## Historic museum samples provide evidence for a recent replacment of *Wolbachia* variants in European *Drosophila melanogaster*.

Anton Strunov1, Sandra Kirchner2, Julia Schindelar2, Luise Kruckenhauser2, Elisabeth Haring2, Martin Kapun1,2,\*

1 Medical university of Vienna, Center for Anatomy and Cell Biology

2 Natural History Museum, Central Research Laboratories

\* Corresponding Author: Martin Kapun [martin.kapun@meduniwien.ac.at](mailto:martin.kapun@meduniwien.ac.at)

### Abstract

### 1. Introduction

*Wolbachia* is a gram-negative alpha-proteobacterium of the order Rickettsiales, which represents one of the most common endosymbionts in animals. *Wolbachia*, which have been detected in XXX percent of all arthropds, can have a substantial impact on the life history and fitness of its host. These effects range from parasitic to mutualistic, since *Wolbachia* often manipulates the host reproduction for it’s own benefit but

### 2. Materials and Methods

In this study, we investigated *Wolbachia* infections in 25 historic *D. melanogaster* museum samples from Sweden, Denmark and Germany, that were collected between 90 and 200 years ago (REF). Taking advantage of recently published whole genomic Illumina deep sequencing data of the samples, we tested for the presence of *Wolbachia*-specific reads in these samples, estimated titer variation and investigated the relatedness to contemporary *Wolbachia* strains. Complementary to the historic samples, we used Oxford Nanopore sequencing technology (ONT) to newly sequence genomic DNA of six strain of freshly collected isofemale lines from wild populations in Portugal and Finland, that were naturally infected with either the wMel or the wMelCS *Wolbachia* variants (REF) and of three lab-strains that were artifically infected previously with wMel, wMelCS and wMelPOP (REF; see Table S1). Complementary to these data, we obtained XXX RefSeq assemblies of *D. melanogaster*-specific *Wolbachia* samples and included raw Illumina sequencing data of XXX *Drosophila melanogaster* samples infected with wMel or wMelCS from the NCBI Short Read Archive (SRA) for phylogenetic analyses (see Table S1). Given the close relationship between wYak and wMel, we obtained the RefSeq sequence information of the wYak *Wolbachia* variant (NZ\_VCEF01000001.1) and the mitochondrion () of *D. yakuba* and used these as outgroups for phylogenetic inference.

#### 2.1 DNA extraction, library preprartion and ONT whole genome sequencing

TBA

#### 2.2 Detection and characterization of *Wolbachia* infections

To obtain estimates of *Wolbachia* titers, we used bwa mem (REF) for Illumina sequencing data or minimap2 (REF) for ONT sequencing data to map all raw FASTQ reads for each sample against a joint reference sequence, which was constructed from the *Drosophila melanogaster* reference genome v.6 (REF) and additional genome sequences of other common microbial symbionts, including the wMel reference genome (see REF for more details). Using the command *samtools depth* of the samtools program (REF) in combination with a custom *python* script (SumReadDepths.py), we calculated average absolute read depths for all *Drosophila* chromosomes and symbiont genomes. Based on this information, we estimated relative *Wolbachia* titers for a given sample by dividing the average read depth at the *Wolbachia* genome by the average absolute read depth across all *Drosophila* autosomes. Based on the average read depth and the proportion of the *Wolbachia* genome covered, we classified *Drosophila* samples as either infected with *Wolbachia*, as uninfected, or as with unclear status.

#### 2.3 De-Novo assembly and draft annotation

In a next step, we pre-filtered the raw FASTQ files of each library with Kraken (REF) using a custom-built databset that consisted of the published genomes of wMel (XXX), wMelCS (XXX) and wMelCS POP (XXX) and only retained Illumina and ONT reads which matched the references in the database. Illumina reads of the historic samples were quality-trimmed (PHRED-scaled basequality >=25) and trimmed for sequencing adapters using cutadapt (REF). We only retained intact read-pairs with a minimum length of 75bp for de-novo assembly with SPAdes (REF) using default parameters. Raw long-fragment reads from ONT sequencing of contemporary flies were assembled with Flye (REF) using default paramters.

Subsequently, we assessed the assembly quality based on common quality statistics such as numbers of contigs, N50 and N90 with QUAST (REF) and tested for assembly completeness using the BUSCO approach, where the proportion of intact, fragmented and missing benchmarking universal single-copy orthologs specific to the bacterial order Rickettsiales (rickettsiales\_odb10) was evaluated in each assembled genome (REF). In addition, we re-mapped the raw reads to the assembled contigs using minimap2 (REF) to assess variation in read-depth and compared all contigs to a local copy of the NCBI *nt* database using *blastn* of the BLAST suite (REF). After that, we visulized the results of these quality assessments with Blobtools (REF).

Finally, we used the published *Wolbachia* wMel reference genome (ENA|AE017196|AE017196.1) as a backbone to align and orient the raw contigs with *nucmer* of the MUMmer package (REF). Based on *show-tiling* of the MUMmer package, we identified the minimum number of unique contigs that span a maximum of the reference backbone. Using a custom python script (), we then combined these contigs into a single scaffold and filled the gaps between each pair of consecutive contigs with a string of ten N’s. Moreover, given that the bacterial genome is circular, we anchored the newly assembled scaffolds at the startpoint of the reference genome and shifted pretailing sequences to the end of the scaffolds. We then calculated multiple genome alignments of all scaffolds and the *Wolbachia* reference using progressive-mauve (REF) and visualized the alignemnt with Mauve GUI (REF).

Since RNASeq data was neither available for the historic nor for the contemporary samples, we computed a draft genome annotation based on comparing the genomic sequences of the assemblies to a reference wMel (ENA|AE017196|AE017196.1) transcriptome by gapped alignment using the *protein2genome* model of exonerate (REF). Conservatively, we only retained gene models with reached at least 80% of the maximum alignment score optainable for a given sequence.

#### 2.4 Phylogenetic analysis

We employed two complementary approaches to explore the evolutionary history of *Wolbachia* based on phylogenetic inference.

##### 2.4.1 Phylogenetic analysis based on candidate genes

In a first approach, we compared the nucleotide sequences of gene models obtained with the BUSCO approach from the denovo assembled genomes of the museum and the contemporary samples. In addition, we supplemented our dataset with XXX published genomes assemblies of *Wolbachia* samples from *D. melanogaster* hosts available at the NCBI RefSeq database (XXX). We included these independently assembled genomes to confirm that phylogenetic signals are neither confounded or biased by our assembly pipeline nor by combining Illumina and Oxford Nanopore sequencing data.

To obtain a core set of orthologous genes, we applied the BUSCO approach as explained above to all assembled genomes and focused on 104 genes, which were identified as complete and which were present in the majority of the assembled genomes in our dataset. Using MAFFT (REF), we aligned their nuclear sequences across all samples and concatenated the alignments with a custom Python script (ConcatenateAlignments.py). Then, we reconstructed a maximum likelihood tree based on the GTR-Gamma substitution model from 20 starting trees using RaXML (REF) and additionally performed 100 rounds of bootstrapping to test for the robustness of each node. The final tree was plotted in *R* (REF) using the *ggtree* package (REF).

##### 2.4.2 SNP-based phylogenetic analysis

Several of the draft *Wolbachia* genomes assembled from raw Illumina reads of historic samples, were charcaterized by very low numbers of complete BUSCO genes (<5 genes). Thus, it was not possible to include these samples in the phylogenetic approach based on candidate gene alignment explained above. We therefore employed a complementary approach based on reference mapping to include more of the historic samples into the phylogenetic analsyis. To this end, we mapped the raw Illumina reads of each sample that we pre-filtered for *Wolbachia* with Kraken, as explained above, against the wMel reference genome (AE017196.1). For the Illumina sequencing data of historic samples and contemporary samples downloaded from NCBI SRA, we mapped paired-end reads using bwa mem (REF) with default settings. Conversely, we used minimap with default paramters to map long-fragment reads from ONT sequencing against the reference *Wolbachia* genome. Raw BAM files were filtered with Samtools (XXX) to contain mapped reads only and sorted by reference position using the *samtools sort* command.

Then, we used the BCFtools (XXX) command *bcftools mpileup* to synchronize the mapped reads of all samples and called SNPs using *bcftools call* assuming haploidy and stored the variants in the VCF file format. Using a custom Python script (BCF2Phylip.py) we converted the VCF file to the phylip format, only considering bi-allelic polymorphic positions where the posterior probability of the most likely genotype was > 50 and of the alternative genotype <30. In addition, we only included positions with more than 5-fold coverage in each of the samples and removed samples, with more than 50% missing information across all SNPs. Similar to above, we used the aligned SNP data to reconstruct a maximum likelihood tree based on the GTR-CAT substitution model from 20 starting trees using RaXML (REF) and additionally performed 100 rounds of bootstrapping to test for the robustness of each node. The final tree was plotted in *R* (REF) using the *ggtree* package (REF).

##### 2.4.3 Classification based on diagnostic SNPs

The two aforementioned phylogenetic approaches allowed to determine the evolutionary relationship of historic *Wolbachia* samples to contemporary *Wolbachia* strains for a handful of the museum samples. However, several other samples which appeared to be infected with *Wolbachia* did not have sufficient coverage to be included in these two analyses. We thus choose a different strategy to estimate if these are more closely related to wMel or wMelCS. We therefore identified 115 SNPs in 15 contemporary samples that were fixed for different alleles in the eight wMel and the seven wMelCS samples. Then, we tested the allelic state at these diagnostic SNPs in the 13 historic samples that were putatively infected with *Wolbachia* whenever the read-depth was >=2 and the allelic state was unambigously identified based on the posterior probablity as explained above.

#### 2.5 Comparsion to mitochondrial phylogeny

*Wolbachia* and the host mitochondria are both transmitted maternally to the offspring and should thus share a similar evolutionary history. This allows to test for genomic signals of horizontal introgression of *Wolbachia* into a host, which would manifest in inconsistencies among the species trees of mitchondria and *Wolbachia*. We therefore employed the SNP-based phylogenetic approach described above based on mitchondrial reads, which we pre-filtered by comparing all raw reads against a custom-built Kraken database consisting of the *D. melanogaster* mitochondrial genome. To this end, we first included all historic samples irrespective of the *Wolbachia* infection status. Subsequently, we restricted the analyses to samples only, which were *Wolbachia* infected and for which mitochondrial reads were available as well. We then repeated the phylogenetic analysis with RaXML as explained above and then used function *as.dendrogram* from the *R* package *phylogram* (REF) to convert the *Wolbachia* and mitochondrial tree files in NEWICK format to ultrametric trees. We first untangled the two trees, i.e., we swapped the branches to best fit the order of the samples using the *untangle* function with the “step1side” method of the *dendextend* package (XXX) and then produced a tanglegram using the *tanglegram* function to visualize the relationship between the two trees.

### 3. Results & Discussion

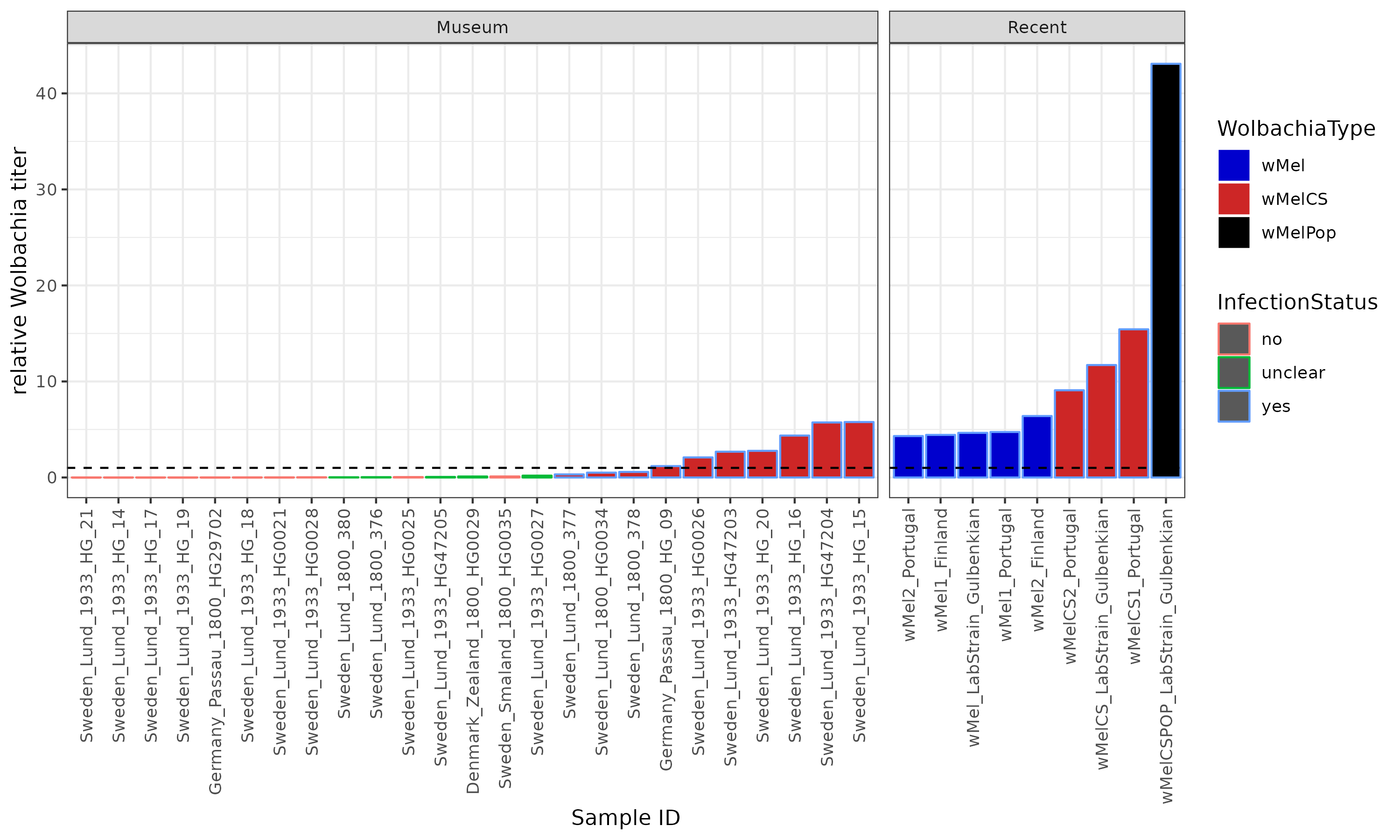
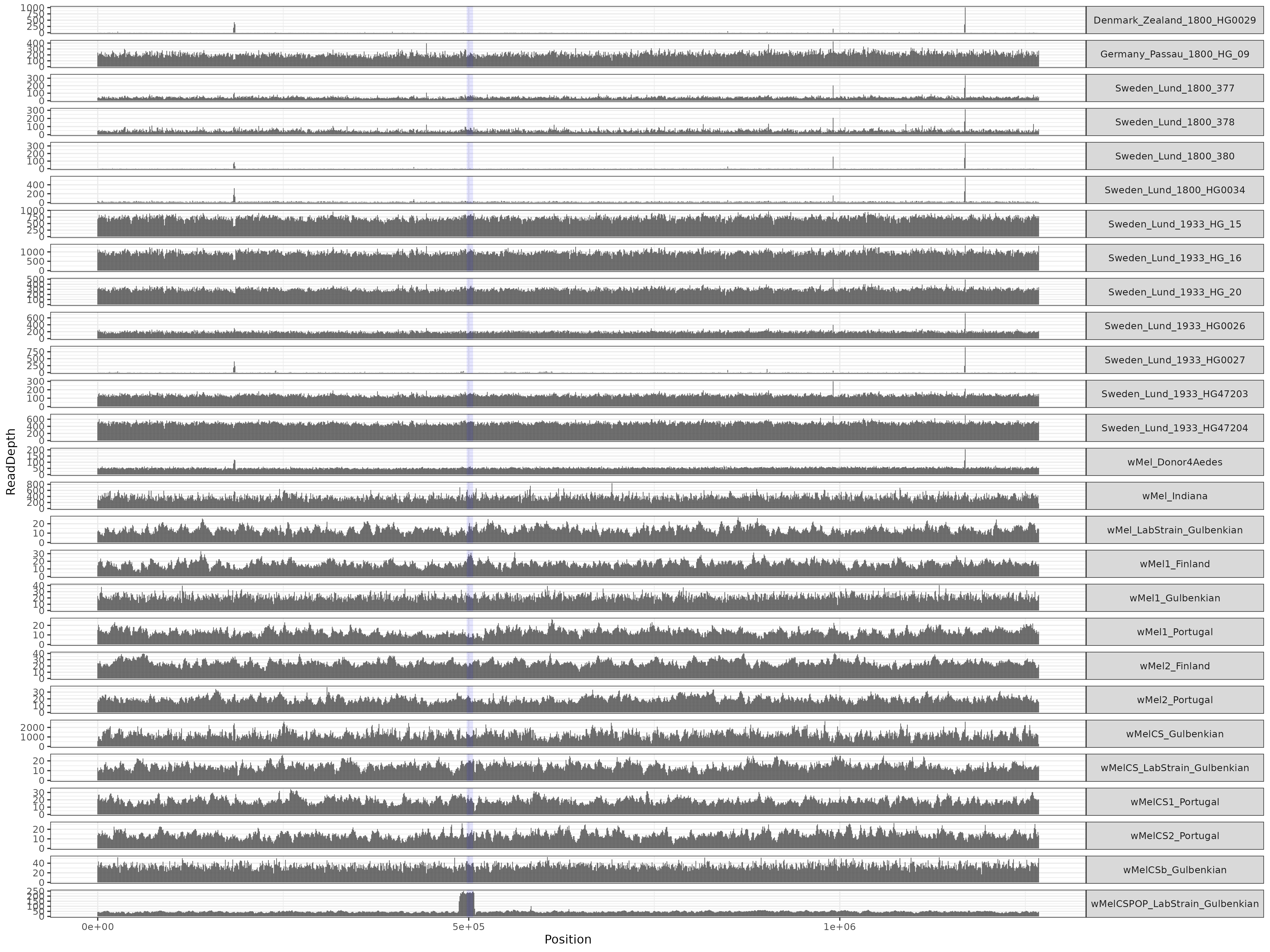
In this study we took advantage of a recently published genomic dataset of 25 museum samples of *D. melanogaster* samples that were collected betwen 220 and 90 years ago in Northern Europe. Besides testing if museomics of century-old *Drosophila* samples allows to identify historic *Wolbachia* infections, we address several long-standing questions concerning the co-evolution of *D. melanogaster* and *Wolbachia*. In particular, these data for the first time allow to test the hypothesis that the wMelCS *Wolbachia* variant, which is now found only at low frequencies in world-wide *D. melanogaster* populations was only recently replaced by the more common wMel variant within the last century.

#### 3.1 *Wolbachia* infections in historic *Drosophila* samples

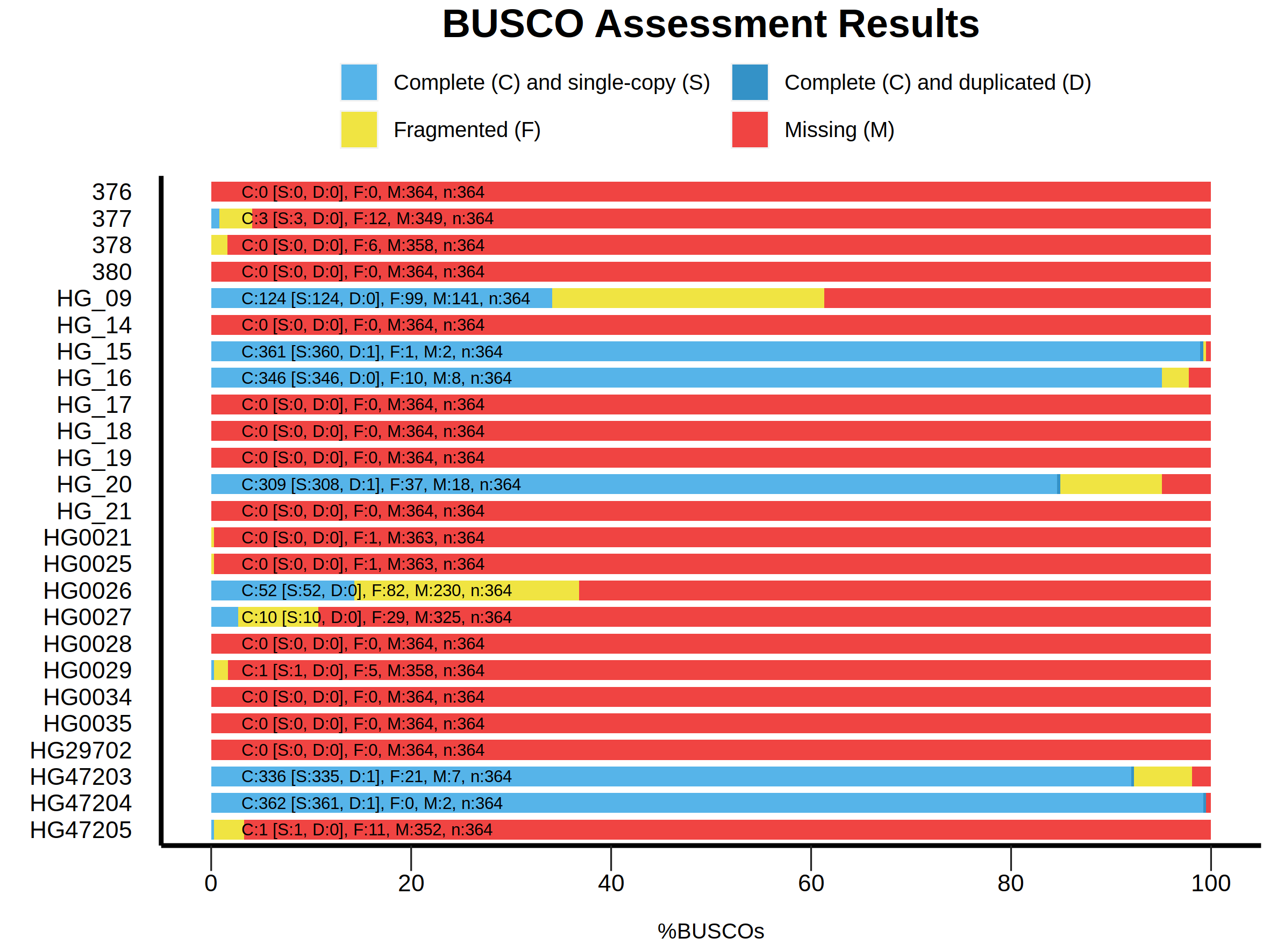
As a first step, we classified the sequenced *Drosophila* samples as infected or uninfected based on reference mapping. We found that read-depths of reads mapping to the *Wolbachia* reference varied dramatically between bascially zero-fold to a thousand-fold read-depths across samples (see Table 1). Similarly, we observed large variations in the proprortion of the reference sequence covered by reads range from as low as 0.9% to 100%. Based on these results, we qualitatively classified 10 historic samples as uninfected (<15% covered by reads and <2-fold average read depth) with very low coverage which could be due to non-specific mapping and 10 historic samples as infected (>50% covered by reads and >10-fold average read depth). Besides these two types,we identified five samples with uncertain infection status, which were characterized by read-depths > 2-fold which covered only parts (15-30%) of the reference genome. We found that these samples were further characterized by very low relative *Wolbachia* titer

**Table 1.** Table showing Sample ID’s, NCBI SRA accession numbers and sequencing technology used for the data generation of samples used in this study. Additionally, this table includes summary statistics for raw reads mapped against the wMel reference genome (ENA|AE017196|AE017196.1) and provides the proportion of the total reference sequence that is covered by reads, the mean read depth, the mean base-quality and the mean mapping-quality of each sample.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| NewID | OriginalID | NCBI Accession | Sequencing Technology | Reference Coverage [%] | Mean ReadDepth | Mean BaseQual | Mean MappingQual | Infection Status |
| Sweden\_Lund\_1800\_376 | 376 | SRR23876563 | Illumina | 30.6 | 4.26 | 36.3 | 50.7 | unclear |
| Sweden\_Lund\_1800\_377 | 377 | SRR23876584 | Illumina | 94.8 | 45.19 | 36.3 | 54.3 | yes |
| Sweden\_Lund\_1800\_378 | 378 | SRR23876583 | Illumina | 88.5 | 51.17 | 36.3 | 53.9 | yes |
| Sweden\_Lund\_1800\_380 | 380 | SRR23876582 | Illumina | 15.4 | 2.10 | 36.3 | 49 | unclear |
| Sweden\_Lund\_1933\_HG0021 | HG0021 | SRR23876586 | Illumina | 3.8 | 0.72 | 36.4 | 36.8 | no |
| Sweden\_Lund\_1933\_HG0025 | HG0025 | SRR23876585 | Illumina | 0.9 | 0.16 | 36.3 | 42 | no |
| Sweden\_Lund\_1933\_HG0026 | HG0026 | SRR23876567 | Illumina | 99.8 | 194.82 | 36.3 | 54.3 | yes |
| Sweden\_Lund\_1933\_HG0027 | HG0027 | SRR23876574 | Illumina | 33.3 | 5.43 | 35.9 | 55 | unclear |
| Sweden\_Lund\_1933\_HG0028 | HG0028 | SRR23876568 | Illumina | 5.2 | 1.92 | 36.3 | 42.6 | no |
| Denmark\_Zealand\_1800\_HG0029 | HG0029 | SRR23876565 | Illumina | 23.7 | 4.31 | 36.1 | 46.5 | unclear |
| Sweden\_Lund\_1800\_HG0034 | HG0034 | SRR23876564 | Illumina | 56.8 | 19.09 | 36.3 | 53.9 | yes |
| Sweden\_Smaland\_1800\_HG0035 | HG0035 | SRR23876566 | Illumina | 5.7 | 1.67 | 36.3 | 50.8 | no |
| Germany\_Passau\_1800\_HG\_09 | HG\_09 | SRR23876562 | Illumina | 99.9 | 217.00 | 36.3 | 54.7 | yes |
| Sweden\_Lund\_1933\_HG\_14 | HG\_14 | SRR23876581 | Illumina | 3.2 | 0.86 | 36.4 | 43.3 | no |
| Sweden\_Lund\_1933\_HG\_15 | HG\_15 | SRR23876580 | Illumina | 100.0 | 723.38 | 36.3 | 55 | yes |
| Sweden\_Lund\_1933\_HG\_16 | HG\_16 | SRR23876579 | Illumina | 100.0 | 976.46 | 36.2 | 54.6 | yes |
| Sweden\_Lund\_1933\_HG\_17 | HG\_17 | SRR23876578 | Illumina | 3.6 | 0.73 | 36.4 | 42.9 | no |
| Sweden\_Lund\_1933\_HG\_18 | HG\_18 | SRR23876577 | Illumina | 2.9 | 1.23 | 36.3 | 41.4 | no |
| Sweden\_Lund\_1933\_HG\_19 | HG\_19 | SRR23876576 | Illumina | 4.1 | 0.90 | 36.3 | 45.2 | no |
| Sweden\_Lund\_1933\_HG\_20 | HG\_20 | SRR23876575 | Illumina | 100.0 | 304.74 | 36.3 | 54.6 | yes |
| Sweden\_Lund\_1933\_HG\_21 | HG\_21 | SRR23876573 | Illumina | 4.0 | 0.40 | 36.4 | 35.4 | no |
| Germany\_Passau\_1800\_HG29702 | HG29702 | SRR23876569 | Illumina | 5.3 | 0.58 | 36.5 | 36.1 | no |
| Sweden\_Lund\_1933\_HG47203 | HG47203 | SRR23876570 | Illumina | 100.0 | 136.47 | 36.4 | 54.1 | yes |
| Sweden\_Lund\_1933\_HG47204 | HG47204 | SRR23876571 | Illumina | 100.0 | 496.44 | 36.3 | 54.9 | yes |
| Sweden\_Lund\_1933\_HG47205 | HG47205 | SRR23876572 | Illumina | 24.0 | 2.14 | 35.9 | 45.4 | unclear |
| wMelCS\_Gulbenkian | wMelCS | SRR945468 | Illumina | 100.0 | 1290.01 | 24.6 | 55.2 | yes |
| wMelCSb\_Gulbenkian | wMelCSb | SRR10438626 | Illumina | 100.0 | 34.80 | 37.3 | 56.8 | yes |
| wMel\_Donor4Aedes | wMel\_donor | SRR17978916 | Illumina | 100.0 | 53.74 | 35.9 | 56.7 | yes |
| wMel\_Indiana | wMel\_Indiana | SRR1645077 | Illumina | 100.0 | 388.55 | 35.5 | 56.6 | yes |
| wMel1\_Gulbenkian | wMel\_run1 | SRR10424182 | Illumina | 100.0 | 22.49 | 34.1 | 56.9 | yes |
| wMel1\_Portugal | Re1\_full | #NA | Oxford Nanopore | 100.0 | 13.58 | 24.3 | 58.6 | yes |
| wMelCS1\_Portugal | Re3 | #NA | Oxford Nanopore | 99.9 | 19.19 | 24.6 | 58.5 | yes |
| wMel2\_Portugal | Re6\_full | #NA | Oxford Nanopore | 100.0 | 20.14 | 24.3 | 58.3 | yes |
| wMelCS2\_Portugal | Re10 | #NA | Oxford Nanopore | 100.0 | 14.61 | 24.6 | 58.3 | yes |
| wMel1\_Finland | Ak7\_full | #NA | Oxford Nanopore | 100.0 | 17.22 | 24.3 | 58.5 | yes |
| wMel2\_Finland | Ak9\_full | #NA | Oxford Nanopore | 100.0 | 25.67 | 24.2 | 58.8 | yes |
| wMel\_LabStrain\_Gulbenkian | MEL\_full | #NA | Oxford Nanopore | 100.0 | 14.20 | 24.2 | 58.5 | yes |
| wMelCS\_LabStrain\_Gulbenkian | CS | #NA | Oxford Nanopore | 100.0 | 15.15 | 24.6 | 58.1 | yes |
| wMelCSPOP\_LabStrain\_Gulbenkian | POP | #NA | Oxford Nanopore | 100.0 | 49.36 | 24.7 | 58.2 | yes |

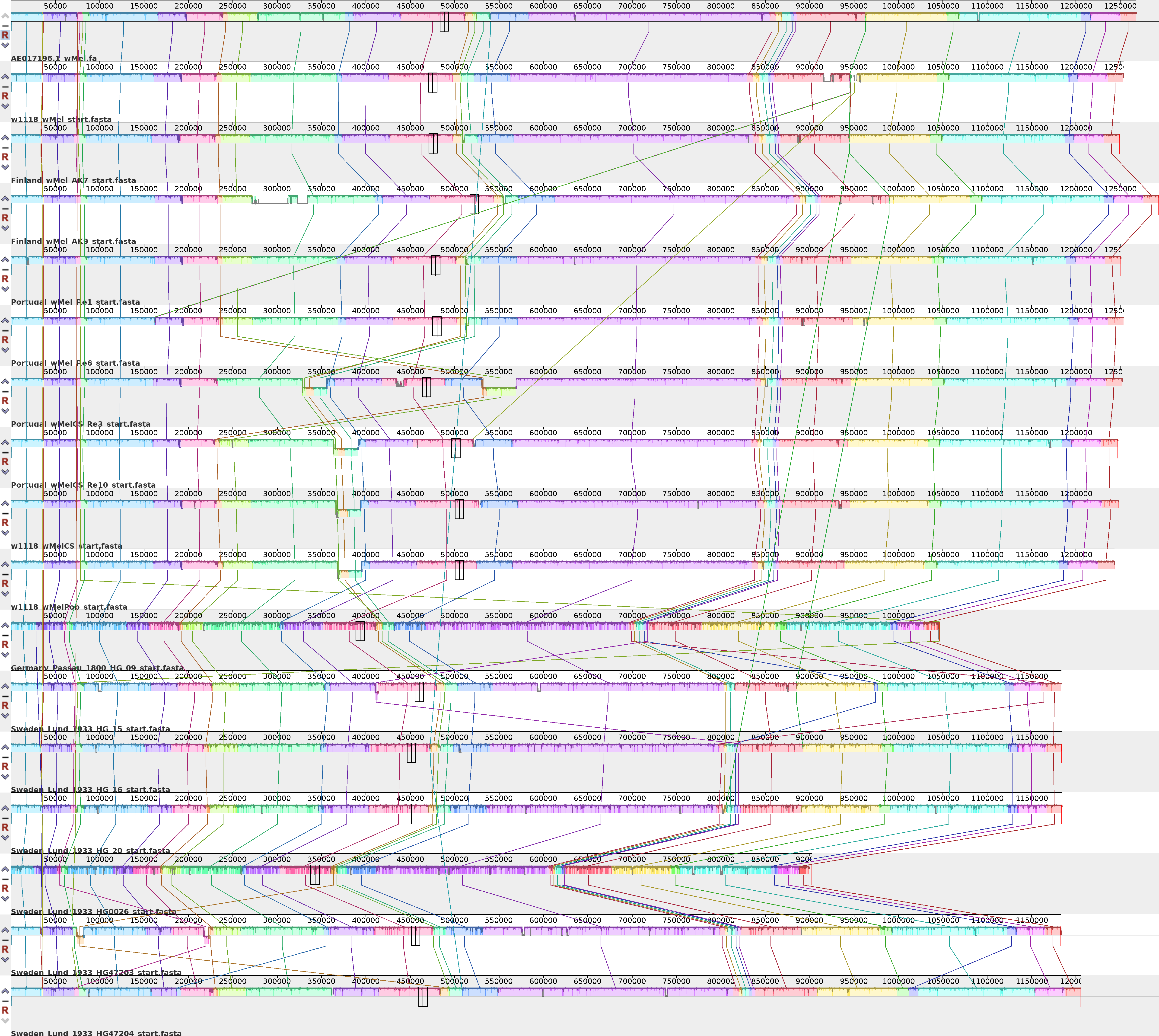
 

#### 3.2 DeNovo Assemblies



BUSCO

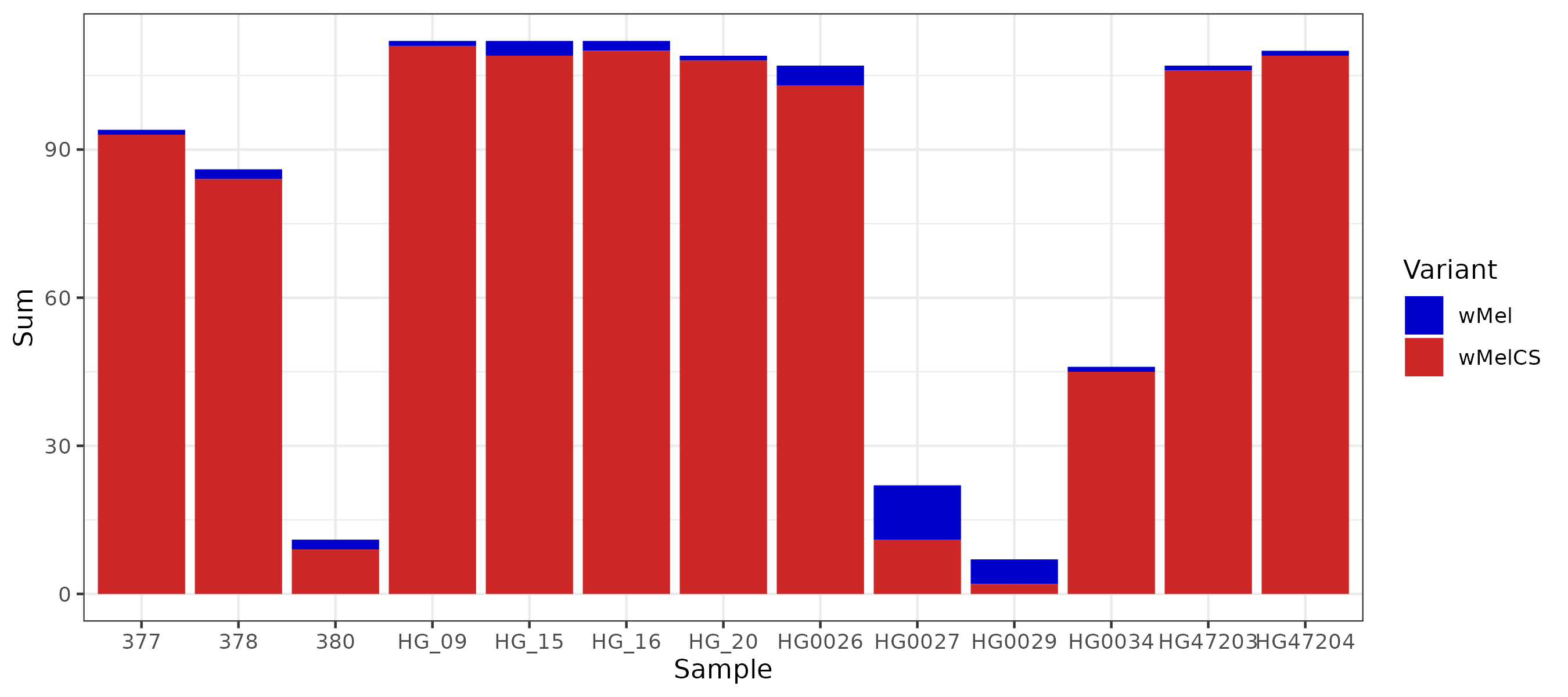
#### 3.3 Comparing genomes



Mauve

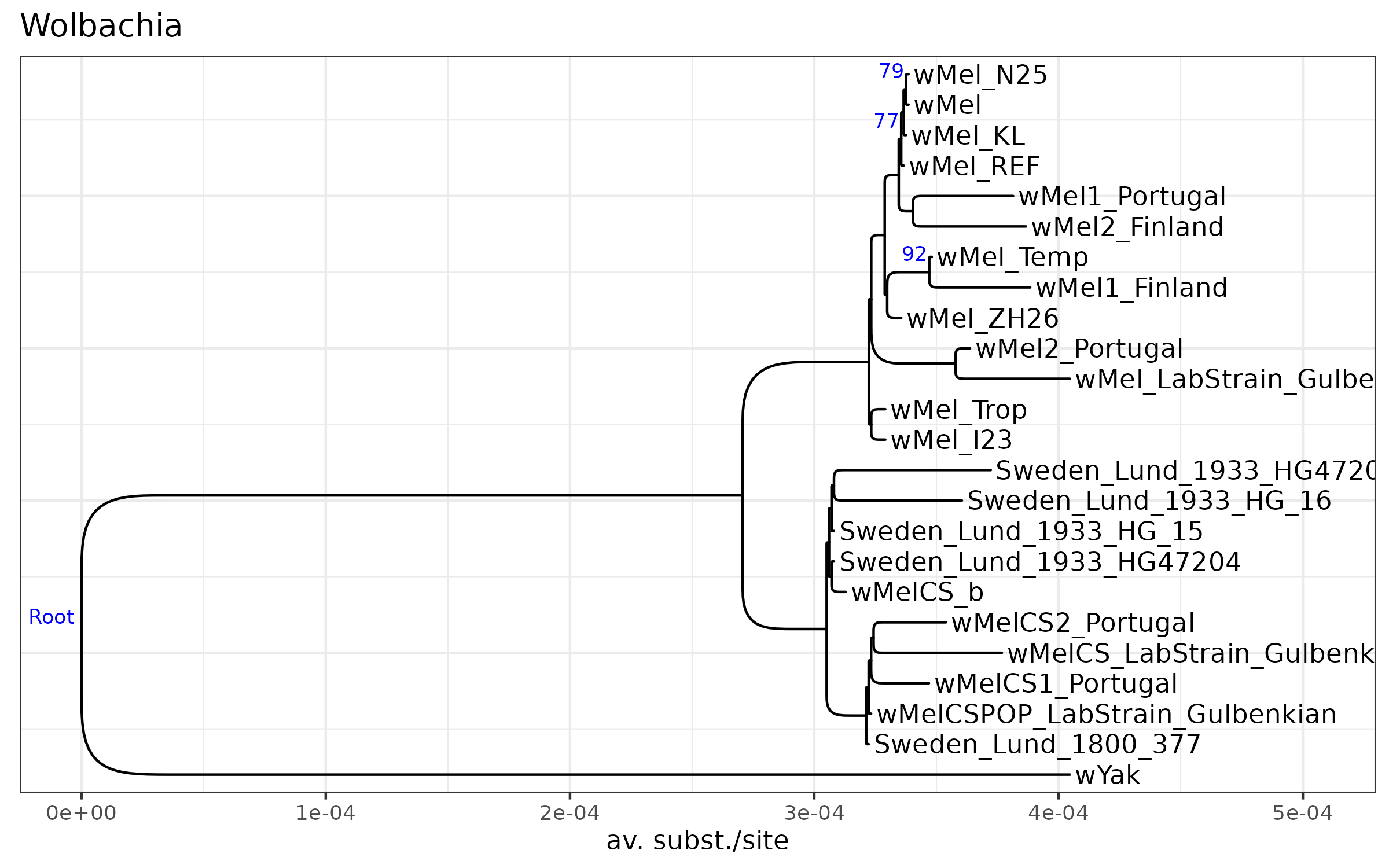
#### 3.3 *Wolbachia* phylogeny

##### 3.3.1 Classification with diagnostic SNPs



Diagnostic

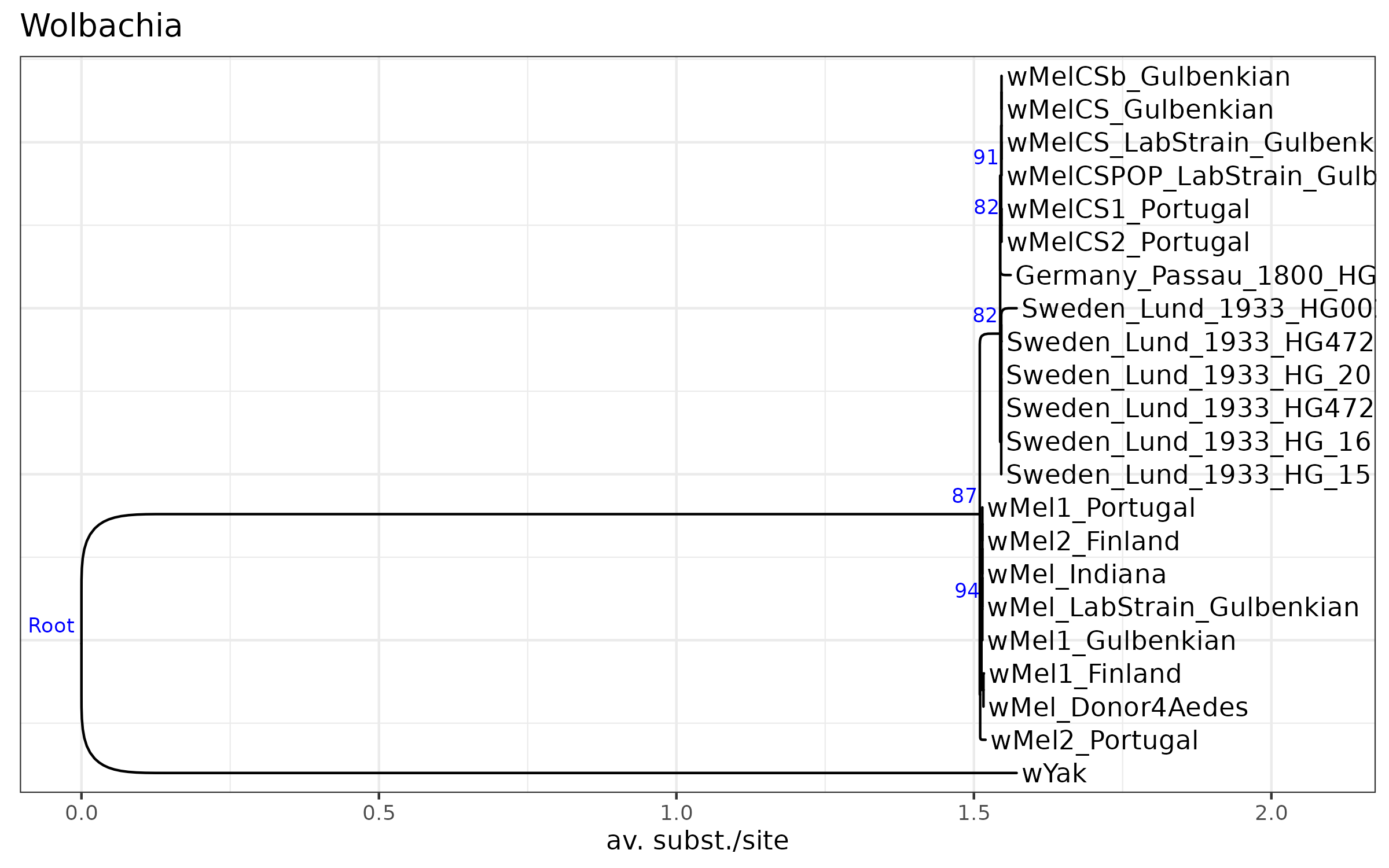
##### 3.3.2 Candidate genes



BUSCO\_Phylo

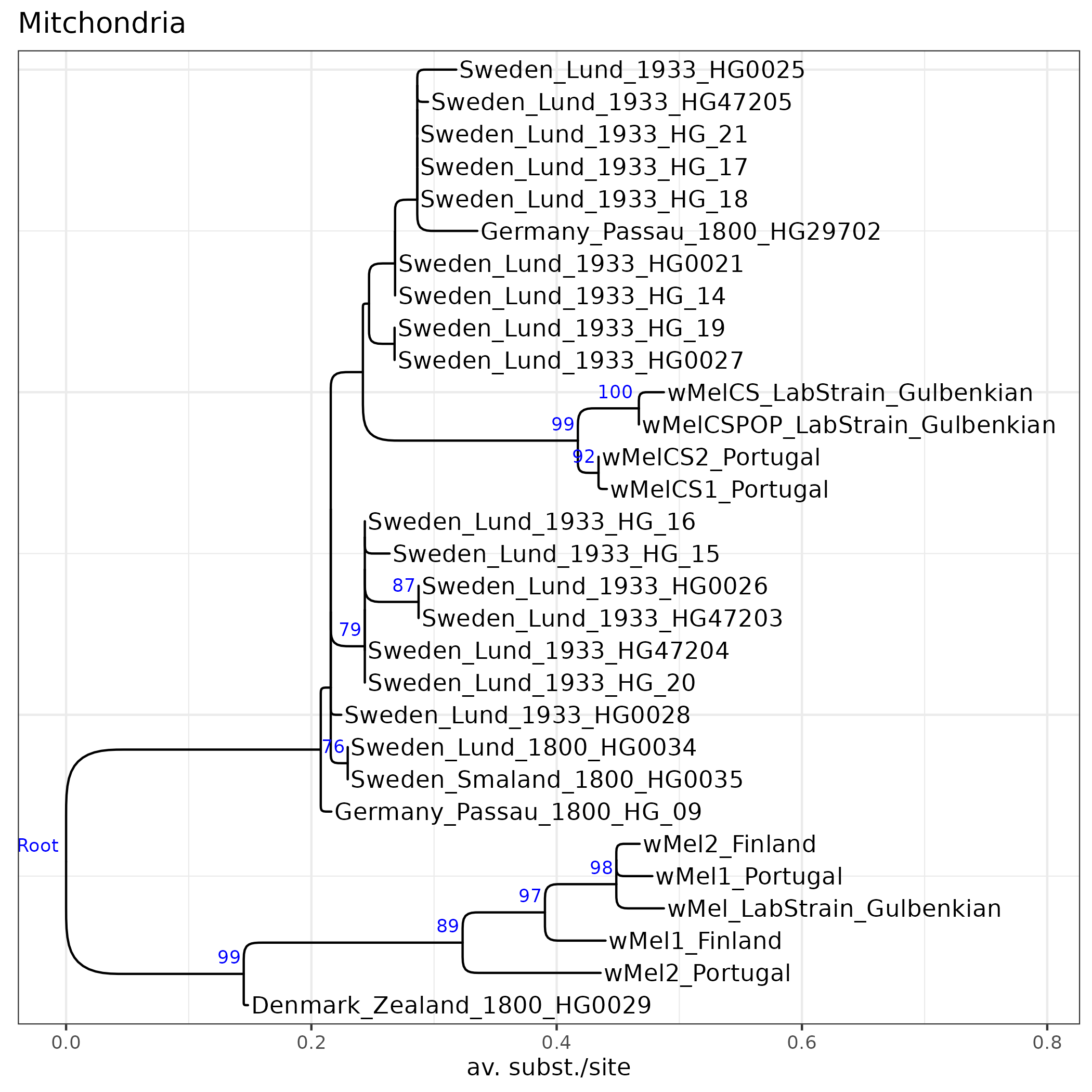
##### 3.3.3 SNP-based

*Wolbachia*



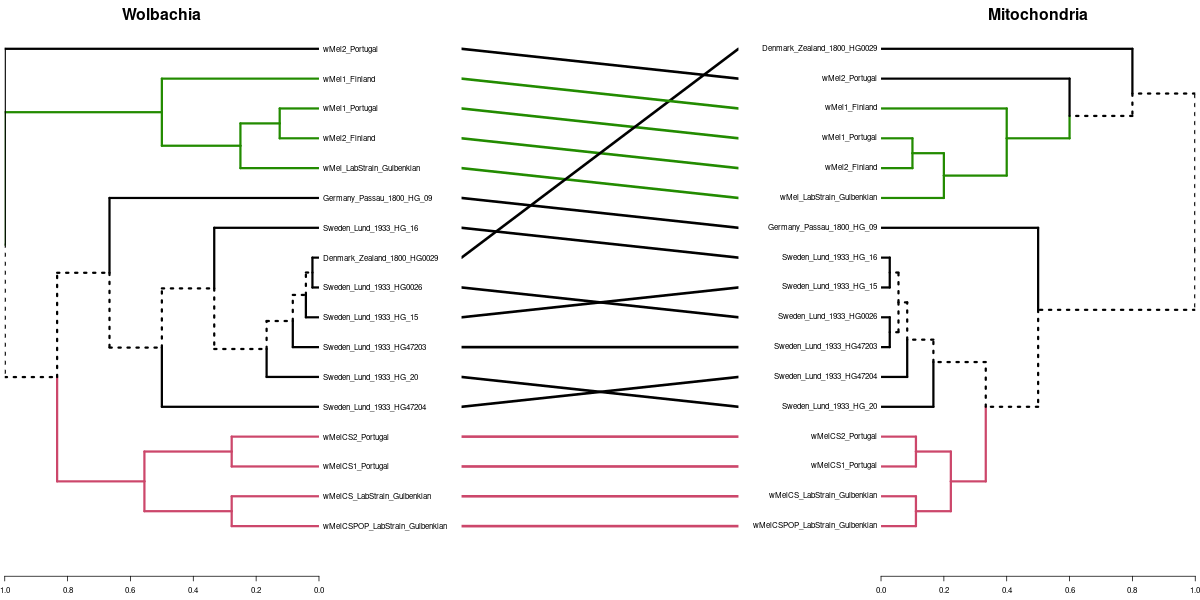
Wolb\_SNPs

Mitochondira



Mito\_SNPs

##### 3.4 Tanglegram



Tanglegram

### Discussion

### Acknowledgments

### References