## Procedure and Methods

#### Pangenome construction

Pangenome was constructed using Roary pipeline with default parameters (95% aa identity). Another pangenome was also constructed at 80% to reduce redundancy even further for intergenic region analysis using Piggy pipeline. Identification of intergenic regions was performed using default parameters in the Piggy pipeline.

#### Analyzed genomes and their quality

1245 initial genomes were selected. Most of the genomes were derived from Trento study (~1161) (not published yet). Genomes were tested for the quality with CheckM (contig files as input). It was chosen 1178 Wolbachia assembly genomes of good quality (out of 1245) (Completeness>=70; Contamination<=10) plus 12 additional genomes, and 2 outgroups (*Anaplasma marginale, Ehrlichia ruminantium*, - due to their wide application in the Wolbachia phylogeny studies(keep in mind **LBA** (long problem which is still unsolved for Wolbachia stains) (in the folder “01.CheckM\_results”)

#### Additional genomes and outgroups

Additional genomes and outgroups were annotated separately from the rest of the dataset (Sasha’s annotation) using Prokka annotation on Galaxy service. Protein sequences were blasted against the pangenome to assign clusters for genes derived from additional genomes. faa and fna files for these genomes were obtained (“02.Additional\_genomes\_outgroups”).

#### Phylogeny

##### *Tree based on the MG40 genes*

Firstly, to build a tree, fetchMG was used. It selects 40 marker genes over 1178+12+2 genomes for further phylogeny. To build a concatenated sequence of 40 marker genes per genome, seqkit tool was used. After that alignment with ClustalOmega was performed followed by implementation of FastTree for phylogenetic tree (saved in Newick format) (“03.Phelogeny/FetchMG”)

##### *Tree based on the MLST genes*

MLST based tree approach has been widely used in literature for *Wolbachia* phylogeny, thus, this method was also applied. Protein MLST database of *gatB, coxA, hcpA, ftsZ, fbpA* genes along with *wsp* was created. Protein BLAST of all sequences of genomes was performed against the MLST database (evalue<1e-6). For this tree, all genomes were used - 1244 (except one, which was too short)+12+2. After concatenation alignment with ClustalO, tree - FastTree. (“03.Phelogeny/MLST”)

Trees were visualized in the EvolView(<https://www.evolgenius.info/>) where the metadata also can be uploaded (iTOL requires to pay for that) (“03.Phelogeny/Meta-data/EvolView”)

Links for EvolView tree for interactive tree visualisation:

Fmg40 tree - <https://www.evolgenius.info/evolview/#shared/%20fmg40_oeyKorGRYi>

Mlst tree - <https://www.evolgenius.info/evolview/#shared/mlst1_mckHRhAZqi/>

#### Gene Annotation due to functions

##### *Description of pangenome clusters*

Main annotation of the pangenome was performed with EggNog (39977 clusters successfully submitted (187 were missing); 27057 clusters annotated). However, there were still a significant number of genes with unknown designation (12921 were needed to annotate). (“04.Gene\_annotation/EggNog”)

###### iPath

iPath was used for visualization of metabolic differences between *Wolbachia* species based on the EggNog annotation. Nonetheless, there were a lot of missing links in paths as incomplete annotation was obtained (files for iPath were obtained after parsing EggNog output: “04.Gene\_annotation/iPath”)

iPath diagrams for some strains:

wPep - <https://pathways.embl.de/selection/yY49yOdxts7tPfNLieR>

wVulC -<https://pathways.embl.de/selection/wyNX5a2nR5Hqqd7Bm9H>

wMel - <https://pathways.embl.de/selection/5N4Ob36y2dhnd2V7gQF>

wMelPop - <https://pathways.embl.de/selection/I4Q0H2fsJjA5Bjejnsd>

wAlbB - <https://pathways.embl.de/selection/J5MGGzUOqZyVRIanZvX>

wNo - <https://pathways.embl.de/selection/EhH8qmqCfxoYbFWG6MN>

wHa - <https://pathways.embl.de/selection/x26651BrD8Zt5Upe53H>

wYak - <https://pathways.embl.de/selection/RQkB9ZSu76VNydPvRTr>

Other resources for annotation that were used and included in analysis: Pnnzr2, BxOm, GoFeat, veupathdb (Galaxy) etc. (“04.Gene\_annotation/Annotation of pangenome.tar.gz”)

The final table of annotation includes data obtained from EggNog, Pnnzr2, BxOm, GoFeat (40165 clusters) (“04.Gene\_annotation/annotated\_pangenome”)

##### 

##### *Mobile elements*

To annotate mobile elements, it was attempted to use IceFinder but there was a problem with instalation. Then the sequence database (MobilomeDB) used for IceFinder was downloaded (<https://db-mml.sjtu.edu.cn/STEP/download.html>) ((not all of them: except *24 Type VI secretion systems representatives; 92 Type VI secretion effectors; class I integron representatives -* they were absent*,* and *322 HMM profiles for core component proteins of prophages* was not fasta format*)*) and pBLAT was carried out with evalue<=1e-6 (queries - all fasta protein for all genomes). Afterward, the trimming of obtained data had continued; the Ha-value was evaluated manually using convenience of the command line (more about Ha-value: <https://db-mml.sjtu.edu.cn/ICEfinder/instruction.html>)

##### *Phages*

Phages were obtained with PhastaF from all contigs over analyzed genomes.

#### Strain identification

Some of the genomes had identified strains from a source database. As a matter of fact, there were a lot of unknown strains mainly because of the Trento dataset. Thus it was attempted to identify strains, at least in *sensu lato* way. For this purpose, nBLAST was used with a dataset of key genes set for known strains (*wsp* and MLST genes) against wsp and MLST genes from all genomes (evalue 1e-6), including known genomes - to verify thoroughness of the approach. 691 genomes out of the analyzed dataset have been identified in terms of the strain designation.

## Analysis 1: Differences between phenotypes due to genomic composition

*Wolbachia* studies often indicate that there is considerable number of evidence that endosymbiont genetical features are not mainly responsible for the strain manifestation as there is, in particular, a strong dependence on the host genetical environment which has been recently assumed as the main reason for different phenotypes evolvement under Wolbachia endosymbiosis. To elaborate on this, it is known that under change of the host different strains can have distinct phenotypes compared to the original host. Also, there are facts that after evolutionary recent transinfection strain has mild or no malign effect on the new host, especially it is true for Drosophila strains. It is worth of noticing that it is true mainly for strains which are derived from A, B clades - parasatic/obligatory endosymbints whcih have bigger genomes and still can have deeper flexibility towards interactions with host genome as their genes were not influenced to Muller ratchet as such in the mutualists.

This incongruence may be explained from the sight of the lack of literature regarding broader spectrum of strains from other clades, restrictedness of gene modification of intracellular bacteria as well as skewness of the definite effects of mutualists under expression of particular favorable for host genes, e.g. biotin operon.

Nonetheless, such effects as nutrition supply, viral defense and different strategies of reproductive manipulation were shown linked to some known genes, based on the experimental proof or predicted from the presence of genetic information within the genome.

The idea was to compare genomic composition of different strains grouped with criteria of both experimentally proved and predicted effects on the host for the sake of either symbiotic or parasitic relationships.

*Phenotypes*

The dataset for which the strains were known, the influence of genes of interest, genomic characteristics, phages, and ME on the phenotype manifestation (CI, FA - fitness advantages, M - mutualism, F - feminisation, MK - male killing, P - parthenogenesis). Overall data size comprised 691 genomes over A, B, C, D, E, and F supergroups (fig.1). The most abundant genomes were those from the supergroup A (623 genomes), mainly included into the FA. F and MK were both present by single genotype for each phenotype (wVulC (*Armadillidium vulgare*) and wBol1-b (*Hypolimnas bolina*), respectively) from B clade. Overall A clade strains are observed among CI, FA, and P phenotypes, when B - CI, F, MK, P. That indicates a wider spectrum of reproductive parasitism for B clade members. C, D, and F clades were only in the M, mutualistic phenotype where hosts are nematodes (in particular, filarial parasites), and bed bug, *Cimex lectularius* (F). From E clade, wFol *Folsomia candida*, conceivably, the parthenogenesis is the dominant way of reproduction in this springtail exactly due to *Wolbachia*,is present along with A and B groups in the P phenotype. Maybe, there is a need to look into the genes of wFol to find possible “parthenogenetic” genes.

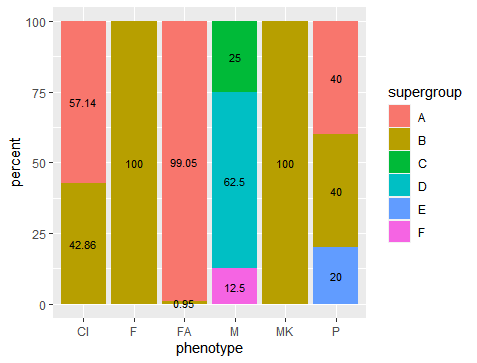
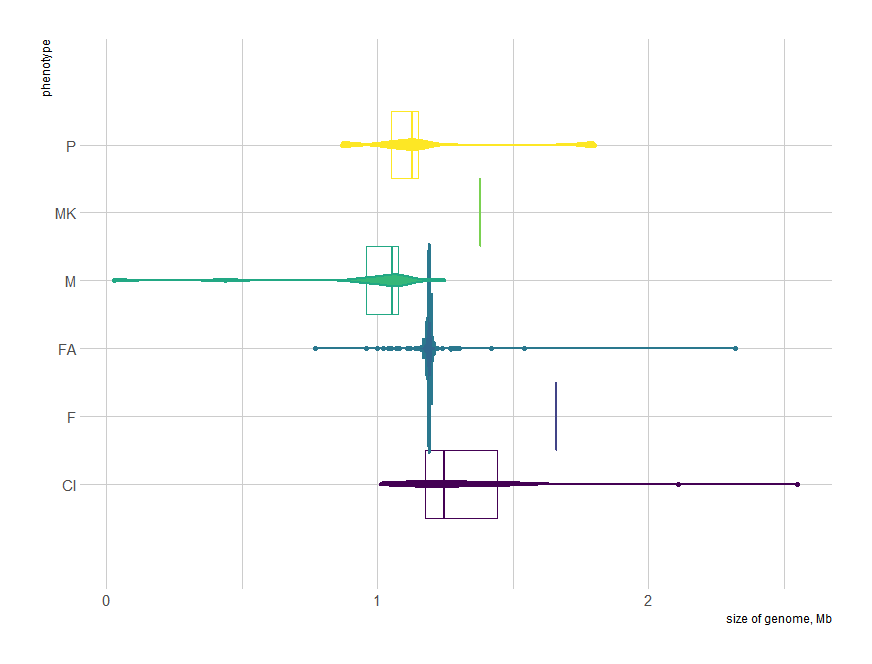


Figure 1. The percent of the supergroups within each phenotype for analyzed *Wolbachia* genomes

The phenotypes for particular strains were identified based on papers: Werren et al., 2008; Hague et al., 2020; Miller and Riegler, 2006; Desjardins et al., 2013; Beckmann and Fallon, 2013; Lindsey et al., 2016; Liu et al., 2019; Kaur et al., 2017; Metcalf et al., 2014; Zhang et al., 2010; International Glossina Genome Initiative, 2014; Graber and Fallon, 2019; Badawi et al., 2015; Çiğdem et al., 2020)

Genomics characteristics indicate that in common the genome sizes are smaller in the strains of M, and bigger for CI strains (fig.2, A). The average size of P genotypes take place between M and CI. Due to abundance of wMel genomes for FA, the distribution is skewed towards wMel strain size (1.32 Mb in average). MK and F strains are presented as solitary, thus they might not be true representatives for these phenotypes. Similar tendency is observed for the number of CDSs (fig.2, B)

B

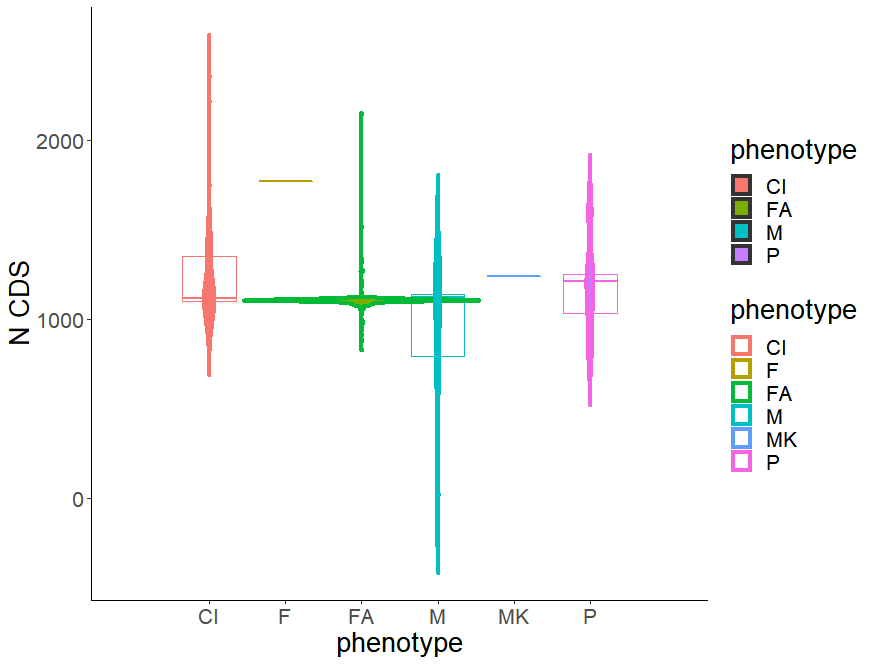
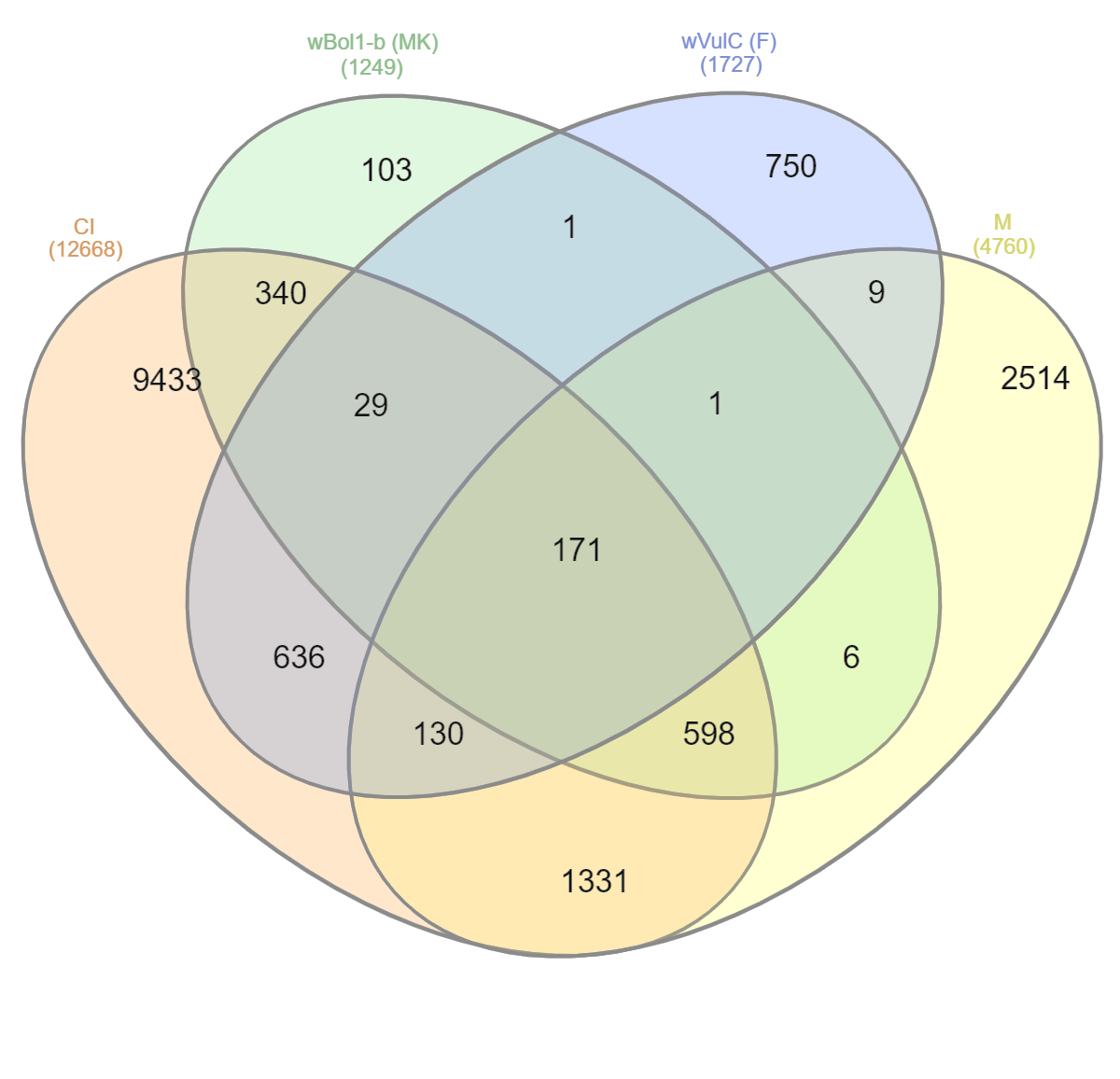
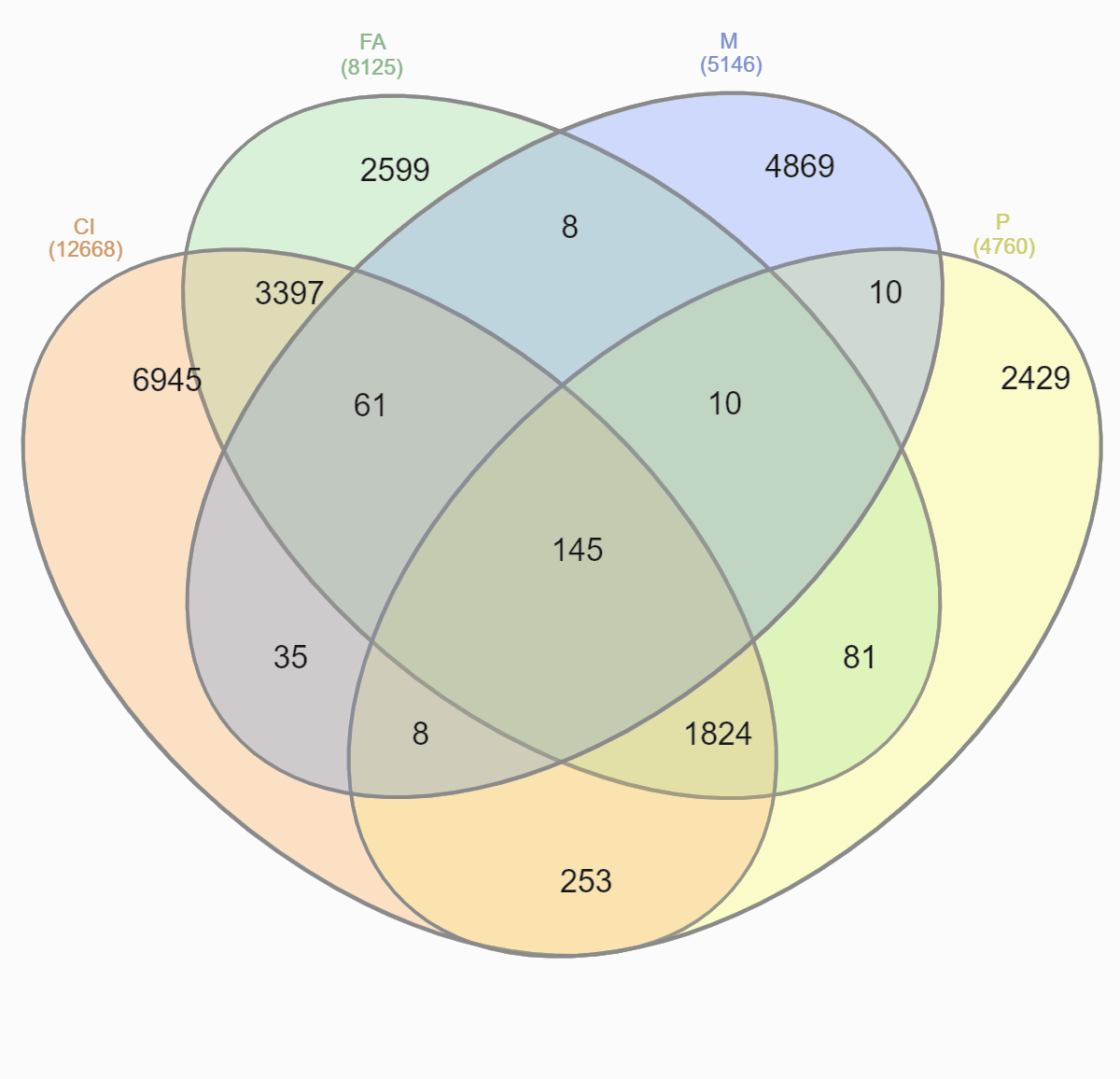


Figure 2. Violin plots for genomic characteristics of the phenotypes for analyzed *Wolbachia* genomes. A - size of the genome; B - number of CDSs.

The closest phenotypes due to genes content are CI and FA (which mainly consist of strains of the supergroup A) (fig.3 A). The significant number of genes P share with both CI and FA. The most distinct phenotype is M group, it has the biggest number of unique genes. F (wVulC) and MK (wBol1-b) were the least conserved groups and have a big share with CI and M groups; may be explained underrepresentation for these phenotypes.



A B

Figure 3. Venn diagrams of genes content for analyzed *Wolbachia* genomes. A - CI, FA, M, P phenotypes; B -CI, MK, F, M phenotypes.

It would be interesting to compare determined phenotypes with the genomes of the strains that share similar characteristics (wPpe has reproductive and mutualistic charact) and have experimentally not confirmed effects like in Nomada strains which can have mutualistic relationships (fig.).

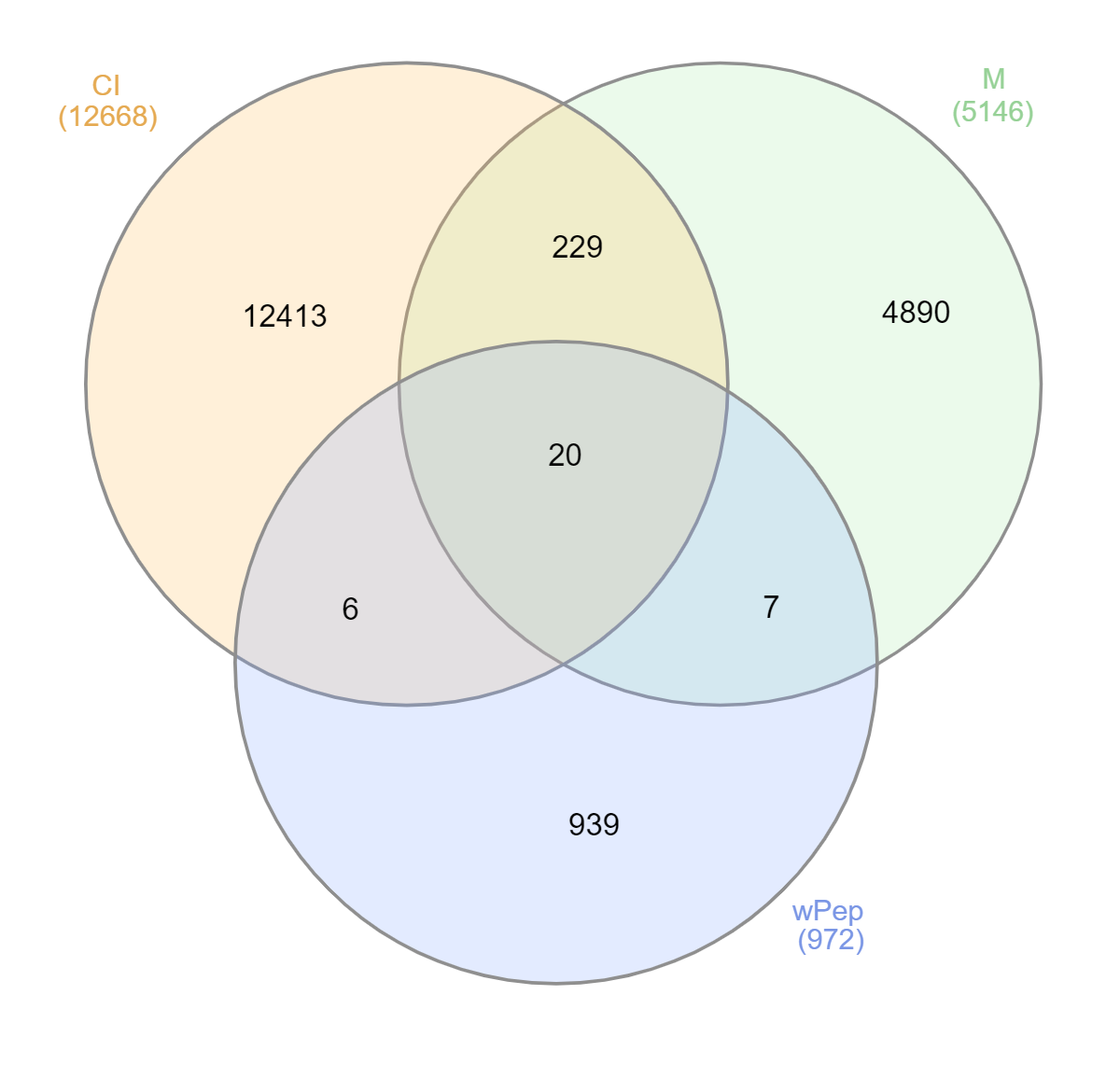


Figure.

Also, it is an unknown fitness effect of wInc, strain that hitchhiked to the Drosophila incompta in the New world, on the host but it doesn't cause CI (Wallau et al., 2016).

Parthenogenesis of springtail Folsomia candida is probably an effect of wFol (Graber and Fallon, 2019; Faddeeva-Vakhrusheva et al., 2017) but it still needs to be proved (we took it as parth.casual). Also, genomically it was supported that wFex of Formica exsecta is CI-inducing (Dhaygude et al., 2019) but we didn’t include it. wDacA and wDacB of Dactylopius coccus can be mutualists (moreover, wDacA doesn’t have candidate genes that are likely involved in cytoplasmic incompatibility as in wDacB), nonetheless, there is no experimental evidence, so they were not included to the phenotypes denomination (Ramírez-Puebla et la., 2016). In the same way it was shown that wCfeT is mutualist due to carrying biotin synthesis operon, and wCfeJ is reproductive parasite, it contains operon similar to the CinA/B TA operon of wPip\_Pel, inCtenocephalides felis (Driscoll et al., 2020), then they weren’t included.

*Genes of interest*

Based on the findings regarding the genes that are responsible or conceivably responsible for different influences on the host, 8 gene groups were chosen (fig.). These genes were sorted out from the clusters dataset based on the presence of key word within the description of the cluster. Accordingly different number of appropriate genes were found for analysed phenotypes. On the fig. , the average coun of genes of interest is presented.

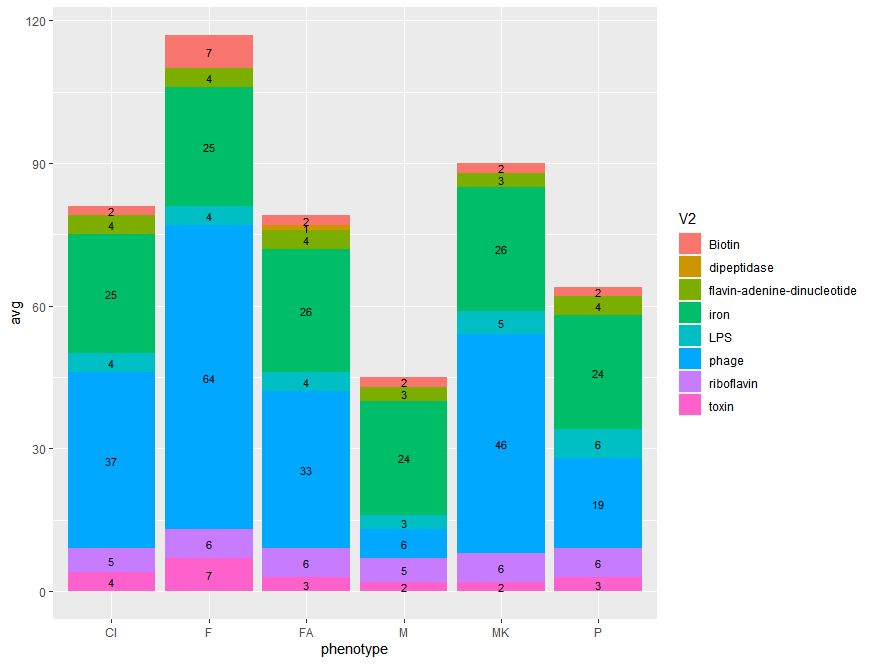


Figure.

Dunn test was used to compare phenotypes due to the genes of interest. M phenotype differs from CI, FA, F (p<0.05).

## Analysis 2: Mobile elements and phages

Overall, the 112 genomes were identified as carriers of the phages. PhastF had shown that the most rich genomes with phages were wAlbA\_HN (12 incidences in the genomes out of 8 types of phage elements; however, wAlbA\_FL strain had only 2); WOLB0035, WOLB0042 - 7 instances; WOLB0040 -6; For < 25 genomes, the presence of phages were more than 2 instances.

Xantho\_33913 (100), Escher\_Sakaiphage (39) were observed the most often when Caulob\_Cr30, Escher\_pro147, Entero\_186, Pseudo\_phi3, Pseudo\_PPpW\_3, Salmon\_SP\_004 were happened as single instances.

summary for **mobilome** table

KHFPKGOH\_01202: 30 ISWpi12\_PEP : 2832

KHFPKGOH\_01213: 26 ISWosp2\_PEP : 2134

EIACEGFN\_00180: 14 ISWpi11\_PEP3: 1986

LMLIPNKF\_00597: 7 ISWpi15\_PEP : 1612

LHNECCLP\_00824: 6 ISWen2\_PEP : 1466

OODCHNIP\_00058: 6 ISWpi11\_PEP : 1425

(Other) :22864 (Other) :11498

**KHFPKGOH\_01202** (the most frequent one) - only one gene in cluster turf1. Translation elongation factor activity (Pnnzr2\_MF), Elongation factor Tu(Prokka\_descrp) (GCA\_902636535.1\_WOLB0025\_genomic)

mob.elem: NC\_002516.2.881718.p01, NC\_002516.2.881697.p01, CP002695.1.gene10.p01, 02 CP002695.1.gene3614.p01, CP004022.1.gene3277.p01, CP004022.1.gene2801.p01, CP001918.1.gene242.p01, CP000647.1.gene4394.p01, CP000647.1.gene3761.p01 ( some have double appearance, within >80, <90 identity)

## To do next?

##### From notes:

Nucleotide divergence and Synteny

"Overall, genomic distances based on nucleotide divergence of orthologues shared between the newly sequenced Wolbachia strains was found to be low (1.2% on average; Fig. 1a), and synteny across genomes high (Fig. 1c). " https://pubmed.ncbi.nlm.nih.gov/28005061/

How to measure divergence levels?

". Nucleotide distances between orthologues present in all genomes were calculated with the R package ‘ape’ (http://ape-package.ird.fr/) 42" https://pubmed.ncbi.nlm.nih.gov/28005061/

synonymous substitution in the certain gene

as the characteristic to compare of this estimate between different genomes (for example, for bioitin operon among different genomes)

"Furthermore, the estimated rate of synonymous substitutions between Nomada-associated Wolbachia strains and wCle was, on average, much lower in the biotin genes (0.17) compared with genome-wide rates (0.41). This also suggested that the operon was not present in the last common ancestor (LCA) of wCle and the strains investigated here, but rather that it was acquired by Nomada-associated Wolbachia more recently."

https://pubmed.ncbi.nlm.nih.gov/28005061/

to evidence recombination between hypothetical stains genes that they share only (build for them trees with synteny)

"B-vitamin-deficient diets (such as blood)"

"Single-gene trees were reconstructed for all orthologous groups with IQ-TREE version 1.3.853 after determining the best-fitting nucleotide substitution model (option ‘-m TEST’)." https://pubmed.ncbi.nlm.nih.gov/28005061/

for visualisation of an alignment

". Using the ‘ape’ package within R, trees with monophyletic bee and bedbug Wolbachia strains were identified, and corresponding alignments were further manually scrutinized using Aliview version 1.17.1" https://pubmed.ncbi.nlm.nih.gov/28005061/

For future: use MLST to built the tree and recognize the strain

for strain identification https://pubmlst.org/wolbachia/

"Strains with similar WSP sequences can have very different MLST allelic profiles and vice versa, indicating the importance of the MLST approach for strain identification. The MLST system provides a universal and unambiguous tool for strain typing, population genetics, and molecular evolutionary studies. "

https://pubmed.ncbi.nlm.nih.gov/16936055/

[05/28/202]

Blast pWCP against all the genomes

pWCP plasmid was identified in wPip strain https://www.nature.com/articles/s41467-019-08973-w.pdf

here are other genes:

https://www.nature.com/articles/srep34955