same information as in Supplementary Table 6.

Supplementary Table 16 – Filtered and shared SNVs (inter-MAG table for each metagenome (Sheet 1; M11, Sheet 2; O11, Sheet 3; O03, Sheet 4; O07, Sheet 5; O12). Table reports the same information as in Supplementary Table 6.

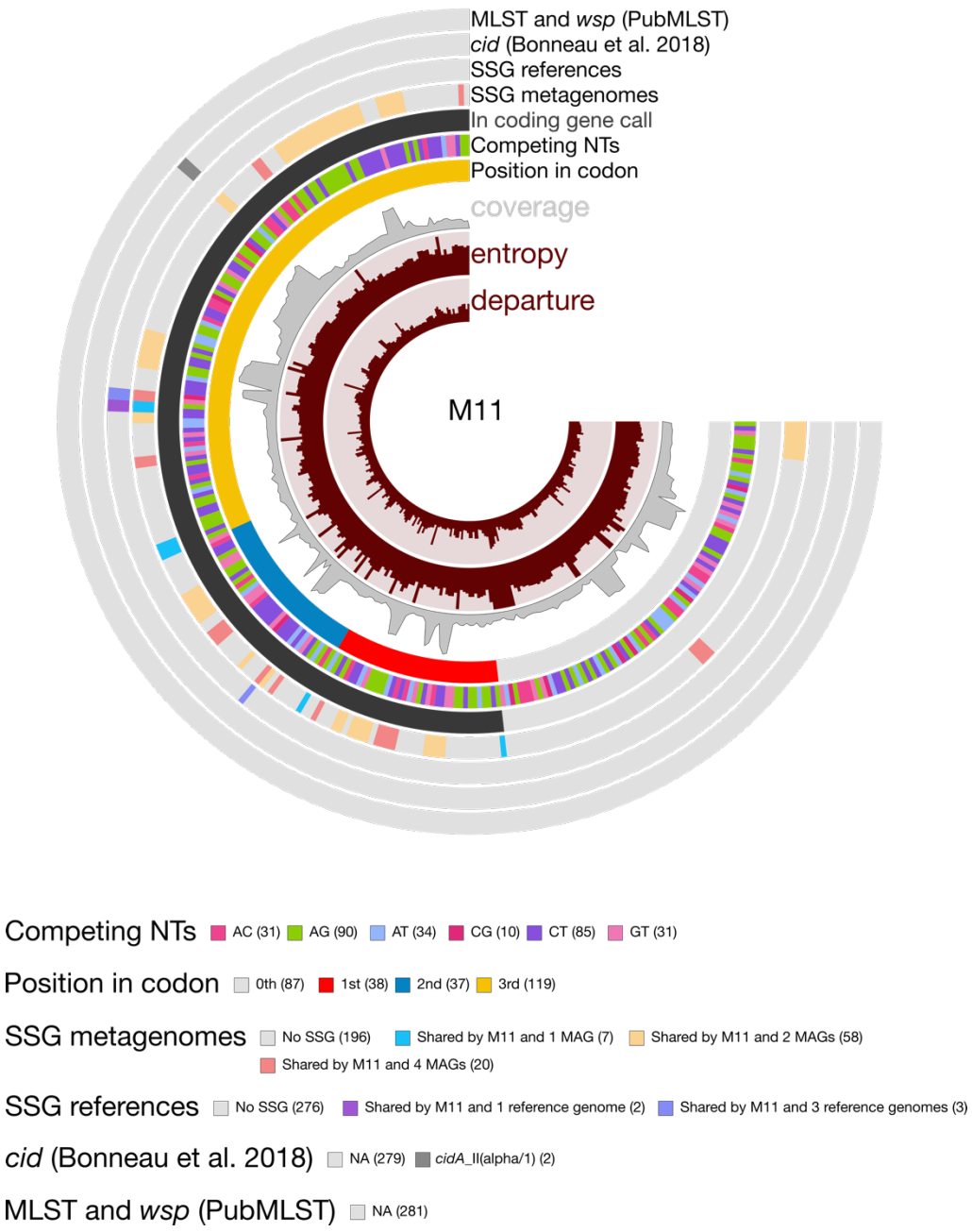
Supplementary Table 17 – Indels table for the five *Wolbachia* MAGs. (Sheet 1&2 ; raw indels and filtered indels (without outliers defined as cov_outlier_in_split or cov_outlier_in_contig equals to "1") in M11 *Wolbachia* MAG, Sheet 3&4; raw indels and filtered indels in O11 *Wolbachia* MAG, Sheets 5&6; raw and filtered indels in splits 98617, 118914 and 120111 from O11 *Wolbachia* MAG, Sheets 7&8; raw and filtered indels from O03 *Wolbachia* MAG, Sheets 9&10; raw and filtered indels from O07 *Wolbachia* MAG, Sheets 11&12; raw and filtered indels from O12 *Wolbachia* MAG). Table reports the sample where the Indel has been identified (sample_id), the split name where the indel has been indentified (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 18 – GC (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.

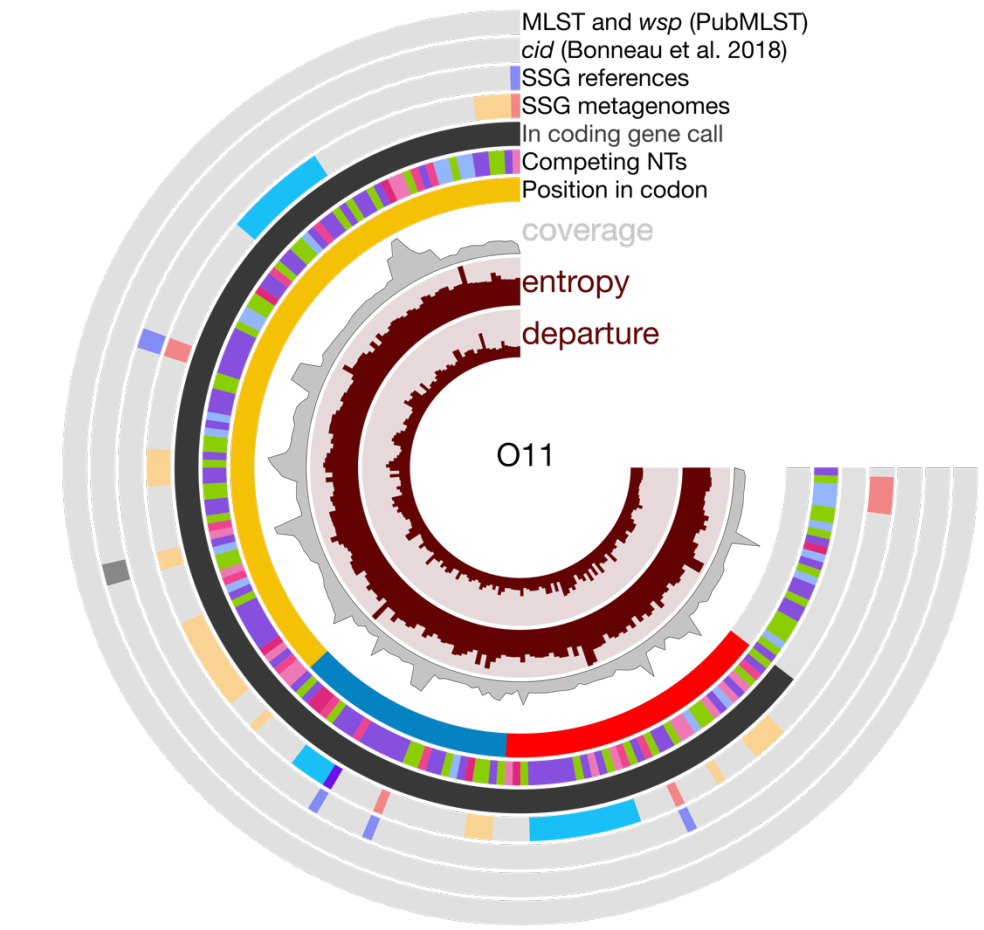
Supplementary Figures



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 eucaryotic 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 – Representation of the filtered SNVs (entropy and departure from consensus >= 0.2) in the *Wolbachia* M11 MAG. From inner to outer and for each SNV, 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 800X), 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 / absence of the SNV in a coding gene call, vii) the presence of the SNV in a gene shared by the sample and at least one another *Wolbachia* MAG, viii) the presence of the SNV in a gene shared by the sample and at least 1 *Wolbachia* reference genome, ix) the presence of the SNV in a gene annotated to a *cid* gene (from Bonneau et al. 2018[1]), x) the presence of the SNV in a gene annotated to a MLST or a *wsp* gene (from PubMLST[2]).



Competing NTs AC (13) AG (52) AT (21) CG (8) CT (76) GT (16)

Position in codon 0th (26) 1st (38) 2nd (30) 3rd (92)

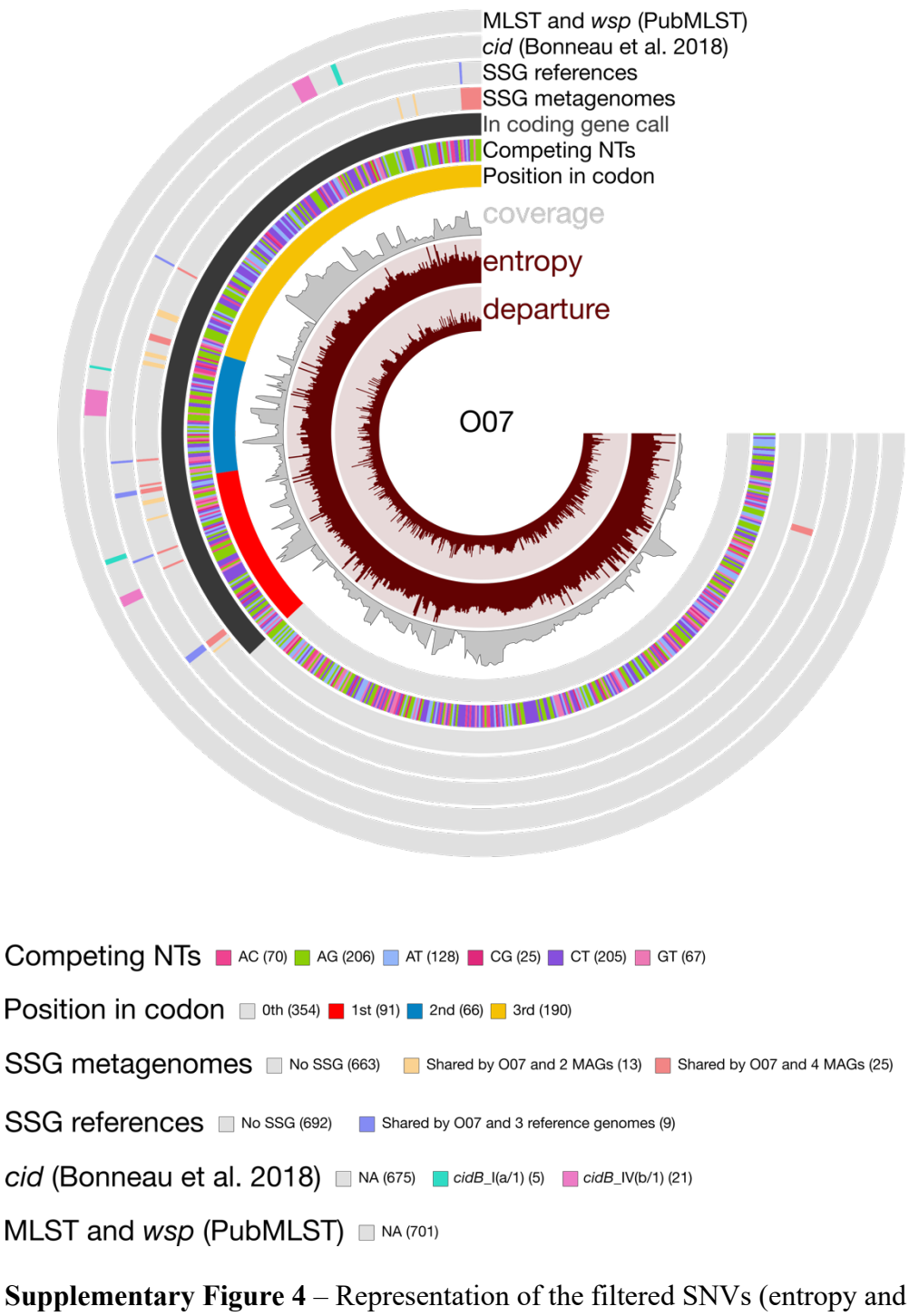
SSG metagenomes No SSG (120) Shared by O11 and 1 MAG (27) Shared by O11 and 2 MAGs (29) Shared by O11 and 3 MAGs (1) Shared by O11 and 4 MAGs (9)

SSG references No SSG (180) Shared by O11 and 3 reference genomes (6)

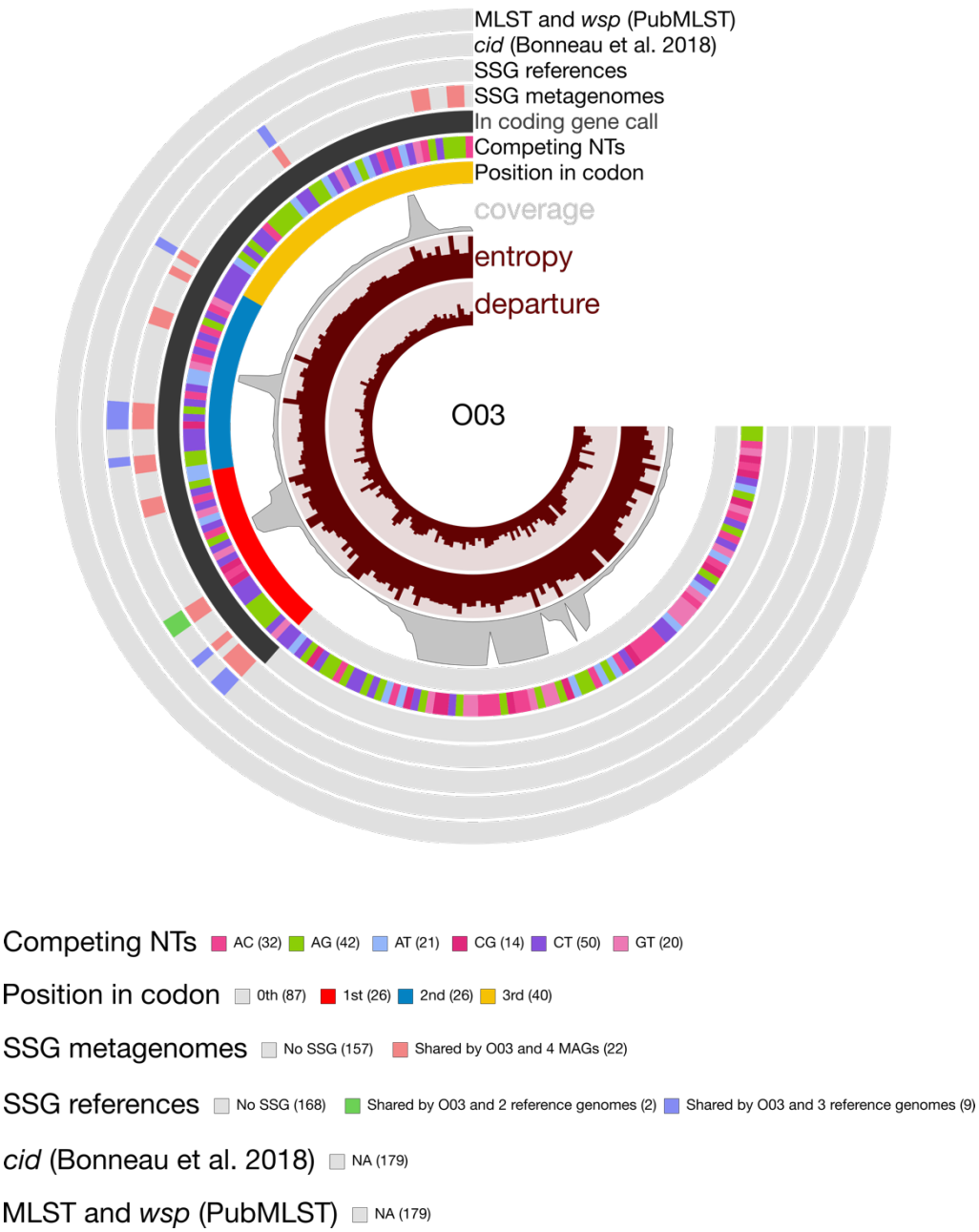
cid (Bonneau et al. 2018) NA (184) *cidA*_II(alpha/1) (2)

MLST and *wsp* (PubMLST) NA (186)

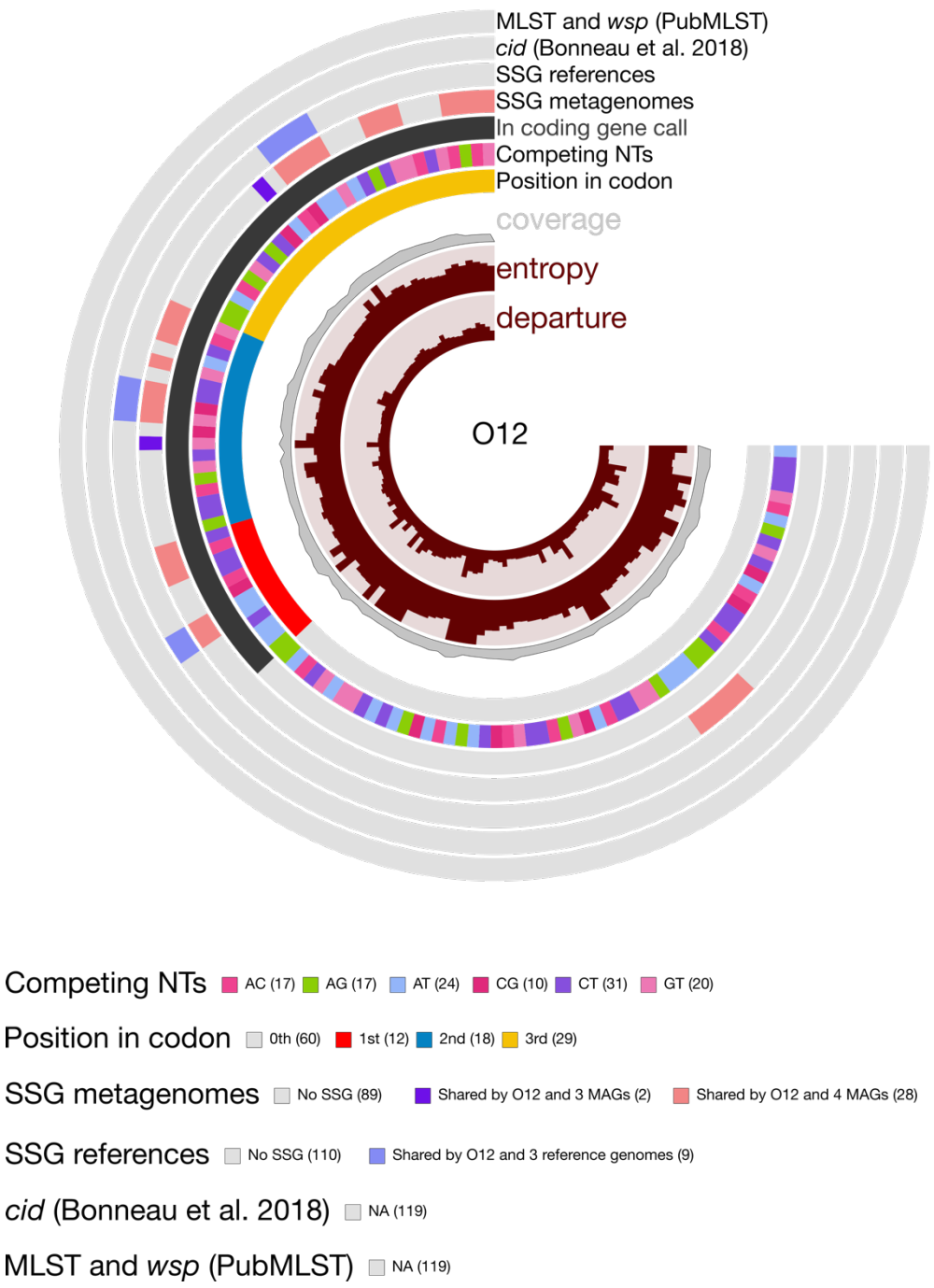
Supplementary Figure 3 – Representation of the filtered SNVs (entropy and departure from consensus >= 0.2) in the *Wolbachia* O11 MAG. From inner to outer and for each SNV, the layers correspond to i) the departure from the consensus (max value set to 1), ii) the Shannon entropy (max value set to 1), 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 / absence of the SNV in a coding gene call, vii) the presence of the SNV in a gene shared by the sample and at least one another *Wolbachia* MAG, viii) the presence of the SNV in a gene shared by the sample and at least 1 *Wolbachia* reference genome, ix) the presence of the SNV in a gene annotated to a *cid* gene (from Bonneau et al. 2018[1]), x) the presence of the SNV in a gene annotated to a MLST or a *wsp* gene (from PubMLST[2]).



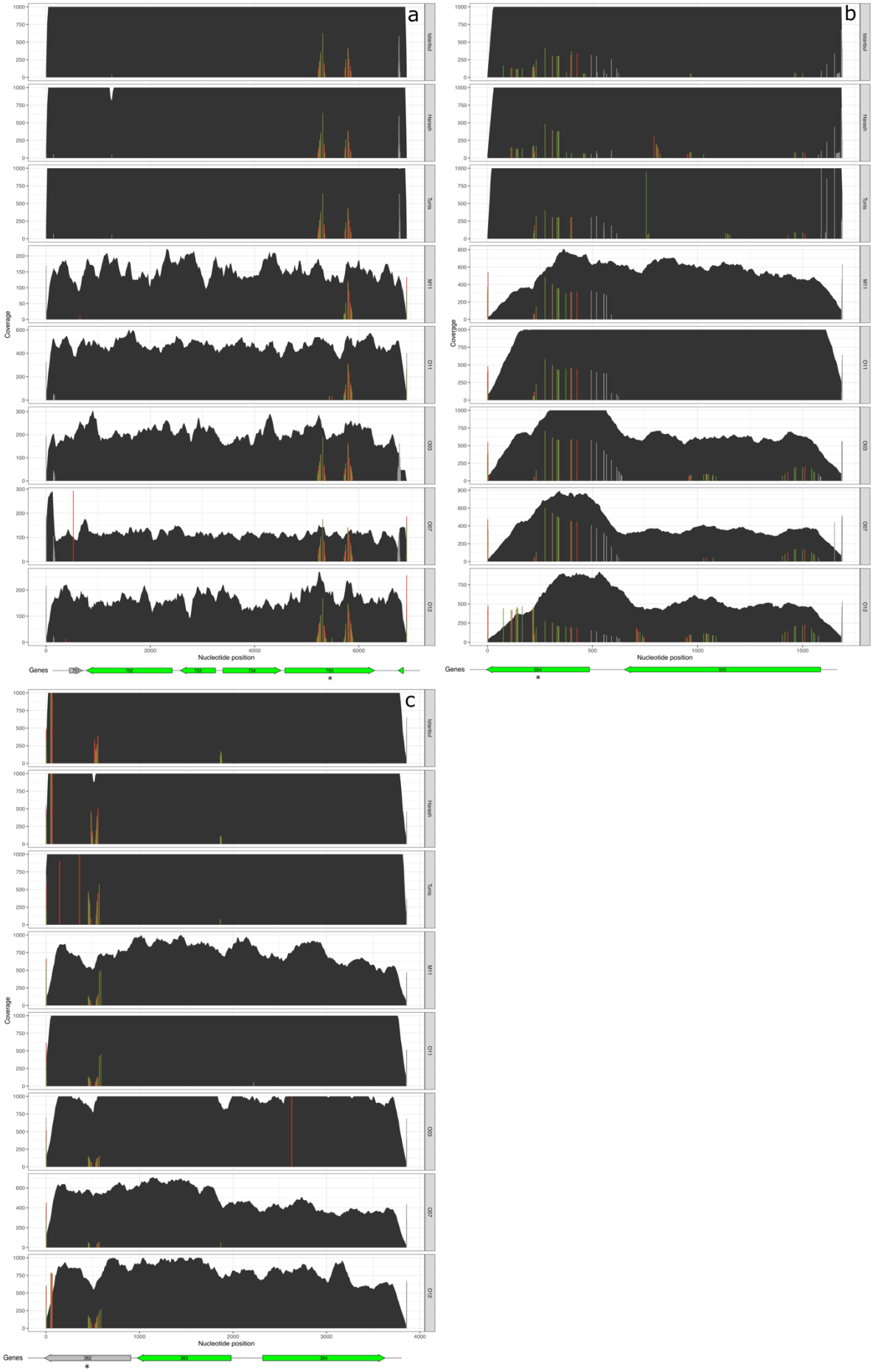
Supplementary Figure 4 – Representation of the filtered SNVs (entropy and departure from consensus >= 0.2) in the *Wolbachia* O07 MAG. From inner to outer and for each SNV, the layers correspond to i) the departure from the consensus (max value set to 1), ii) the Shannon entropy (max value set to 1), iii) the coverage (max value set to 800X), 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 / absence of the SNV in a coding gene call, vii) the presence of the SNV in a gene shared by the sample and at least one another *Wolbachia* MAG, viii) the presence of the SNV in a gene shared by the sample and at least 1 *Wolbachia* reference genome, ix) the presence of the SNV in a gene annotated to a *cid* gene (from Bonneau et al. 2018[1]), x) the presence of the SNV in a gene annotated to a MLST or a *wsp* gene (from PubMLST[2]).



Supplementary Figure 5 – Representation of the filtered SNVs (entropy and departure from consensus >= 0.2) in the *Wolbachia* O03 MAG. From inner to outer and for each SNV, 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 / absence of the SNV in a coding gene call, vii) the presence of the SNV in a gene shared by the sample and at least one another *Wolbachia* MAG, viii) the presence of the SNV in a gene shared by the sample and at least 1 *Wolbachia* reference genome, ix) the presence of the SNV in a gene annotated to a *cid* gene (from Bonneau et al. 2018[1]), x) the presence of the SNV in a gene annotated to a MLST or a *wsp* gene (from PubMLST[2]).



Supplementary Figure 6 – Representation of the filtered SNVs (entropy and departure from consensus >= 0.2) in the *Wolbachia* O12 MAG. From inner to outer and for each SNV, 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 800X), 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 / absence of the SNV in a coding gene call, vii) the presence of the SNV in a gene shared by the sample and at least one another *Wolbachia* MAG, viii) the presence of the SNV in a gene shared by the sample and at least 1 *Wolbachia* reference genome, ix) the presence of the SNV in a gene annotated to a *cid* gene (from Bonneau et al. 2018[1]), x) the presence of the SNV in a gene annotated to a MLST or a *wsp* gene (from PubMLST[2]).


a
b
c
Istanbul
Hatayah
Tunis
M11
O11
O03
O07
O12
Coverage
Nucleotide position
Genes

Supplementary Figure 7 - Read recruitments for *Wolbachia* MAG 011 across four metagenomes from organs extracted in mosquitoes collected at the same location (M11, O11, O03, O07, O11, O12) and three metagenomes from egg Culex mosquitoes collected at geographical distant locations (Istanbul, Harash and Tunis from[1]) in three selected splits: a. Split 98617 containing gene 755 defined as SSG (metagenomes and references) annotated to an "Ankyrin repeat function" by COG20 (Supplementary Table 13, sheet 2), b. Split 118914 containing gene 654 defined as SSG-metagenomes and annotated to a "DNA repair protein RadC" function by COG20 (Supplementary Table 13, sheet 2), c. Split 120111 containing the gene 362 defined as SSG-metagenomes and that we have annotated to a *cidA* variant, namely *cidA_II(alpha/1)* from[1] (Supplementary Table 13, sheet 2). 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). SNVs are not filtered because the analyzed metagenomes have very different coverage values.

References

1. Bonneau, M. *et al. Culex pipiens* crossing type diversity is governed by an amplified and polymorphic operon of *Wolbachia*. *Nat. Commun.* **9**, 319 (2018).
2. Jolley, K. A., Bray, J. E. & Maiden, M. C. J. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res.* **3**, 124 (2018).