One is not enough: on the effects of reference genome for the mapping and subsequent analyses of short-reads

Carlos Valiente-Mullor[1], Beatriz Beamud[1,*], Iván Ansari[1], Carlos Francés-Cuesta[1], Neris García-González[1], Lorena Mejía[1], Paula Ruiz-Hueso[1], Fernando González-Candelas[1,2,*]

1) Joint Research Unit "Infection and Public Health" FISABIO-University of Valencia, Institute for Integrative Systems Biology (I2SysBio), Valencia, Spain
2) CIBER in Epidemiology and Public Health, Valencia, Spain

*Authors for correspondence

Beatriz Beamud, beatriz.beamud@uv.es, +34 961925966
Fernando González-Candelas, fernando.gonzalez@uv.es, +34 961925961

Abstract

Mapping of high-throughput sequencing (HTS) reads to a single arbitrary reference genome is a frequently used approach in microbial genomics. However, the choice of a reference may represent a source of errors that may affect subsequent analyses such as the detection of single nucleotide polymorphisms (SNPs) and phylogenetic inference. In this work, we evaluated the effect of reference choice on short-read sequence data from five clinically and epidemiologically relevant bacteria (*Klebsiella pneumoniae*, *Legionella pneumophila*, *Neisseria gonorrhoeae*, *Pseudomonas aeruginosa* and *Serratia marcescens*). Publicly available whole-genome assemblies encompassing the genomic diversity of these species were selected as reference sequences, and read alignment statistics, SNP calling, recombination rates, d*N*/d*S* ratios, and phylogenetic trees were evaluated depending on the mapping reference. The choice of different reference genomes proved to have an impact on almost all the parameters considered in the five species. In addition, these biases had potential epidemiological implications such as including/excluding isolates of particular clades and the estimation of genetic distances. These findings suggest that the single reference approach might introduce systematic errors during mapping that affect subsequent analyses, particularly for data sets with isolates from genetically diverse backgrounds. In any case, exploring the effects of different references on the final conclusions is highly recommended.

Author summary

Mapping consists in the alignment of reads (i.e., DNA fragments) obtained through high-throughput genome sequencing to a previously assembled reference sequence. It is a common practice in genomic studies to use a single reference for mapping, usually the ‘reference genome’ of a species —a high-quality assembly. However, the selection of an optimal reference is hindered by intrinsic intra-species genetic variability, particularly in bacteria. Biases/errors due to reference choice for mapping in bacteria have been identified. These are mainly originated in alignment errors due to genetic differences between the reference genome and the read sequences. Eventually, they could lead to misidentification of variants and biased reconstruction of phylogenetic trees (which reflect ancestry between different bacterial lineages). However, a systematic work on the effects of reference choice in different bacterial species is still missing, particularly regarding its impact on phylogenies. This work intended to fill that gap. The impact of reference choice has proved to be pervasive in the five bacterial species that we have studied and, in some cases, alterations in phylogenetic trees could lead to incorrect epidemiological inferences. Hence, the use of different reference genomes may be prescriptive to assess the potential biases of mapping.

Introduction

The development and increasing availability of high-throughput sequencing (HTS) technologies, along with bioinformatic tools to process large amounts of genomic data, has facilitated the in depth study of evolutionary and epidemiological dynamics of microorganisms [1–3]. Whole-genome sequencing (WGS)-based approaches are useful to infer phylogenetic relationships between large sets of clinical isolates [4–7], showing improved resolution for molecular epidemiology [8–11] compared to traditional typing methods [12–14]. Short-read mapping against a single reference sequence is a commonly used approach in bacterial genomics for genome reconstruction of sequenced isolates and variant detection [4,6,15–17]. Nevertheless, there are grounds for suspecting that this approach might introduce biases depending on the reference used for mapping. Most of these errors originate in the genetic differences between the reference and the read sequence data [18–21], and they can affect subsequent analyses [22–28]. These include the identification of variants throughout the genome (mainly single nucleotide polymorphisms [SNPs]) and phylogenetic tree construction, which are essential steps for epidemiological and evolutionary inferences.

Sequencing status, completeness, assembly quality and annotation are relevant factors in reference selection, which explain the widespread use of the NCBI-defined reference genome of a species for mapping [26,28]. However, these criteria do not necessarily account for the amount of genetic information shared between the reference and subject sequences [29], neither the intrinsic genomic variability of the different bacterial species, which is reflected in their pangenomes (i.e., the total gene set within a species or within a particular sequence data set) [30]. It has been suggested that the impact of reference selection in clonal bacteria such as *Mycobacterium tuberculosis [31]* could be ameliorated by its limited variability at the intra-species level [25,28], although its effects on epidemiological inferences have been described [32]. In contrast, we expect a greater impact of reference choice in species with open pangenomes (e.g., *Pseudomonas aeruginosa [33]*) and/or highly recombinogenic bacteria (e.g., *Neisseria gonorrhoeae [34]* or *Legionella pneumophila [35]*). In spite of the awareness of the problem of reference selection considering the high genomic diversity of most bacterial species, systematic studies on the effect of reference choice in bacterial data sets are still missing, particularly if we are concerned with the consequences on epidemiological or evolutionary inferences. In addition, previous studies considering reference selection explicitly have been mainly focused on biases in SNP calling [23,24,28] and have not addressed other possible implications.

De novo assembly of read sequence data dispense with the need of using a reference genome. However, this requires higher sequencing coverage and longer reads in order to obtain enough read overlap at each position of the genome. Therefore, obtaining unfinished or fragmented assemblies is a major drawback, particularly when using short-reads (which still are the most frequently used in HTS-based studies) [36]. Complementarily, *de novo* assembled isolates could be used as reference genomes if previously assembled, high-quality references are found to be suboptimal in terms of genetic relatedness to the newly sequenced isolates [12,32,37]. However, this solution still has to deal with the additional costs of long-read sequencing and mapping errors derived from using a low-quality or fragmented reference.

In this work, we have analyzed the effect of reference selection on the analysis of short-read sequence data sets from five clinically and epidemiologically relevant bacteria (*Klebsiella pneumoniae*, *Legionella pneumophila*, *Neisseria gonorrhoeae*, *Pseudomonas aeruginosa* and *Serratia marcescens*) with different core and pangenome sizes [38–41]. WGS data sets were mapped to different complete and publicly available reference genomes, encompassing the currently sequenced genomic diversity of each species. We have studied the effect of reference choice on mapping statistics (mapped reads, reference genome coverage, average depth), SNP calling, phylogenetic inference (tree congruence and topology) as well as parameters of interest from an evolutionary perspective such as the inference of natural selection and recombination rates. Particular emphasis has been given to the effects of reference selection that result in misleading epidemiological inferences.

Results

Selection of reference genomes

Complete genome sequences of five pathogenic bacterial species were downloaded from GenBank. These included *K. pneumoniae* (270 genomes), *L. pneumophila* (91 genomes), *N. gonorrhoeae* (15 genomes), *P. aeruginosa* (150 genomes) and *S. marcescens* (39 genomes). Only one strain from *P. aeruginosa* (KU, accession number CP014210.1) was discarded because of low assembly quality (32% of ambiguous positions). We built a ML core genome tree showing the phylogenetic relationships between the available assemblies for each species (S1 Fig). Based on this phylogenetic information and the strains commonly used in the literature, we selected 8 reference genomes for *K. pneumoniae*, 7 for *L. pneumophila*, 3 for *N. gonorrhoeae*, 6 for *P. aeruginosa* and 4 for *S. marcescens* (S1 Table), including the NCBI reference genome of each species. The strains 342 and AR_0080 (*K. pneumoniae*), and U8W and Lansing 3 (two *L. pneumophila* strains not included in subsp. *pneumophila*), and PA7 (a known 'taxonomic outlier' of *P. aeruginosa*) showed ANI values <95% in pairwise comparisons with the remaining selected references (S3 Table) and long branches separating them from the other references in their corresponding phylogenies (S1 File).

In silico MLST typing was performed for all the reference genomes except those of *S. marcescens*. The only cases of shared STs were found in strains HS09565, HS102438 and NTUH-K2044 of *K. pneumoniae* (ST 23), and in strains 32867 and CAV1761 of *N. gonorrhoeae* (ST 1901).

Mapping to different references

We randomly sampled 20 isolates from different whole-genome sequencing data sets of the five bacterial species. Next, filtered and trimmed paired-end reads of each isolate were mapped to each reference genome from the same species. We computed different parameters for each mapping (S4 Table). The proportion of mapped reads and coverage of the reference genome (i.e., the percentage of reference genome covered by the aligned reads) showed variability depending on the reference used for mapping (Figs 1 and 2). Both parameters followed a roughly similar trend, as they presumably depend on the genetic distance between isolates and reference genomes. Moreover, we observed overlaps between the values obtained from mappings of the same isolates against different reference sequences in the five species. In most cases, the lowest median values were obtained in the alignments against the most genetically distant reference genomes (see 'Selected isolates and reference genomes'). However, the largest gap between median values depending on reference choice was found in the *S. marcescens* data set: the alignments to the outbreak-related reference UMH9 showed a high proportion of mapped reads (96.7%) and genome coverage (97.7%), whereas the alignment against the remaining references resulted in median values lower than 89% for both parameters. This was probably due to the high proportion of mapped reads and genome coverage resulting from mappings of outbreak isolates against a very close reference genome. Differences in both parameters were found to be significant (Kruskal-Wallis, $P < 0.05$) depending on the reference used for mapping in all species but *N. gonorrhoeae*. In the case of genome coverage, most pairwise comparisons (50%-100% in the four species) were found to be significant (Wilcoxon, $P < 0.05$), whereas the number of significant comparisons was lower for the proportion of mapped reads (Table 1). For example, in the case of *K. pneumoniae*, only 2 (out of 28) comparisons, involving the most genetically divergent reference genomes, showed significant differences in the proportion of mapped reads.

***** Place Fig 1. around here *****

***** Place Fig 2. around here *****

Table 1. Proportion of significant ($P<0.05$) comparisons depending on reference choice.

		Proportion (%) of significant comparisons

Species	**Comparisons**	**Mapped reads**[a]	**Genome coverage**[a]	**SNPs**[a]	**ρ**[b]	**d*N*/d*S***[a]
K. pneumoniae	28	7.1	75.0	53.6	42.9	60.7
L. pneumophila	21	19.0	52.4	95.2	23.8	47.6
N. gonorrhoeae	3	0.0	0.0	66.7	0.0	66.7
P. aeruginosa	15	26.7	93.3	86.7	73.3	53.3
S. marcescens	6	50.0	100	83.3	83.3	83.3

[a] Pairwise Bonferroni-corrected Wilcoxon tests.

[b] Pairwise Kolmogorov-Smirnov tests.

The average coverage depth (i.e., mean number of reads covering each position of the reference genome) was only slightly affected by reference choice (Fig 3, S4 Table). Its effect was noticeable when reads were mapped to the most divergent reference genomes of the different species, as in the previous parameters. However, the average depth seemed to be more dependent on other factors such as the total number of reads (sequencing coverage) of the isolates rather than on the genetic distance to the reference genome. One such example is isolate NG-VH-50 (*N. gonorrhoeae*), which had a low total number of reads and also showed low average depth values regardless the reference selected for mapping (S5 Table). Differences in this parameter depending on the reference used for mapping were found to be non-significant in all the species, according to Kruskal-Wallis tests.

***** Place Fig 3. around here *****

SNP calling

SNPs were called and quality-filtered from the different mappings to each reference of the five species. The number of quality SNPs showed high variability depending on the reference used.

Overlapping ranges of the number of called SNPs were found when comparing the results of the same isolates aligned to different reference sequences (Fig 4). Thus, considering that the number of SNPs between sequences is directly related to their genetic distance, SNP-calling results reflect genetic heterogeneity among isolates selected from the same species, as individual isolates showed different genetic relatedness to the different references.

***** Place Fig 4. around here *****

An overall inverse relationship between SNP count and the previously discussed alignment parameters (mapped reads and genome coverage) was also observed (see Figs 1, 2 and 4). This implies that, in most cases, more SNP calls were expected in alignments with a lower proportion of mapped reads and genome coverage (which is roughly indicative of a worse performance of the read mapping process).

A relationship between the genetic distance of isolates to the reference sequence and the total number of SNPs called was clearly observed in the alignments against the most distant reference genomes of *K. pneumoniae*, *L. pneumoniae* and *P. aeruginosa*. These sequences, whose distances to all the isolates were expected to be high, showed SNP counts one order of magnitude larger than to other reference sequences (S4 Table).

In the case of *S. marcescens*, the alignments to strain UMH9 resulted in significantly fewer SNP calls when compared to mappings against the remaining reference sequences. This is explained by the presence of nearly identical isolates (outbreak isolates) to strain UMH9 (<160 SNPs detected). A similar case was found in *L. pneumophila* isolates 28HGV and 91HGV, which appeared to be nearly identical to the reference strain Paris, as less than 100 SNPs were detected in their respective mappings to this sequence. In all the species, most comparisons (53%-95%) between called SNPs from mappings against different references were significant (Wilcoxon, $P < 0.05$) (Table 1).

Phylogenetic analyses and tree comparisons

We obtained a collection of MSAs including the same isolates and reference sequences, but differing only in the reference used for mapping by removing the regions absent in each mapping reference. We also obtained a 'core' genome MSA by removing simultaneously all the regions absent from any of the reference genomes for each species. Then, ML trees were inferred from each MSA. Due to methodology used to obtain the MSAs, the comparison between phylogenies strictly implies assessing the impact of reference selection.

Firstly, we quantified the topological distances between phylogenetic trees from each species with Robinson-Foulds clusters (RF) and matching clusters (MC) metrics. Tree distances spanned a variable range of values depending on the species (Table 2, S6 Table). The normalized values of both metrics for the same tree comparisons were not equal (in most cases) but followed a similar global trend (Fig 5).

***** Place Fig 5. around here *****

Table 2. Descriptive statistics of topological distances per species.

	Matching clusters					Robinson-Foulds clusters				
Species	**Mean**	**Median**	**SD**	**Min**	**Max**	**Mean**	**Median**	**SD**	**Min**	**Max**
K. pneumoniae	57.7	49	34.4	0	99	12.4	11	6.2	0	20
L. pneumophila	43.9	42	16.0	5	67	9.7	11	3.5	1	14
N. gonorrhoeae	40.0	44	13.5	25	51	8.7	10	3.2	5	11
P. aeruginosa	49.9	47	16.1	25	80	12.3	12	4.2	7	19
S. marcescens	31.3	29.5	7.9	21	43	6.5	6.5	2.9	3	10

The comparisons involving phylogenies that include sequences mapped to the most divergent reference genomes of *K. pneumoniae* and *P. aeruginosa* showed the largest distance values. However, in most cases there was not a straightforward relationship between the genetic distance to the reference genomes and the topological distance between the corresponding trees (Fig 6). For example, *K. pneumoniae* trees using sequences from mappings to strains 342 and AR_0080 showed an identical topology (RF=0, MC=0), despite the ANI value between these references was <94%.

***** Place Fig 6. around here *****

The congruence between different tree topologies was rejected in most comparisons by ELW tests (Table 3). The few cases in which congruence was not rejected could be explained by the close phylogenetic relationship between the reference genomes involved.

Table 3. Congruent comparisons according to ELW test. All the other pairwise comparisons were not congruent.

Species	**Reference**	**Congruent pair**
K. pneumoniae	HS09565	HS09565, NTUH-K2044
	HS102438	HS102438, NTUH-K2044
	NTUH-K2044	NTUH-K2044, HS09565
	342	342, AR_0080
	AR_0080	AR_0080, 342
L. pneumophila	Lansing 3	Lansing 3, U8W

Finally, in order to assess in detail the effects of reference selection on phylogenetic inference, trees from the same species were compared qualitatively. Changes in the phylogenetic relationships were found when using different reference sequences in almost all cases except for two identical topologies. In some cases, the changes only affected branches in clades including closely related isolates (Fig 7A and 7B), while others implied more profound changes in the resulting topologies. Moreover, the alignments against a single reference genome seemed to underestimate the genetic distance between the consensus sequences of the isolates and the reference sequence. Branch lengths were thus shortened between the leaves involved. In some extreme cases (when mapping to genetically distant genomes 342, AR_0080 [*K. pneumoniae*], Lansing 3, U8W [*L. pneumophila*] and PA7 [*P. aeruginosa*]), this 'attraction' effect led to the clustering of reference genomes not used as references for mapping in a single clade, regardless their genetic distance to the isolates (Fig 7D). These differences were also observed when only the core genome was used to obtain the phylogenetic tree (S3 File). Additional species-specific differences are described next.

***** Place Fig 7. around here ******

K. pneumoniae. The topologies inferred with KP1768, NTUH-K2044, HS09565 and HS102438 as reference sequences revealed the same phylogenetic relationships between clusters of isolates, although there were some differences within clusters depending on the reference used for the MSA. Isolates HGV2C-06 and HCV1-10 (not associated to any of these clusters) changed their placement in the topologies with HS11286 and AR_0143 as reference sequences (Fig 8). The tree topologies using 342 and AR_0080 as reference genomes were identical and markedly different to the phylogenies derived with the other reference strains (S2 File).

**** Place Fig 8. around here *****

L. pneumophila. The tree topologies using Lansing 3 and U8W as reference genomes were the most similar ones for this species (RF=1, MC=5) despite the large genetic distance between these sequences (ANI < 94%). Their topology was markedly different from the remaining topologies, where isolates grouped in three clades associated with reference genomes Paris, Alcoy and Philadelphia 1, respectively (see Fig 7, S2 File). Notably, because of the epidemiological implications discussed below, isolates 28HGV and 91HGV were included in the Alcoy clade only when mapped to this reference genome (Fig 7C), whereas in all other cases (excluding U8W and Lansing 3) the isolates grouped with the Paris strain.

N. gonorrhoeae. The most similar topologies resulted from using FA 1090 and 32867 as reference genomes, despite that 32867 and NCTC13798 had larger ANI values. Three clades of isolates could be identified in all the phylogenies. However, those isolates not included in any of these clusters changed their position in the tree when using NCTC13798 as reference sequence in comparison with the two other trees. As an exception, isolate NG-VH-50 always grouped close to the reference sequence it was mapped to (S2 File). This artifact was due to the low total number of reads obtained in sequencing this strain.

P. aeruginosa. Three clades were clearly identified in all the trees, with the exception of the one inferred using PA7 as reference sequence. In this tree, PA7 was placed in a cluster of isolates, whereas the remaining reference sequences clustered together (S2 File). The main topological differences depending on the reference were: (a) the placement of reference genome M18 and the isolate P5M1 in the tree, and (b) the phylogenetic relationships within the clade of reference genomes and P6M6, where the sequence chosen as reference for mapping occupied a basal position in the clade (Fig 9).

***** Place Fig 9. around here ******

S. marcescens. Outbreak isolates grouped with strain UMH9 in all the trees. Branch lengths within this clade were practically null when UMH9 was used as the reference sequence, but these lengths increased when other reference sequences were used (Fig 10). As expected, the control isolate SMElx20 grouped with its closest reference (Db11) in all the cases. The phylogenetic relationships between reference genomes, isolates and clades changed depending on the reference used. The reference genome WW4 grouped with isolate CNH62 in all the topologies except when this strain was used as reference (S2 File).

***** Place Fig 10. around here *****

Distribution of recombination rates

Population recombination rates (*ρ*) were computed for 1000 bp sliding windows of the MSAs (S4 Table) and the corresponding distributions were compared. Those regions that were not present in all the sequences of a species were removed from the alignments for these analyses.

Overall, the distributions of recombination rates were very similar regardless the reference genome used in each case. However, relevant differences in some peaks were found in different MSAs from the same species. For example, the MSAs built with 32867 or NCTC13798 (*N. gonorrhoeae*) as reference sequences showed at least two clearly observable peaks that were absent when FA 1090 was the reference (Fig 11).

***** Place Fig 11. around here *****

The number of significant pairwise comparisons between distributions of recombination rates (Kolmogorov-Smirnov, $P < 0.05$) differed widely depending on the species. While none of the comparisons between distributions of *N. gonorrhoeae* sequences showed significant results (although, as described previously, relevant differences were found), almost all *S. marcescens* estimated distributions were found to be significantly different (83.3%) (Table 1). In most cases, the significance of the comparisons between recombination rates could be explained by the phylogenetic relationships among the reference genomes. For example, the comparisons involving the most distant reference sequences of *K. pneumoniae*, *L. pneumophila* and *P. aeruginosa* showed significant differences, with the exception of the mutual comparisons between U8W and Lansing 3 (*L. pneumophila*), as well as AR_0080 and 342 (*K. pneumoniae*). Moreover, the significant comparisons in *P. aeruginosa* roughly reflected genetic distances between reference sequences, because using phylogenetically close reference sequences (M18 and PAO1 or UCBPP-PA14, Pa124 and 12939) resulted in non-significant differences between recombination rate distributions. In the case of *S. marcescens*, generalized significant comparisons could reflect nearly homogeneous divergence among the four reference genomes (S1 File).

Analysis of natural selection

Changes in the ratio ω (= d*N*/d*S*) due to reference choice could affect inferences on how natural selection has acted throughout the genome. This parameter was estimated in pairwise comparisons between concatenated CDS extracted from consensus sequences obtained from the mappings (S4 Table).

In all cases, the d*N*/d*S* values computed for each gene were <1. Differences in d*N*/d*S* depending on the reference used (Fig 12) were significant (Kruskal-Wallis, $P < 0.05$) for all the species. The proportion of significant pairwise comparisons (Wilcoxon, $P < 0.05$) depended on the species, ranging from 47.7% (*L. pneumophila*) to 83.3% (*S. marcescens*) (Table 1). In contrast with the results obtained in the parameters discussed previously, some of the comparisons involving the most genetically distant reference genomes (e.g., 342 strain of *K. pneumoniae*) as mapping references were not significant. Therefore, in this case it is difficult to explain the variability of ω based on the genetic distances between reference sequences for most species. *N. gonorrhoeae* could be treated as an exception, because the comparisons involving the reference strain FA 1090 (the most genetically distinct one) were the only significant ones. These differences were also observed when only the core genome was used to compute ω.

***** Place Fig 12. around here *****

Discussion

The impact of using different reference sequences for mapping NGS data sets has been studied previously in clinically relevant bacteria such as *Escherichia coli [22]*, *Salmonella enterica [26]*, *Listeria monocytogenes [23,24,28,42]* or *Mycobacterium tuberculosis [25,28]*, as well as in eukaryotes [21,43,44], including *Homo sapiens [45]*. However, a systematic analysis of the evolutionary and epidemiological implications of reference choice, encompassing different bacterial species and diverse reference genomes is still missing. This work has been aimed at filling this gap. Indeed, in some cases, reference selection analysis is incidental, spanning a restricted number of reference sequences [46]. Among the species included in this work, the influence of reference diversity on SNP calling has been previously assessed in *K. pneumoniae* and *N. gonorrhoeae [28]*, whereas *L. pneumophila*, *P. aeruginosa* (both showing high genomic variability [33,35]) and *S. marcescens* have not been studied under this perspective.

Statistics on raw mapping data such as the proportion of mapped reads and the coverage of the reference genome can provide preliminary information on the effect of reference choice and its effects on subsequent analyses, because these parameters reflect the performance of read alignment. As suggested previously, the genetic distance between short-read data and the reference genome is directly related to incorrect read alignment and unmapped reads due to mismatches between the sequence of the reads and the homologous positions in the reference [19,20,22]. This is also confirmed by our results on read alignment statistics. The percentage of the reference genome covered by mapped reads may be affected not only by genetic differences in homologous regions, but also by the presence of strain-specific genomic regions [21], because genes absent in the reference genome are expected to be lost during the mapping and in the subsequent multiple alignment. Moreover, as proposed by Lee and Behr [25], there might exist a coverage threshold beyond which subsequent phylogenetic analyses would be strongly affected, thus reducing the accuracy of evolutionary and epidemiological inferences derived from such inaccurate mappings.

The effect of sequencing coverage of the isolates on mapping seems to be generally independent of reference choice, as shown by the values of average coverage depth obtained in this study. Similarly to Pightling *et al. [23]*, we have not observed any relationship between sequencing coverage and other variables during HTS data processing. However, as shown by one *N. gonorrhoeae* isolate (NG-VH-

50), the reference mapping approach could strongly underestimate the genetic distance between the assembly of the genome of a particular isolate and that of the reference genome below a certain threshold of total reads, thus affecting subsequent phylogenetic inferences.

Benchmarking of SNP calling performance for HTS data seems to be more common compared to other steps of genomic analyses [27,47–54]. Although most of these works are focused on assessing the effect of the selected pipeline (and its underlying algorithm), the use of different reference sequences has also been identified as a potential source of biases that could interact with other variables of the pipeline such as selection of the variant caller and read alignment software [23,24,28]. The number of SNPs is often used as a criterion for defining clusters of epidemiologically related isolates [55]. Our results confirm the existence of a systematic and significant influence of reference choice on the number of identified SNPs in all the species analyzed. They also reflect the correlation between genetic distance of isolates to the reference genome and the number of called variants which, as highlighted in previous studies, could be associated with the increase of false positives when the precision of SNP calling decreases [23,28,42]. Overlapping ranges in the number of SNPs called depending on the reference sequence used for mapping reflects the genomic heterogeneity within the sets of isolates selected from each species.

Recovering phylogenetic relationships between organisms or strains within a species represents an essential procedure in evolutionary and epidemiological studies. Biases in how and how many SNPs are called as well as in the gene content of the final assemblies due to reference choice could affect phylogenetic inferences [47]. The overall negative results obtained in congruence tests also reflect the existence of a systematic effect of reference choice on tree topologies: the only statistically concordant comparisons (6 out of 73) between topologies of the same species were found when references chosen for mapping were (a) closely related sequences (*K. pneumoniae* ST 23 strains), or (b) extremely distant sequences, showing ANI values close to the boundaries for species delimitation. The topologies resulting from using phylogenetically unrelated, extremely divergent genomes were mutually similar while, in contrast, generally showed high topological distance values when compared to trees built using non-extreme references. This kind of loss in tree resolution has already been observed (although limited to clonal bacteria [25]). In our case, it may be originated from a reduced proportion of shared gene content between isolates and extremely divergent sequences, along with the existence of barriers to recombination between populations, as the ability for recombination and its frequency is expected to decrease with genetic distance [56]. However, these differences were also observed when considering only the core genome. This suggests that the effect of the reference on phylogenetic inference is not only due to the presence/absence of genes in the accessory genome. It might be due also to differences in core genome sequences arising from biased/erroneous identification of variants.

The effect of reference choice on phylogenetic inferences is pervasive in these five species. However, despite the differences between topologies and even lack of congruence, these changes might not be necessarily associated with altered epidemiological inferences. A similar situation was studied by Usongo *et al. [26]* on a *S. enterica* epidemiological data set, in which two different topologies (RF=24) were resolutive enough to distinguish different outbreak clusters. However, we have observed that the use of different reference sequences affects phylogenetic relationships between clades and even to the association of specific isolates to transmission clusters, thus potentially affecting epidemiological inferences. This has been observed even when using phylogenetically related strains from the same non-clonal species as a reference, in contrast with previous studies in clonal bacteria [25] where differences in phylogenetic inference appeared when using reference genomes from close but different species. This is most obvious in the *L. pneumophila* data set, in which two isolates changed their positions and were placed in the same cluster of the reference sequence used for mapping, while the overall topology remained practically unchanged.

Differences between trees were quantified by topological distance metrics, reflecting, in most cases, lack of correlation between tree distances and genetic distances of the corresponding reference genomes. As suggested previously [22,27], when working with a genetically diverse set of isolates, it is impossible to select a single reference close to all of them, and single-reference mapping biases are expected to increase with genomic divergence. Therefore, these differences in tree topologies could be partially explained by the use of genetically heterogeneous data sets. Moreover, its impact on tree reconstruction may be alleviated by using multiple references or a reference pangenome instead [22,57–60]. If data sets of isolates were homogenous (i.e., the isolates are equally close to the same reference) as the one employed by Lee and Behr [25], we would expect that read alignment performance and tree resolution would decrease as we select progressively distant reference genomes [23,24,28].

However, we could not ignore that the presence of recombination (particularly in highly recombinogenic species such as *K. pneumoniae* and *L. pneumophila*) could reduce accuracy in phylogenetic reconstruction [22], thus explaining to some extent the topological incongruence or the differences in branch lengths [61].

Selecting one reference or another for mapping can also affect the estimates of phylogenetic distance between isolates [22,26], which is reflected in the branch lengths of the trees. This is clearly illustrated by the phylogenetic analysis of the *S. marcescens* data set, which reveals that tree branches connecting outbreak isolates increased their lengths when consensus sequences were calculated from alignments using reference genomes that were phylogenetically unrelated to the isolates (different from strain UMH9). Similar findings were observed for *Listeria monocytogenes* sequences by Pightling *et al. [23]*.

The development and increasing availability of high-throughput, whole-genome sequencing technologies have allowed assessing evolutionary rates and dynamics at the genome level which, in turn, contribute to a better understanding of emerging diseases and transmission patterns [62]. Therefore, the study of natural selection and recombination, frequent processes in bacteria [63], is relevant not only from an evolutionary point of view but also in its application to molecular epidemiology [64]. The impact of reference selection on the inference of evolutionary parameters such as substitution and recombination rates at the genome level has not been explored thoroughly previously. In this work, variations in d*N*/d*S* and *ρ* have been detected in all the species depending on the reference sequence used for mapping. This might have an effect in subsequent inferences on the action of natural selection and the detection of recombination events. Significant differences in *ρ* seemed to be more strongly correlated with the genetic distance between the genomes used as reference for mapping than d*N*/d*S*.

Short-read mapping of HTS data against a reference genome is a common approach in bacterial genomics. Our results show that the impact of selecting a single reference is pervasive in the genomic analyses of five different bacterial species, and likely in many others. All the parameters evaluated were affected by the usage of different reference sequences for mapping and, notably, alterations in phylogenetic trees modified in some cases the epidemiological inferences. Furthermore, working with heterogeneous sets of isolates seems to be a particularly challenging scenario for the selection of a single reference genome. Mapping simultaneously to multiple references or against a reference pangenome may reduce the effect of reference choice. In any case, exploring the effects of different references on the final conclusions is highly recommended.

Methods

The workflow used in this study is summarized in Fig 13.

***** Place Fig 13. around here *****

For each species, we selected different (3-8) publicly available closed whole-genome sequences as references and 20 sets of short-reads from whole-genome sequencing projects. Reads were mapped to each selected reference genome per species and consensus sequences were obtained from quality SNPs of each mapping. Consensus sequences from the mappings to the same reference genome were added to the MSA of all references of each species. For the analysis of each MSA, (a) we considered only those genome regions present in the reference used for mapping and (b) we obtained a ‘core’ MSA by removing all the regions absent from any of the reference sequences. Finally, we studied the impact of reference choice on the ML trees inferred from each MSA, recombination rates calculated on 'core' MSAs and d*N*/d*S* ratios calculated considering only coding sequences.

Selection of reference genomes

Closed whole-genome sequences of *K. pneumoniae*, *L. pneumophila*, *N. gonorrhoeae*, *P. aeruginosa* and *S. marcescens* available in June, 2018 were downloaded from NCBI GenBank [65] in fasta format. Plasmids were removed with seqtk v1.0 (https://github.com/lh3/seqtk) (subseq command). Genome sequences were annotated using Prokka v1.12 [66] (with default settings) and the set of intra-species co-orthologous genes was inferred using Proteinortho v5.11 [67] (option -p=blastn+). Coding sequences (CDS) of orthologous genes in each species were aligned with MAFFT v7.402 [68] (with default settings) and concatenated to obtain a CDS-coding core genome multiple sequence alignment (MSA) for each species.

A maximum-likelihood (ML) tree was inferred from each MSA with IQ-TREE v1.6.6 [69] using the GTR substitution model and 1000 fast bootstrap replicates [70]. After consideration of the core genome phylogenies (distance between strains and clusters) and the usage of different references in the literature, we selected a set of genomes to be employed as reference genomes for each species. The number of reference sequences selected was roughly proportional (≈10%) to the initial number of publicly available sequences from each species. In brief, we included (a) the NCBI reference genome of the species, (b) relevant or commonly used references for mapping, and (c) representative sequences of different lineages. Detailed information about the selected reference genomes is provided in S1 Table.

The selected reference genomes of each species were aligned with progressiveMauve v2.4 [71] and gaps were added to regions where homologous sequences were absent in any genome in the alignment (see 'Code availability'). The XMFA output alignment was converted into fasta format with xmfa2fasta.pl (https://github.com/kjolley/seq_scripts/blob/master/xmfa2fasta.pl).

To evaluate the genetic divergence between the selected reference sequences, we used three different procedures: (a) we built ML trees with IQ-TREE, as above, (b) we computed Average Nucleotide Identities [72] (ANIs) using FastANI v1.1 [73], and (c) we performed an *in silico* multi-locus sequence typing (MLST) using mlst v1.15.1 (https://github.com/tseemann/mlst) for *K. pneumoniae*, *N. gonorrhoeae* and *P. aeruginosa*; and using BLAST+ [74] and the EWGLI [75] database for *L. pneumophila*. This procedure was not used with *S. marcescens*.

Selection of isolates for analysis

20 sets of short-reads from whole genome sequencing projects of the five species (S2 Table) were randomly selected (with the R [76] function sample_n) among those obtained in our laboratory and/or deposited at the SRA as detailed next. Sequences in our laboratory were obtained with Illumina MiSeq 300x2 paired-ends (*P. aeruginosa*) or NextSeq 150x2 paired-ends (the remaining species). The *K. pneumoniae* data set included isolates of 9 different STs obtained in a surveillance study of ESBL-producing strains in the Comunitat Valenciana (Spain). The *L. pneumophila* data set comprised isolates obtained from environmental surveillance at 2 hospitals of the Comunitat Valenciana. The *N. gonorrhoeae* data set includes isolates obtained in a surveillance study in different regions of Spain (Comunitat Valenciana, Madrid and Barcelona). The *P. aeruginosa* data set included isolates from 2 outbreaks detected in the Comunitat Valenciana. Finally, the *S. marcescens* data set included 9 almost identical outbreak isolates genetically close to strain UMH9, one isolate close to the reference of the species, Db11, and 10 unrelated isolates downloaded from the SRA repository.

Quality control analysis and sequence read processing

The quality of the reads (before and after trimming and filtering) was assessed using FastQC v0.11.8 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and quality reports were merged with MultiQC v1.7 [77]. Illumina, Truseq and Nextera adapters were removed with cutadapt v1.18 [78]. Reads were trimmed and filtered using Prinseq-lite v0.20.4 [79]. 3'-end read positions with quality <20 were trimmed and reads with overall quality <20, >10% ambiguity content and total length <50 bp were removed.

Mapping, variant calling and consensus sequences

Reads passing the above filters were mapped to each selected reference of each species using BWA MEM v0.7.17 [80] (with default settings). SAM files were converted to binary format (BAM), sorted and indexed with samtools v1.6 [81] (commands sort and index). Mapping statistics were obtained using samtools (commands flagstats and depth).

SNPs were identified in each alignment with samtools and bcftools v1.6 [82] (commands mpileup and call, respectively). Indels were excluded from the analysis (option --skip-variants indels). Remaining SNPs after filtering (quality >40, mapping quality [MQ] >30, depth >10 and under twice the average depth and distance of >10 pb to any indel) were counted with bcftools (command stats).

Consensus sequences were obtained from quality-filtered SNPs and the appropriate reference sequence using bcftools (command consensus) for every possible combination of isolates and reference genomes from the same species.

Multiple sequence alignment of reference genomes and consensus sequences

The MSAs of the reference sequences from each species were used as ‘backbones’ on which the consensus sequences from the mappings to the same reference genome were added using a custom Python script (see ‘Code availability’). XMFA-formatted MSAs were converted to fasta format as described previously. Finally, for the analysis of each MSA we considered only those genome regions present in the reference genome, using a custom Python script (see ‘Code availability’) to mask the absent regions from the global MSA. This procedure (see Fig 13) allowed us to obtain a collection of MSAs (one per each reference sequence) including the same isolates and reference genomes (per species), differing only in the reference sequence used for mapping. In addition, we also obtained a ‘core’ genome MSA by removing all the regions absent from any of the reference sequences.

Analysis of natural selection

We explored the effect of reference choice on the inference of natural selection at the whole genome level by computing pairwise d*N*/d*S* ratios with the PAML package 4.9i [83] between concatenated CDSs of consensus sequences that were built using the same reference. CDSs were extracted using coordinates of the corresponding reference obtained with Prokka (see 'Selection of reference genomes'). A custom Python script (see 'Code availability') and the emboss package v6.6.0 [84] were used. We also computed pairwise dN/dS values between consensus sequences considering only the core genome CDSs (i.e., shared by all the selected references from each species).

Distribution of recombination rates

Population recombination rates ($\rho = 4N_e r$; where N_e is the effective population size and r is the recombination rate per base pair and generation) were estimated using LDJump [85] (with a window of 1000 pb) from the 'core' genome MSAs. The distributions of recombination rates along MSAs were compared for the different reference genomes of each species and were represented graphically with the R package ggplot2 [86].

Comparisons of phylogenetic trees

ML trees were inferred from each MSA with IQ-TREE as described above, and visualized with iTOL v4 [87].

Congruence tests. We used expected likelihood weight (ELW) tests [88], as implemented in IQ-TREE, to assess the congruence between phylogenies that differed only in the genome chosen as mapping reference. The ELW test computes weights for each topology based on its likelihood given a MSA, with the total sum of weights being equal to 1 and higher weights assumed to be those best supported by the data. Decreasing weights are progressively collected to build a confidence set until their cumulative sum is equal to or higher than 0.95. At this point, the trees included in the confidence set are accepted as congruent.

Topological distances. Pairwise distances between tree topologies obtained with the different mapping references were assessed using TreeCmp v2.0 [89]. Robinson-Foulds [90] clusters (RF) and matching clusters [89] (MC) metrics were calculated for each comparison. The RF distance reflects the number of bipartitions differing between topologies, whereas the MC distance computes the minimal number of moves needed to convert a topology into another. Therefore, two identical topologies will receive a value equal to 0 with both metrics. Conversely, distance values will increase as the compared trees become more different.

Qualitative comparison of trees. Finally, a qualitative assessment of trees was performed in order to identify specific changes in the phylogenetic relationships between isolates due to the choice of different reference genomes. Particularly, we focused on clustering of isolates and alterations that could affect epidemiological inferences (e.g., including/excluding one particular sample in an outbreak).

Statistical analyses

To study the effect of using different reference genomes on mapping statistics (proportion of mapped reads, genome coverage, average depth), number of called SNPs, and d*N*/d*S* values, non-parametric Kruskal-Wallis [91] tests were performed with R 3.5 (function kruskal.test). If a Kruskal-Wallis test showed significant differences between groups (reference sequence), we performed pairwise Wilcoxon [92] tests with Bonferroni-corrected p-value for multiple comparisons (with the R function pairwise.wilcox.test) in order to identify significant differences between specific reference sequences. Pairwise Kolmogorov-Smirnov [93] tests (R function pairwise_ks_test [https://github.com/netlify/NetlifyDS]), which compare observed distributions of data, were performed in order to identify significant differences in the distributions of recombination rates depending on the mapping reference.

Code availability

Custom scripts used in this work are available in https://github.com/cvmullor/reference.

Supporting information

S1 Fig. Core genome trees of the complete whole-genome sequences downloaded from GenBank. The circles at the tips denote the sequence type (ST) of the different strains in the trees of the species with an MLST scheme available for *in-silico* typing. The black triangles denote the branches with bootstrap support values <70. (A) *K. pneumoniae*, (B) *L. pneumophila* and (C) *P. aeruginosa* trees were rooted on their corresponding longest branches. As all the branches connecting the different clades of (D) *S. marcescens* and (E) *N. gonorrhoeae* trees were approximately the equal length, they were rooted arbitrarily for a better visualization.

S1 Table. Strains selected as references for mapping.

S2 Table. Isolates (short-read sequence data) selected for mapping.

S3 Table. ANI (%) calculated between the selected reference genomes.

S4 Table. Summary statistics per reference and species. Median, minimum and maximum values are shown.

S5 Table. Mapping and SNP statistics per reference and species.

S6 Table. RF and MC distances.

S1 File. Phylogenetic trees of the reference genomes selected for each species.

S2 File. Phylogenetic trees per reference and species. Strain selected as reference for mapping in each tree is indicated in the corresponding newick file name.

S3 File. 'Core' genome phylogenetic trees per reference and species. Strain selected as reference for mapping in each tree is indicated in the corresponding newick file name.

Funding

This project was partly funded by projects BFU2017-89594R from MICIN (Spanish Government) and PROMETEO2016-0122 (Generalitat Valenciana, Spain). WGS was performed at Servicio de Secuenciación Masiva y Bioinformática de la Fundación para la Investigación Sanitaria y Biomédica de la Comunitat Valenciana (FISABIO) and co-financed by the European Union through the Operational Program of European Regional Development Fund (ERDF) of Valencia Region (Spain) 2014-2020. CV is recipient of contract FPU2018/02579, BB of contract FPU2016/02139 and CF of

FPI contract BES-2015-074204 from MICIN (Spanish Government). LM benefits of a fellowship from Fundación Carolina. The sponsors had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Brockhurst MA, Colegrave N, Rozen DE. Next-generation sequencing as a tool to study microbial evolution. Mol Ecol. 2011 Mar;20(5):972–80.

2. Quainoo S, Coolen JPM, van Hijum SAFT, Huynen MA, Melchers WJG, van Schaik W, et al. Whole-Genome Sequencing of Bacterial Pathogens: the Future of Nosocomial Outbreak Analysis. Clin Microbiol Rev. 2017 Oct;30(4):1015–63.

3. Bentley SD, Parkhill J. Genomic perspectives on the evolution and spread of bacterial pathogens. Proc Biol Sci. 2015 Dec 22;282(1821):20150488.

4. Harris SR, Feil EJ, Holden MTG, Quail MA, Nickerson EK, Chantratita N, et al. Evolution of MRSA during hospital transmission and intercontinental spread. Science. 2010 Jan 22;327(5964):469–74.

5. Holt KE, Baker S, Weill F-X, Holmes EC, Kitchen A, Yu J, et al. Shigella sonnei genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe. Nat Genet. 2012 Sep;44(9):1056–9.

6. Kaiser T, Finstermeier K, Häntzsch M, Faucheux S, Kaase M, Eckmanns T, et al. Stalking a lethal superbug by whole-genome sequencing and phylogenetics: Influence on unraveling a major hospital outbreak of carbapenem-resistant Klebsiella pneumoniae. Am J Infect Control. 2018 Jan;46(1):54–9.

7. David S, Reuter S, Harris SR, Glasner C, Feltwell T, Argimon S, et al. Epidemic of carbapenem-resistant Klebsiella pneumoniae in Europe is driven by nosocomial spread. Nat Microbiol. 2019 Nov;4(11):1919–29.

8. Laabei M, Recker M, Rudkin JK, Aldeljawi M, Gulay Z, Sloan TJ, et al. Predicting the virulence of MRSA from its genome sequence. Genome Res. 2014 May;24(5):839–49.

9. Golparian D, Donà V, Sánchez-Busó L, Foerster S, Harris S, Endimiani A, et al. Antimicrobial resistance prediction and phylogenetic analysis of Neisseria gonorrhoeae isolates using the Oxford Nanopore MinION sequencer. Sci Rep. 2018 Dec 4;8(1):17596.

10. Nikolayevskyy V, Niemann S, Anthony R, van Soolingen D, Tagliani E, Ködmön C, et al. Role and value of whole genome sequencing in studying tuberculosis transmission. Clin Microbiol Infect. 2019 Nov;25(11):1377–82.

11. Sánchez-Busó L, Harris SR. Using genomics to understand antimicrobial resistance and transmission in Neisseria gonorrhoeae. Microb Genom [Internet]. 2019 Feb;5(2). Available from: http://dx.doi.org/10.1099/mgen.0.000239

12. Harris SR, Clarke IN, Seth-Smith HMB, Solomon AW, Cutcliffe LT, Marsh P, et al. Whole-genome analysis of diverse Chlamydia trachomatis strains identifies phylogenetic relationships masked by current clinical typing. Vol. 44, Nature Genetics. 2012. p. 413–9.

13. Allard MW, Strain E, Melka D, Bunning K, Musser SM, Brown EW, et al. Practical Value of Food Pathogen Traceability through Building a Whole-Genome Sequencing Network and Database. Vol. 54, Journal of Clinical Microbiology. 2016. p. 1975–83.

14. Pérez-Losada M, Arenas M, Castro-Nallar E. Microbial sequence typing in the genomic era. Vol. 63, Infection, Genetics and Evolution. 2018. p. 346–59.

15. McAdam PR, Templeton KE, Edwards GF, Holden MTG, Feil EJ, Aanensen DM, et al. Molecular tracing of the emergence, adaptation, and transmission of hospital-associated methicillin-resistant Staphylococcus aureus [Internet]. Vol. 109, Proceedings of the National Academy of Sciences. 2012. p. 9107–12.

16. Mentasti M, Cassier P, David S, Ginevra C, Gomez-Valero L, Underwood A, et al. Rapid detection and evolutionary analysis of Legionella pneumophila serogroup 1 sequence type 47. Clin Microbiol Infect. 2017 Apr;23(4):264.e1–264.e9.

17. Ellington MJ, Heinz E, Wailan AM, Dorman MJ, de Goffau M, Cain AK, et al. Contrasting patterns of longitudinal population dynamics and antimicrobial resistance mechanisms in two priority bacterial pathogens over 7 years in a single center. Genome Biol. 2019 Sep 2;20(1):184.

18. Alkan C, Sajjadian S, Eichler EE. Limitations of next-generation genome sequence assembly. Nat Methods. 2011 Jan;8(1):61–5.

19. Landan G, Graur D. Characterization of pairwise and multiple sequence alignment errors. Vol. 441, Gene. 2009. p. 141–7.

20. Farrer RA, Henk DA, MacLean D, Studholme DJ, Fisher MC. Using false discovery rates to benchmark SNP-callers in next-generation sequencing projects. Sci Rep. 2013;3:1512.

21. Hurgobin B, Edwards D. SNP Discovery Using a Pangenome: Has the Single Reference Approach Become Obsolete? Biology [Internet]. 2017 Mar 11;6(1). Available from: http://dx.doi.org/10.3390/biology6010021

22. Bertels F, Silander OK, Pachkov M, Rainey PB, van Nimwegen E. Automated reconstruction of whole-genome phylogenies from short-sequence reads. Mol Biol Evol. 2014 May;31(5):1077–88.

23. Pightling AW, Petronella N, Pagotto F. Choice of reference sequence and assembler for alignment of Listeria monocytogenes short-read sequence data greatly influences rates of error in SNP analyses. PLoS One. 2014 Aug 21;9(8):e104579.

24. Pightling AW, Petronella N, Pagotto F. Choice of reference-guided sequence assembler and SNP caller for analysis of Listeria monocytogenes short-read sequence data greatly influences rates of error. BMC Res Notes. 2015 Dec 8;8:748.

25. Lee RS, Behr MA. Does Choice Matter? Reference-Based Alignment for Molecular Epidemiology of Tuberculosis. J Clin Microbiol. 2016 Jul;54(7):1891–5.

26. Usongo V, Berry C, Yousfi K, Doualla-Bell F, Labbé G, Johnson R, et al. Impact of the choice of reference genome on the ability of the core genome SNV methodology to distinguish strains of Salmonella enterica serovar Heidelberg. PLoS One. 2018 Feb 5;13(2):e0192233.

27. Carroll LM, Wiedmann M, Mukherjee M, Nicholas DC, Mingle LA, Dumas NB, et al. Characterization of Emetic and Diarrheal Bacillus cereus Strains From a 2016 Foodborne Outbreak Using Whole-Genome Sequencing: Addressing the Microbiological, Epidemiological, and Bioinformatic Challenges. Vol. 10, Frontiers in Microbiology. 2019.

28. Bush SJ, Foster D, Eyre DW, Clark EL, De Maio N, Shaw LP, et al. Genomic diversity affects the accuracy of bacterial single-nucleotide polymorphism-calling pipelines. Gigascience [Internet]. 2020 Feb 1;9(2). Available from: http://dx.doi.org/10.1093/gigascience/giaa007

29. Gil N, Fiser A. The choice of sequence homologs included in multiple sequence alignments has a dramatic impact on evolutionary conservation analysis. Vol. 35, Bioinformatics. 2019. p. 12–9.

30. Tettelin H, Riley D, Cattuto C, Medini D. Comparative genomics: the bacterial pan-genome. Vol. 11, Current Opinion in Microbiology. 2008. p. 472–7.

31. Dos Vultos T, Mestre O, Rauzier J, Golec M, Rastogi N, Rasolofo V, et al. Evolution and diversity of clonal bacteria: the paradigm of Mycobacterium tuberculosis. PLoS One. 2008 Feb 6;3(2):e1538.

32. Lee RS, Proulx J-F, McIntosh F, Behr MA, Hanage WP. Previously undetected super-spreading of Mycobacterium tuberculosis revealed by deep sequencing [Internet]. Vol. 9, eLife. 2020. Available from: http://dx.doi.org/10.7554/elife.53245

33. Silby MW, Winstanley C, Godfrey SAC, Levy SB, Jackson RW. Pseudomonas genomes: diverse and adaptable. FEMS Microbiol Rev. 2011 Jul;35(4):652–80.

34. Hanage WP. Fuzzy species revisited. BMC Biol. 2013 Apr 15;11:41.

35. David S, Sánchez-Busó L, Harris SR, Marttinen P, Rusniok C, Buchrieser C, et al. Dynamics and impact of homologous recombination on the evolution of Legionella pneumophila [Internet]. Vol. 13, PLOS Genetics. 2017. p. e1006855. Available from: http://dx.doi.org/10.1371/journal.pgen.1006855

36. Koren S, Phillippy AM. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. Curr Opin Microbiol. 2015 Feb;23:110–20.

37. Bryant JM, Grogono DM, Greaves D, Foweraker J, Roddick I, Inns T, et al. Whole-genome sequencing to identify transmission of Mycobacterium abscessus between patients with cystic fibrosis: a retrospective cohort study. Lancet. 2013 May 4;381(9877):1551–60.

38. Holt KE, Wertheim H, Zadoks RN, Baker S, Whitehouse CA, Dance D, et al. Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in Klebsiella pneumoniae, an urgent threat to public health. Proc Natl Acad Sci U S A. 2015 Jul 7;112(27):E3574–81.

39. D'Auria G, Jiménez-Hernández N, Peris-Bondia F, Moya A, Latorre A. Legionella pneumophila pangenome reveals strain-specific virulence factors. Vol. 11, BMC Genomics. 2010. p. 181.

40. Freschi L, Vincent AT, Jeukens J, Emond-Rheault J-G, Kukavica-Ibrulj I, Dupont M-J, et al. The Pseudomonas aeruginosa Pan-Genome Provides New Insights on Its Population Structure, Horizontal Gene Transfer, and Pathogenicity. Genome Biol Evol. 2019 Jan 1;11(1):109–20.

41. Abreo E, Altier N. Pangenome of Serratia marcescens strains from nosocomial and environmental origins reveals different populations and the links between them. Sci Rep. 2019 Jan 10;9(1):46.

42. Kwong JC, Mercoulia K, Tomita T, Easton M, Li HY, Bulach DM, et al. Prospective Whole-Genome Sequencing Enhances National Surveillance of Listeria monocytogenes. J Clin Microbiol. 2016 Feb;54(2):333–42.

43. Gopalakrishnan S, Samaniego Castruita JA, Sinding M-HS, Kuderna LFK, Räikkönen J, Petersen B, et al. The wolf reference genome sequence (Canis lupus lupus) and its implications

for Canis spp. population genomics [Internet]. Vol. 18, BMC Genomics. 2017. Available from: http://dx.doi.org/10.1186/s12864-017-3883-3

44. Wu X, Heffelfinger C, Zhao H, Dellaporta SL. Benchmarking variant identification tools for plant diversity discovery. BMC Genomics. 2019 Sep 9;20(1):701.

45. Yang X, Lee W-P, Ye K, Lee C. One reference genome is not enough [Internet]. Vol. 20, Genome Biology. 2019. Available from: http://dx.doi.org/10.1186/s13059-019-1717-0

46. Leekitcharoenphon P, Nielsen EM, Kaas RS, Lund O, Aarestrup FM. Evaluation of whole genome sequencing for outbreak detection of Salmonella enterica. PLoS One. 2014 Feb 4;9(2):e87991.

47. Olson ND, Lund SP, Colman RE, Foster JT, Sahl JW, Schupp JM, et al. Best practices for evaluating single nucleotide variant calling methods for microbial genomics. Front Genet. 2015 Jul 7;6:235.

48. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Vol. 12, Nature Reviews Genetics. 2011. p. 443–51.

49. Petkau A, Mabon P, Sieffert C, Knox NC, Cabral J, Iskander M, et al. SNVPhyl: a single nucleotide variant phylogenomics pipeline for microbial genomic epidemiology. Microb Genom. 2017 Jun 30;3(6):e000116.

50. Hwang S, Kim E, Lee I, Marcotte EM. Systematic comparison of variant calling pipelines using gold standard personal exome variants. Sci Rep. 2015 Dec 7;5:17875.

51. Li H. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics. 2014 Oct 15;30(20):2843–51.

52. Liu X, Han S, Wang Z, Gelernter J, Yang B-Z. Variant Callers for Next-Generation Sequencing Data: A Comparison Study [Internet]. Vol. 8, PLoS ONE. 2013. p. e75619. Available from: http://dx.doi.org/10.1371/journal.pone.0075619

53. Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, Efremova M, et al. A survey of tools for variant analysis of next-generation genome sequencing data. Brief Bioinform. 2014 Mar;15(2):256–78.

54. Yu X, Sun S. Comparing a few SNP calling algorithms using low-coverage sequencing data. BMC Bioinformatics. 2013 Sep 17;14:274.

55. Jajou R, de Neeling A, van Hunen R, de Vries G, Schimmel H, Mulder A, et al. Epidemiological links between tuberculosis cases identified twice as efficiently by whole genome sequencing than conventional molecular typing: A population-based study [Internet]. Vol. 13, PLOS ONE. 2018. p. e0195413. Available from: http://dx.doi.org/10.1371/journal.pone.0195413

56. Coscollá M, Comas I, González-Candelas F. Quantifying nonvertical inheritance in the evolution of Legionella pneumophila. Mol Biol Evol. 2011 Feb;28(2):985–1001.

57. Valenzuela D, Norri T, Välimäki N, Pitkänen E, Mäkinen V. Towards pan-genome read alignment to improve variation calling. BMC Genomics. 2018 May 9;19(Suppl 2):87.

58. Computational Pan-Genomics Consortium. Computational pan-genomics: status, promises and challenges. Brief Bioinform. 2018 Jan 1;19(1):118–35.

59. Jandrasits C, Kröger S, Haas W, Renard BY. Computational pan-genome mapping and pairwise SNP-distance improve detection of Mycobacterium tuberculosis transmission clusters. PLoS

Comput Biol. 2019 Dec 9;15(12):e1007527.

60. Chen N-C, Solomon B, Mun T, Iyer S, Langmead B. Reducing reference bias using multiple population reference genomes. BioRxiv:2020.03.03.975219 [Preprint]. 2020 [cited 2010 March 21]. Available from: http://dx.doi.org/10.1101/2020.03.03.975219

61. Hedge J, Wilson DJ. Bacterial Phylogenetic Reconstruction from Whole Genomes Is Robust to Recombination but Demographic Inference Is Not [Internet]. Vol. 5, mBio. 2014. Available from: http://dx.doi.org/10.1128/mbio.02158-14

62. Duchêne S, Holt KE, Weill F-X, Le Hello S, Hawkey J, Edwards DJ, et al. Genome-scale rates of evolutionary change in bacteria. Microb Genom. 2016 Nov;2(11):e000094.

63. Didelot X, Maiden MCJ. Impact of recombination on bacterial evolution. Trends Microbiol. 2010 Jul;18(7):315–22.

64. von Wintersdorff CJH, Penders J, van Niekerk JM, Mills ND, Majumder S, van Alphen LB, et al. Dissemination of Antimicrobial Resistance in Microbial Ecosystems through Horizontal Gene Transfer. Front Microbiol. 2016 Feb 19;7:173.

65. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Ostell J, Pruitt KD, et al. GenBank. Nucleic Acids Res. 2018 Jan 4;46(D1):D41–7.

66. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014 Jul 15;30(14):2068–9.

67. Lechner M, Findeiss S, Steiner L, Marz M, Stadler PF, Prohaska SJ. Proteinortho: detection of (co-)orthologs in large-scale analysis. BMC Bioinformatics. 2011 Apr 28;12:124.

68. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013 Apr;30(4):772–80.

69. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015 Jan;32(1):268–74.

70. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol Biol Evol. 2018 Feb 1;35(2):518–22.

71. Darling AE, Mau B, Perna NT. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One. 2010 Jun 25;5(6):e11147.

72. Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol. 2007 Jan;57(Pt 1):81–91.

73. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun. 2018 Nov 30;9(1):5114.

74. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST : architecture and applications [Internet]. Vol. 10, BMC Bioinformatics. 2009. p. 421. Available from: http://dx.doi.org/10.1186/1471-2105-10-421

75. Fry NK, Bangsborg JM, Bergmans A, Bernander S, Etienne J, Franzin L, et al. Designation of the European Working Group on Legionella Infection (EWGLI) Amplified Fragment Length Polymorphism Types of Legionella pneumophila Serogroup 1 and Results of Intercentre

Proficiency Testing Using a Standard Protocol. Vol. 21, European Journal of Clinical Microbiology & Infectious Diseases. 2002. p. 722–8.

76. R Core Team. R: A language and environment for statistical computing [Internet]. R Foundation for Statistical Computing; 2018. Available from: https://www.R-project.org/

77. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016 Oct 1;32(19):3047–8.

78. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. Vol. 17, EMBnet.journal. 2011. p. 10. Available from: http://dx.doi.org/10.14806/ej.17.1.200

79. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. Bioinformatics. 2011 Mar 15;27(6):863–4.

80. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Vol. 25, Bioinformatics. 2009. p. 1754–60.

81. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078–9.

82. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Vol. 27, Bioinformatics. 2011. p. 2987–93.

83. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007 Aug;24(8):1586–91.

84. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 2000 Jun;16(6):276–7.

85. Hermann P, Heissl A, Tiemann-Boege I, Futschik A. LDJump: Estimating variable recombination rates from population genetic data. Mol Ecol Resour. 2019 May;19(3):623–38.

86. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer; 2016. Available from: https://ggplot2.tidyverse.org.

87. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 2019 Jul 2;47(W1):W256–9.

88. Strimmer K, Rambaut A. Inferring confidence sets of possibly misspecified gene trees. Proc Biol Sci. 2002 Jan 22;269(1487):137–42.

89. Bogdanowicz D, Giaro K, Wróbel B. TreeCmp: Comparison of Trees in Polynomial Time [Internet]. Vol. 8, Evolutionary Bioinformatics. 2012. p. EBO.S9657. Available from: http://dx.doi.org/10.4137/ebo.s9657

90. Robinson DF, Foulds LR. Comparison of phylogenetic trees. Vol. 53, Mathematical Biosciences. 1981. p. 131–47.

91. Kruskal WH, Allen Wallis W. Use of Ranks in One-Criterion Variance Analysis. Vol. 47, Journal of the American Statistical Association. 1952. p. 583.

92. Rey D, Neuhäuser M. Wilcoxon-Signed-Rank Test. International Encyclopedia of Statistical Science. 2011. p. 1658–9.

93. Massey FJ. The Kolmogorov-Smirnov Test for Goodness of Fit. Vol. 46, Journal of the

American Statistical Association. 1951. p. 68–78.

LEGENDS TO FIGURES

Fig 1. Distribution of proportion of mapped reads depending on reference choice.

Fig 2. Distribution of coverage of the reference genome depending on reference choice.

Fig 3. Distribution of the average depth depending on reference choice.

Fig 4. Distribution of the number of SNPs depending on reference choice.

Fig 5. Comparison of Robinson-Foulds (RF) and Matching Clusters (MC) normalized distances calculated between trees from the same species.

Fig 6. Comparison of RF distances against ANI calculated between the reference genomes selected for each species.

Fig 7. Impact of reference choice on phylogenetic trees of *L. pneumophila*. ML trees included the selected reference sequences of *L. pneumophila* and the consensus sequences obtained from mappings against strains (A) Philadelphia 1, (B) Paris, (C) Alcoy and (D) Lansing 3. Clusters of isolates related with references Paris (red) and Alcoy (blue) are coloured in the first three phylogenies. Isolates 28HGV and 91HGV (highlighted in yellow) were placed in different clades in the trees when using references Paris and Alcoy. Clade of references resulting from using Lansing 3 as reference genome is coloured in red.

Fig 8. Impact of reference choice on phylogenetic trees of *K. pneumoniae*. ML trees included the selected reference sequences from *K. pneumoniae* and the consensus sequences obtained from mappings against strains (A) HS11286 and (B) NTUH-K2044. Isolates HGV2C-06 and HCV1-10 (yellow) changed their placement depending on reference choice.

Fig 9. Impact of reference choice on phylogenetic trees of *P. aeruginosa*. ML trees included the selected reference sequences of *P. aeruginosa* and the consensus sequences obtained from mappings against strains (A) M18 and (B) 12939. Reference M18 and isolate P5M1 (yellow) alter their phylogenetic relationships depending on reference choice.

Fig 10. Impact of reference choice on phylogenetic trees of *S.marcescens*. ML trees included the selected reference sequences from *S. marcescens* and the consensus sequences calculated from alignments against strains (A) UMH9 and (B) WW4. Outbreak clade is shown in red.

Fig 11. Recombination rate distribution depending on reference choice between 'core' MSAs

including sequences from *N. gonorrhoeae*.

Fig 12. Distribution of d*N*/d*S* depending on reference choice.

Fig 13. Overview of the workflow used.

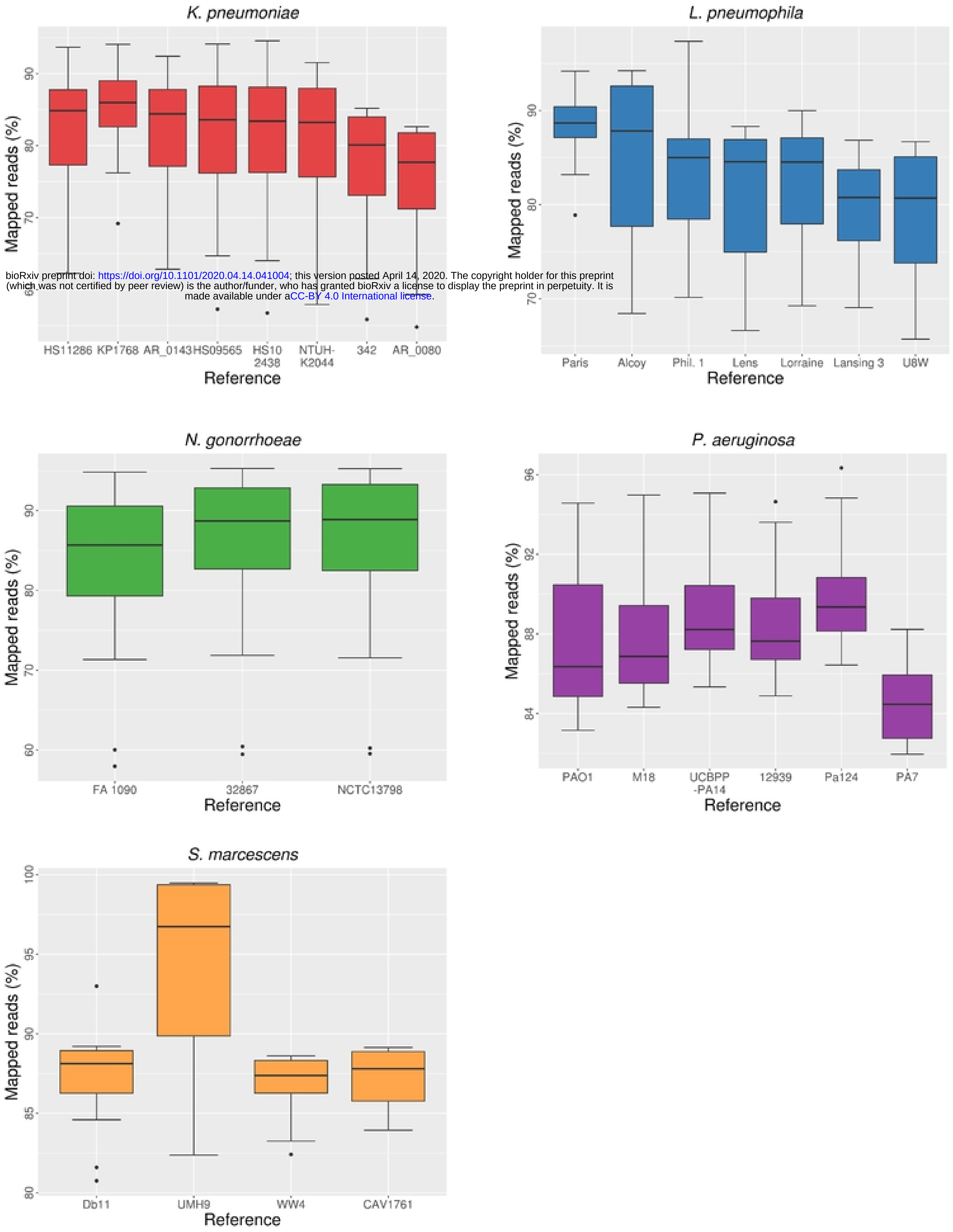

K. pneumoniae
Mapped reads (%)
60
70
80
90
HS11286
KP1768
AR_0143
HS09565
HS10 2438
NTUH-K2044
342
AR_0080
Reference
L. pneumophila
Mapped reads (%)
70
80
90
Paris
Alcoy
Phil. 1
Lens
Lorraine
Lansing 3
U8W
Reference
N. gonorrhoeae
Mapped reads (%)
60
70
80
90
FA 1090
32867
NCTC13798
Reference
P. aeruginosa
Mapped reads (%)
84
88
92
96
PAO1
M18
UCBPP -PA14
12939
Pa124
PA7
Reference
S. marcescens
Mapped reads (%)
80
85
90
95
100
Db11
UMH9
WW4
CAV1761
Reference

Fig1

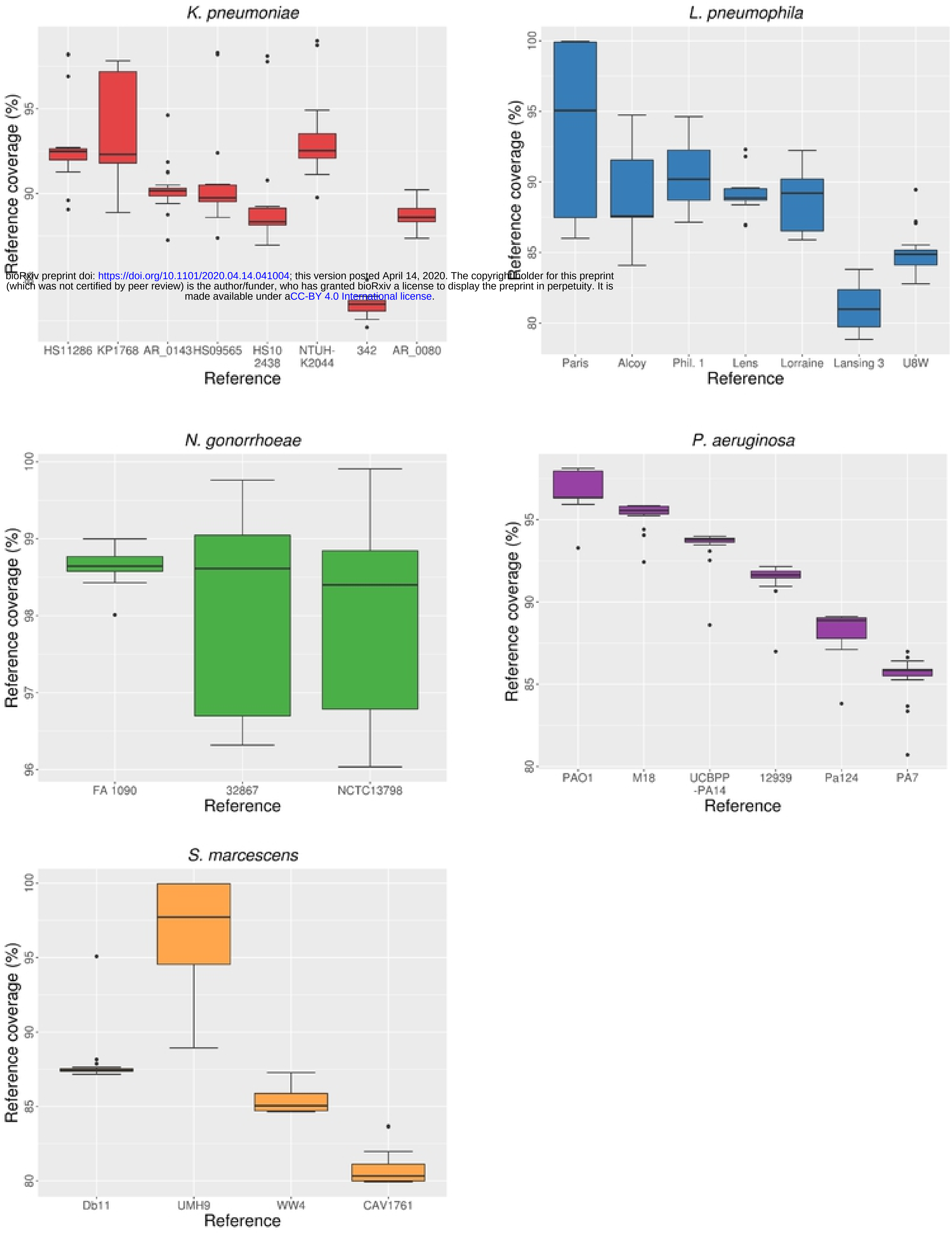

K. pneumoniae
Reference coverage (%)
95
90
85
HS11286
KP1768
AR_0143
HS09565
HS10 2438
NTUH-K2044
342
AR_0080
Reference
L. pneumophila
Reference coverage (%)
100
95
90
85
80
Paris
Alcoy
Phil. 1
Lens
Lorraine
Lansing 3
U8W
Reference
N. gonorrhoeae
Reference coverage (%)
100
99
98
97
96
FA 1090
32867
NCTC13798
Reference
P. aeruginosa
Reference coverage (%)
95
90
85
80
PAO1
M18
UCBPP-PA14
12939
Pa124
PA7
Reference
S. marcescens
Reference coverage (%)
100
95
90
85
80
Db11
UMH9
WW4
CAV1761
Reference

Fig2

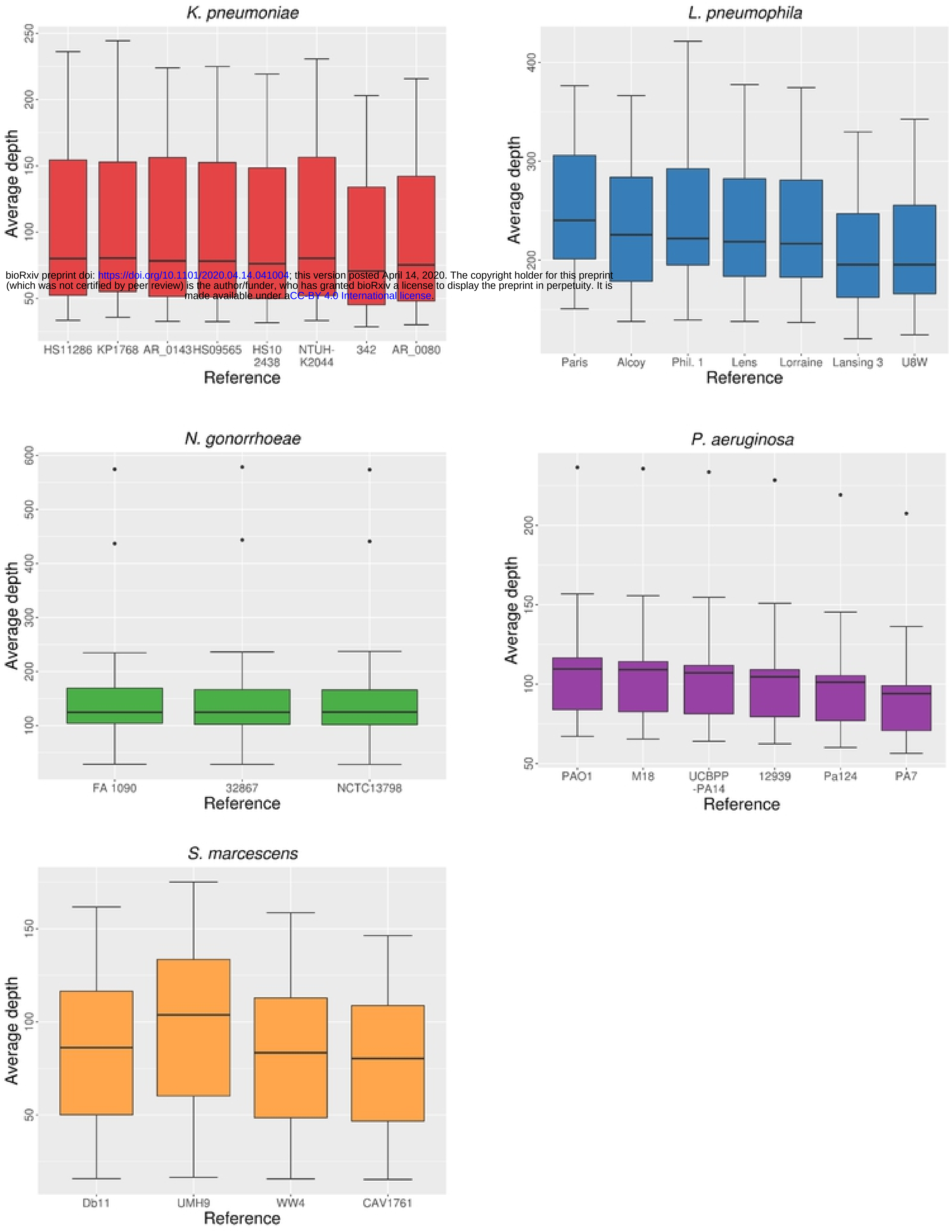

K. pneumoniae
Average depth
250
200
150
100
50
HS11286
KP1768
AR_0143
HS09565
HS10 2438
NTUH-K2044
342
AR_0080
Reference
L. pneumophila
Average depth
400
300
200
Paris
Alcoy
Phil. 1
Lens
Lorraine
Lansing 3
U8W
Reference
N. gonorrhoeae
Average depth
600
500
400
300
200
100
FA 1090
32867
NCTC13798
Reference
P. aeruginosa
Average depth
200
150
100
50
PAO1
M18
UCBPP -PA14
12939
Pa124
PA7
Reference
S. marcescens
Average depth
150
100
50
Db11
UMH9
WW4
CAV1761
Reference

Fig3

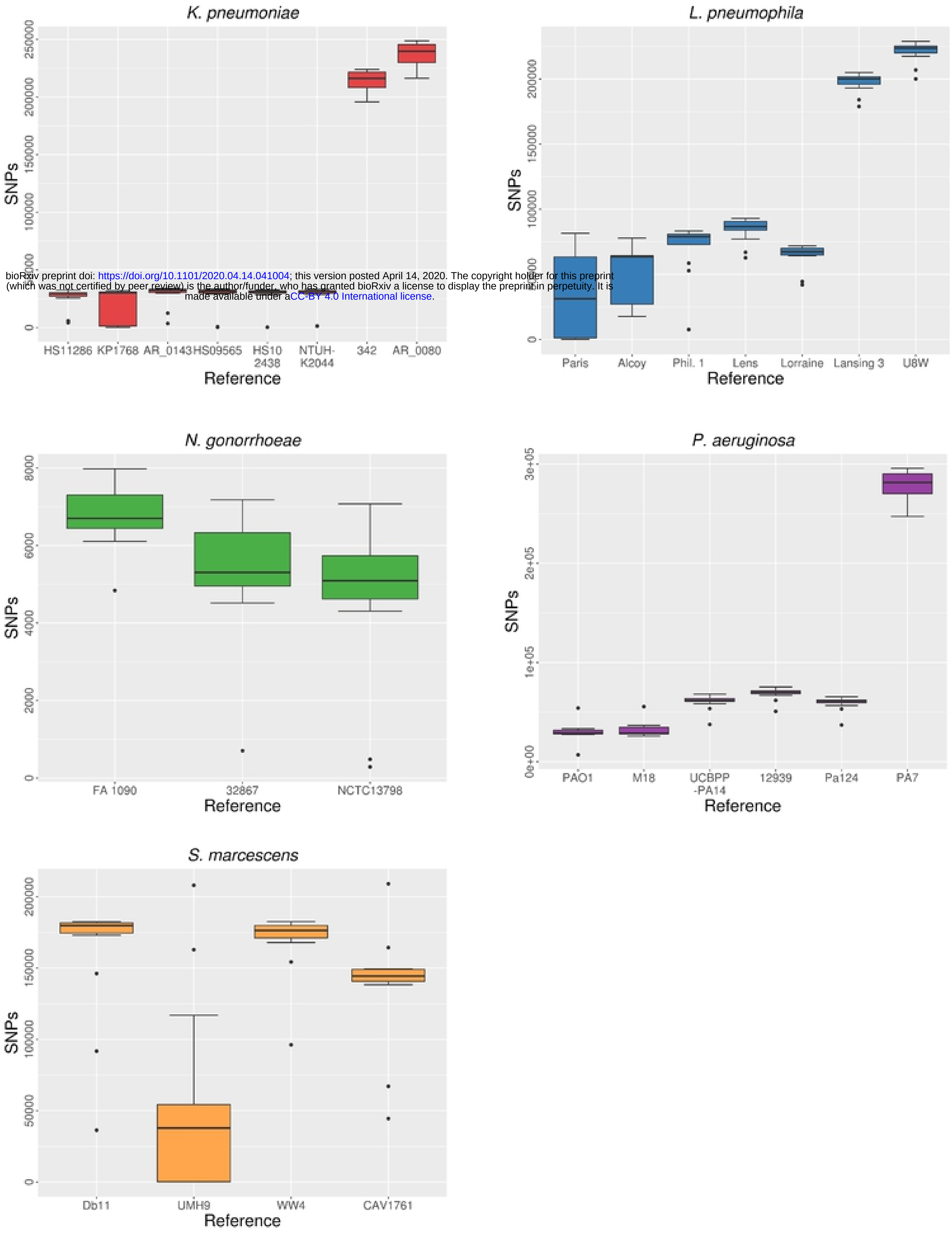

K. pneumoniae
SNPs
250000
200000
150000
100000
0
HS11286
KP1768
AR_0143
HS09565
HS10 2438
NTUH-K2044
342
AR_0080
Reference
L. pneumophila
SNPs
200000
150000
100000
0
Paris
Alcoy
Phil. 1
Lens
Lorraine
Lansing 3
U8W
Reference
N. gonorrhoeae
SNPs
8000
6000
4000
2000
0
FA 1090
32867
NCTC13798
Reference
P. aeruginosa
SNPs
3e+05
2e+05
1e+05
0e+00
PAO1
M18
UCBPP -PA14
12939
Pa124
PA7
Reference
S. marcescens
SNPs
200000
150000
100000
50000
0
Db11
UMH9
WW4
CAV1761
Reference

Fig4

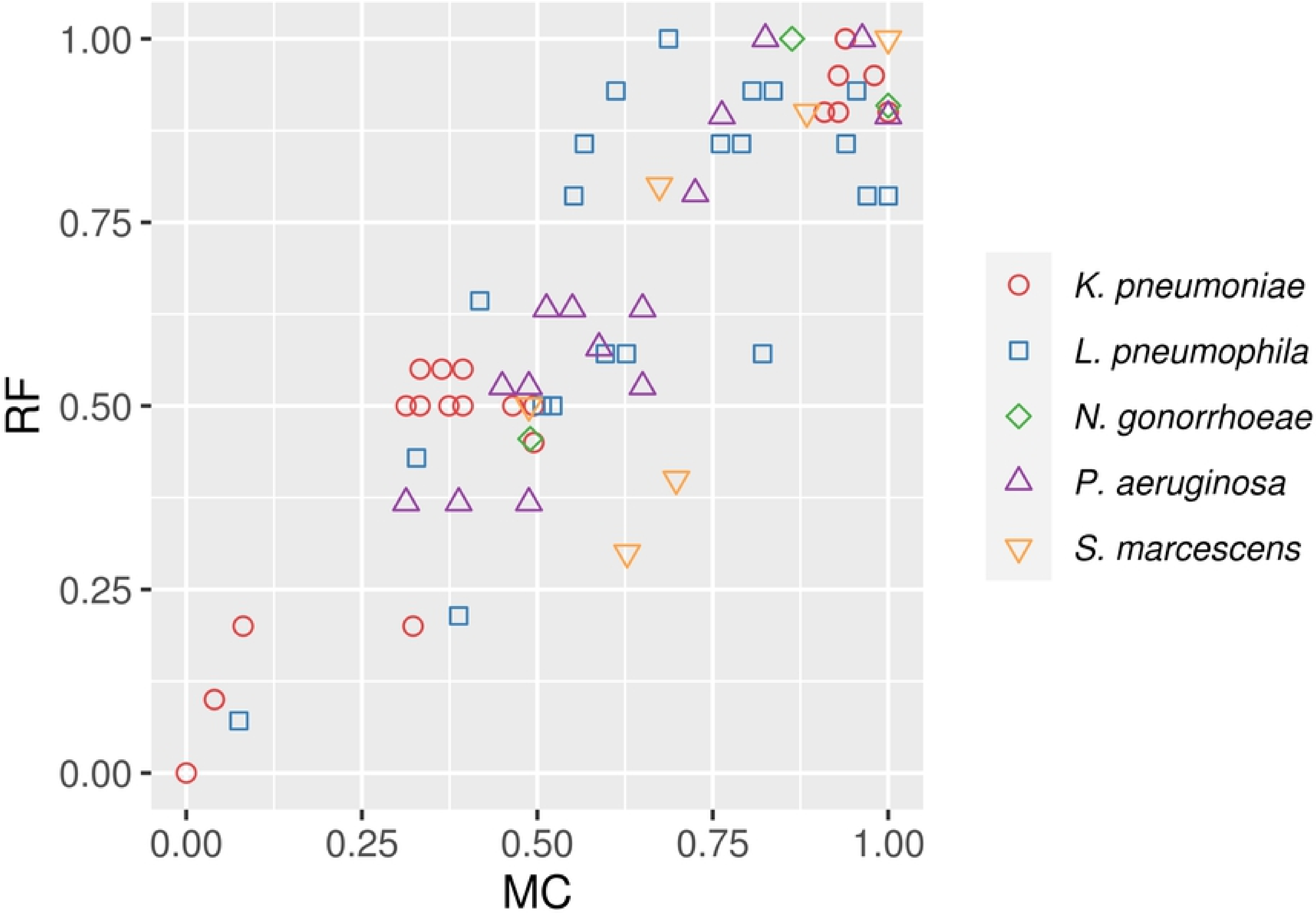
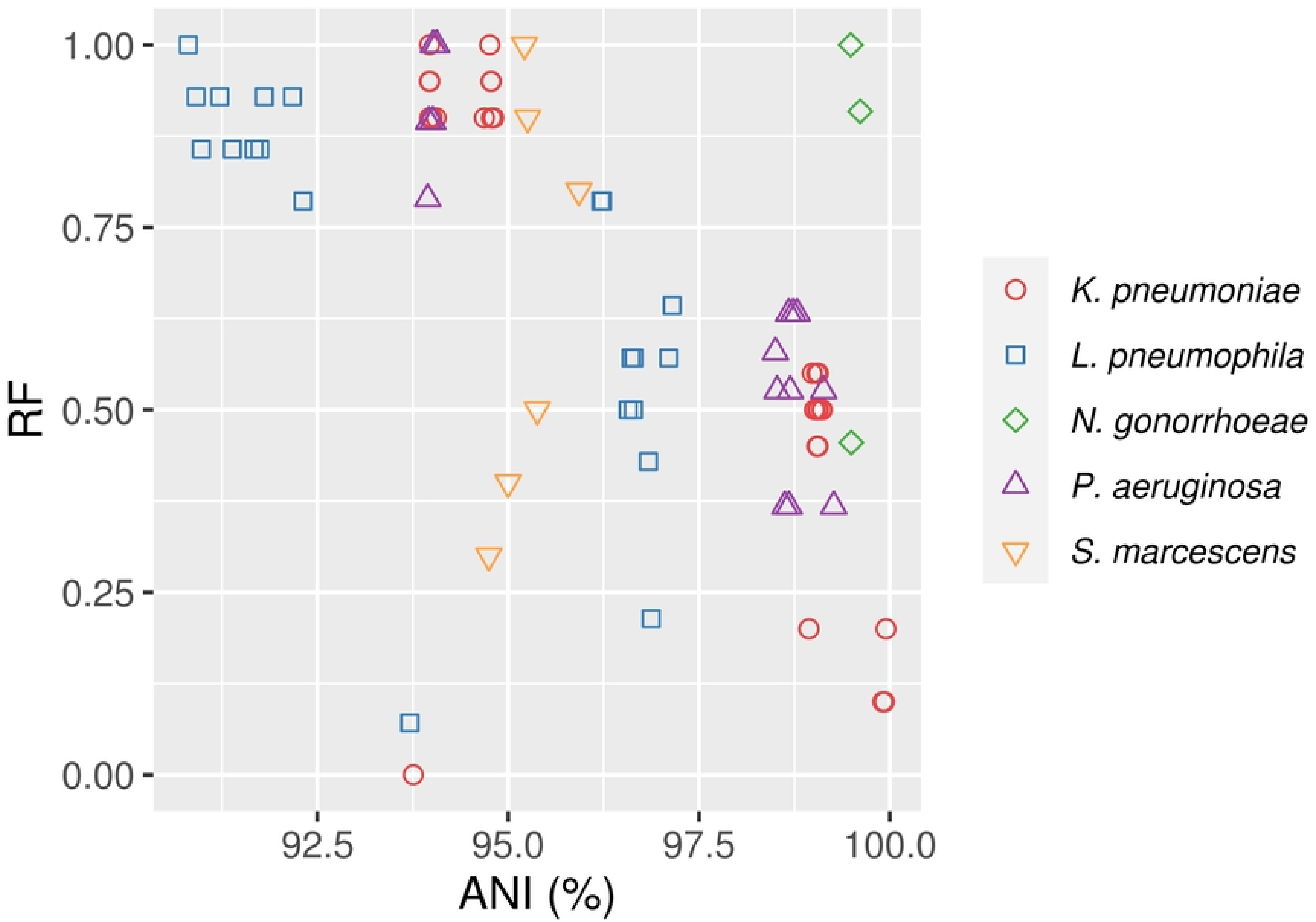

RF
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
MC
○ K. pneumoniae
□ L. pneumophila
◇ N. gonorrhoeae
△ P. aeruginosa
▽ S. marcescens

Fig5



RF
1.00
0.75
0.50
0.25
0.00
92.5
95.0
97.5
100.0
ANI (%)
○ K. pneumoniae
□ L. pneumophila
◇ N. gonorrhoeae
△ P. aeruginosa
▽ S. marcescens

Fig6

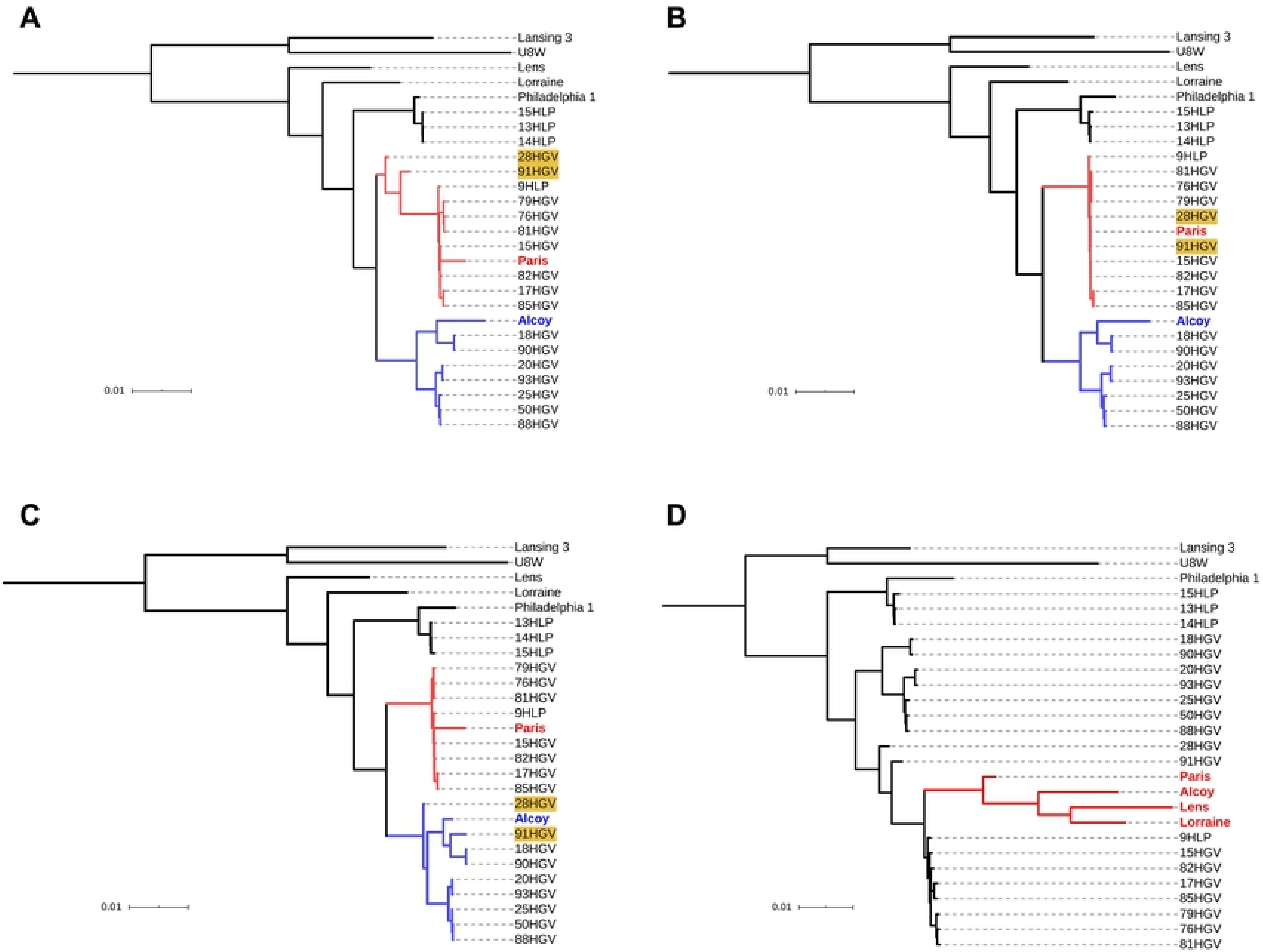

A
Lansing 3
U8W
Lens
Lorraine
Philadelphia 1
15HLP
13HLP
14HLP
28HGV
91HGV
9HLP
79HGV
76HGV
81HGV
15HGV
Paris
82HGV
17HGV
85HGV
Alcoy
18HGV
90HGV
20HGV
93HGV
25HGV
50HGV
88HGV
0.01
B
Lansing 3
U8W
Lens
Lorraine
Philadelphia 1
15HLP
13HLP
14HLP
9HLP
81HGV
76HGV
79HGV
28HGV
Paris
91HGV
15HGV
82HGV
17HGV
85HGV
Alcoy
18HGV
90HGV
20HGV
93HGV
25HGV
50HGV
88HGV
0.01
C
Lansing 3
U8W
Lens
Lorraine
Philadelphia 1
13HLP
14HLP
15HLP
79HGV
76HGV
81HGV
9HLP
Paris
15HGV
82HGV
17HGV
85HGV
28HGV
Alcoy
91HGV
18HGV
90HGV
20HGV
93HGV
25HGV
50HGV
88HGV
0.01
D
Lansing 3
U8W
Philadelphia 1
15HLP
13HLP
14HLP
18HGV
90HGV
20HGV
93HGV
25HGV
50HGV
88HGV
28HGV
91HGV
Paris
Alcoy
Lens
Lorraine
9HLP
15HGV
82HGV
17HGV
85HGV
79HGV
76HGV
81HGV
0.01

Fig7

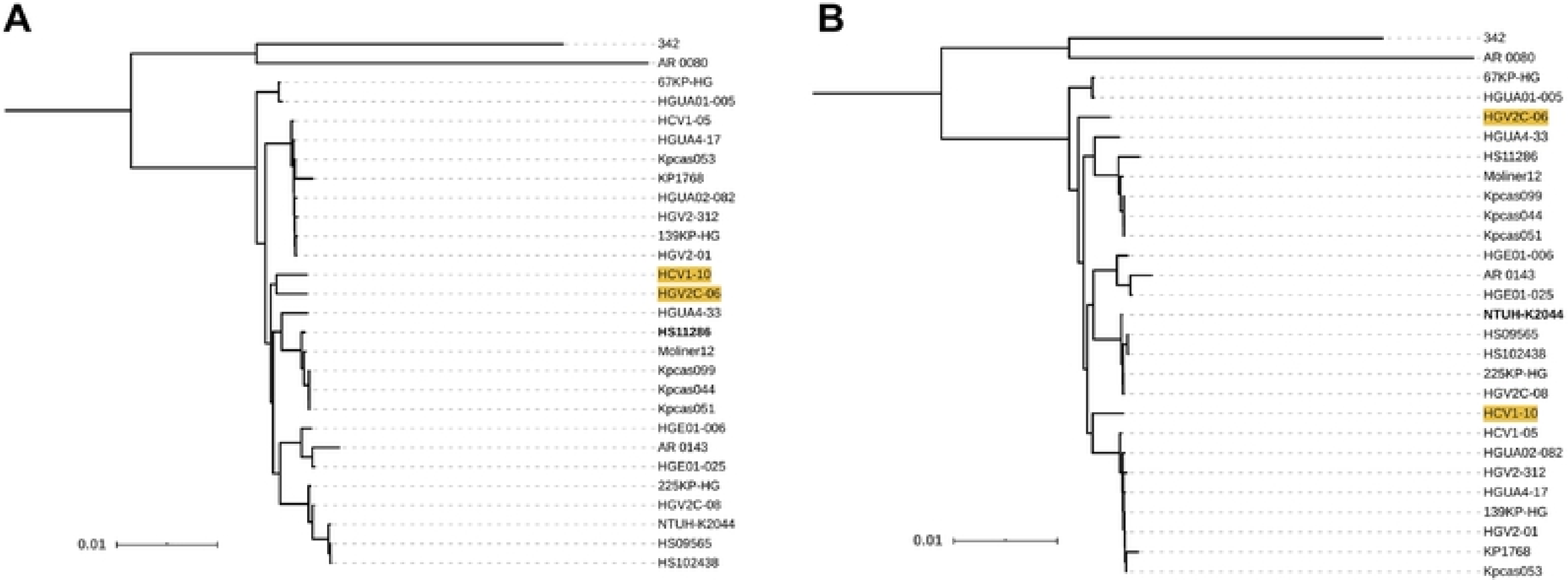

A
342
AR 0080
67KP-HG
HGUA01-005
HCV1-05
HGUA4-17
Kpcas053
KP1768
HGUA02-082
HGV2-312
139KP-HG
HGV2-01
HCV1-10
HGV2C-06
HGUA4-33
HS11286
Moliner12
Kpcas099
Kpcas044
Kpcas051
HGE01-006
AR 0143
HGE01-025
225KP-HG
HGV2C-08
NTUH-K2044
HS09565
HS102438
0.01
B
342
AR 0080
67KP-HG
HGUA01-005
HGV2C-06
HGUA4-33
HS11286
Moliner12
Kpcas099
Kpcas044
Kpcas051
HGE01-006
AR 0143
HGE01-025
NTUH-K2044
HS09565
HS102438
225KP-HG
HGV2C-08
HCV1-10
HCV1-05
HGUA02-082
HGV2-312
HGUA4-17
139KP-HG
HGV2-01
KP1768
Kpcas053
0.01

Fig8

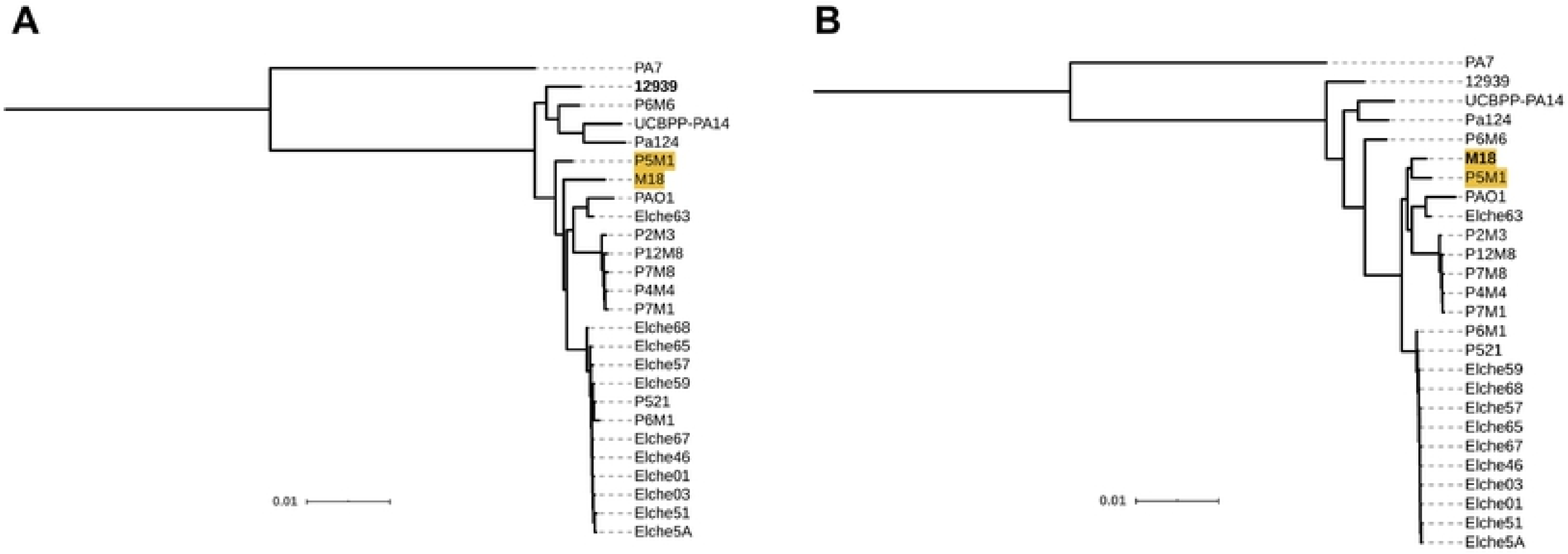

A
PA7
12939
P6M6
UCBPP-PA14
Pa124
P5M1
M18
PAO1
Elche63
P2M3
P12M8
P7M8
P4M4
P7M1
Elche68
Elche65
Elche57
Elche59
P521
P6M1
Elche67
Elche46
Elche01
Elche03
Elche51
Elche5A
0.01
B
PA7
12939
UCBPP-PA14
Pa124
P6M6
M18
P5M1
PAO1
Elche63
P2M3
P12M8
P7M8
P4M4
P7M1
P6M1
P521
Elche59
Elche68
Elche57
Elche65
Elche67
Elche46
Elche03
Elche01
Elche51
Elche5A
0.01

Fig9

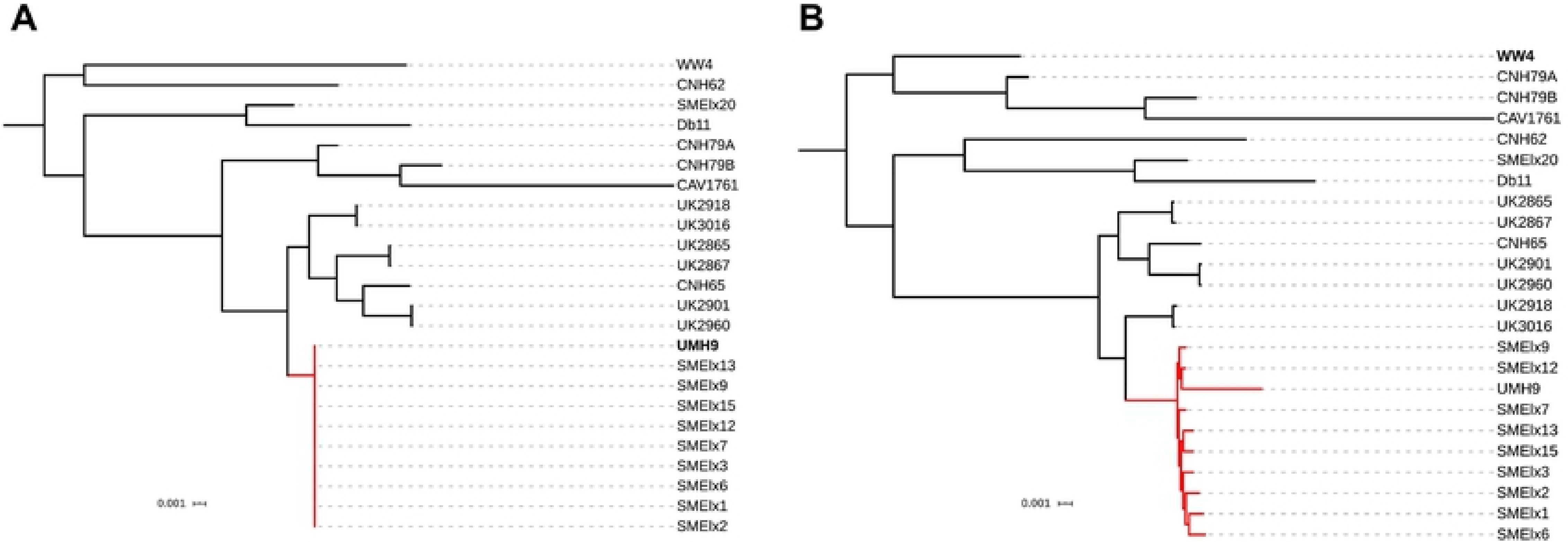

A
WW4
CNH62
SMEIx20
Db11
CNH79A
CNH79B
CAV1761
UK2918
UK3016
UK2865
UK2867
CNH65
UK2901
UK2960
UMH9
SMEIx13
SMEIx9
SMEIx15
SMEIx12
SMEIx7
SMEIx3
SMEIx6
SMEIx1
SMEIx2
0.001
B
WW4
CNH79A
CNH79B
CAV1761
CNH62
SMEIx20
Db11
UK2865
UK2867
CNH65
UK2901
UK2960
UK2918
UK3016
SMEIx9
SMEIx12
UMH9
SMEIx7
SMEIx13
SMEIx15
SMEIx3
SMEIx2
SMEIx1
SMEIx6
0.001

Fig10

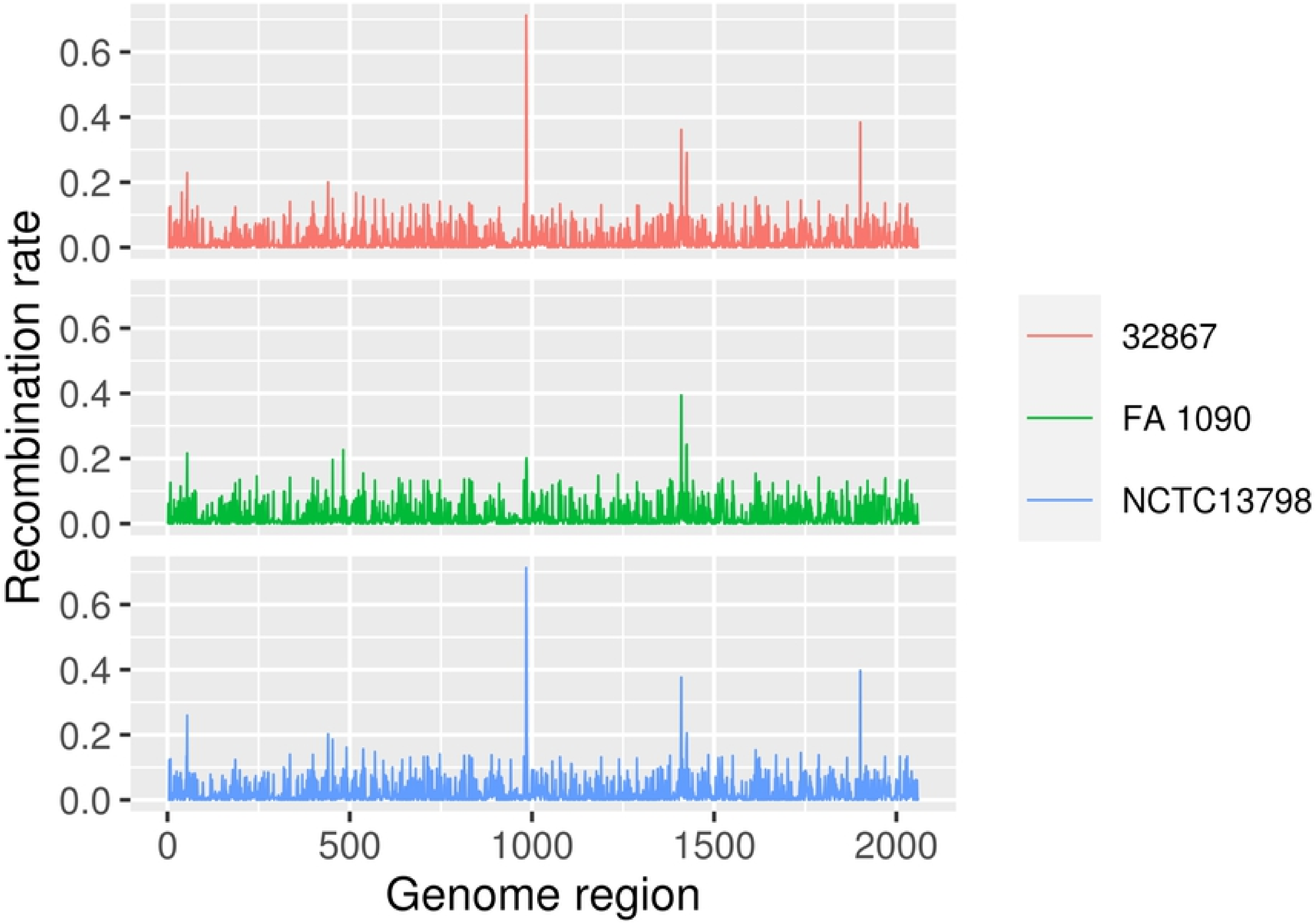

Recombination rate
0.6
0.4
0.2
0.0
0.6
0.4
0.2
0.0
0.6
0.4
0.2
0.0
0
500
1000
1500
2000
Genome region
32867
FA 1090
NCTC13798

Fig11

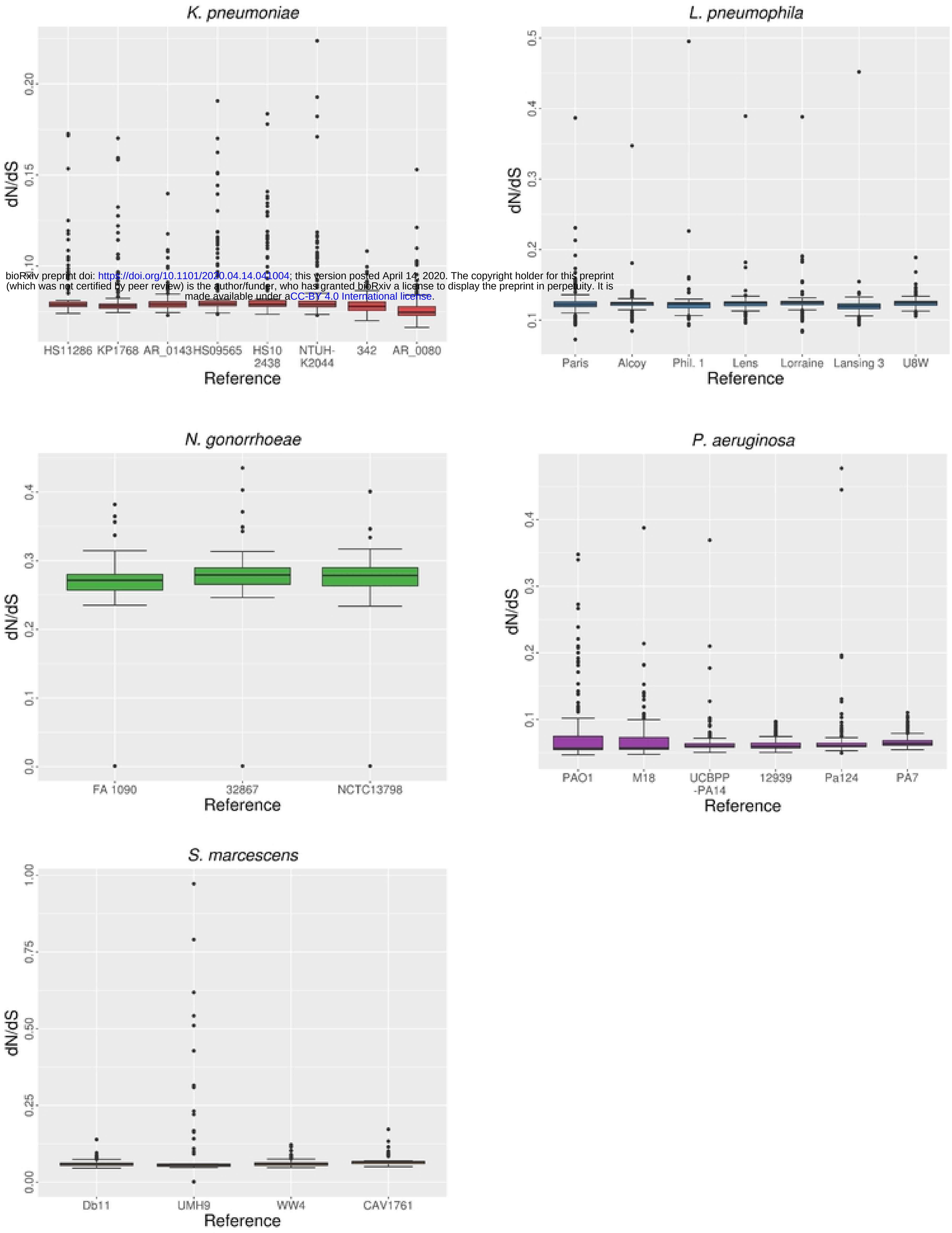

K. pneumoniae
dN/dS
0.10
0.15
0.20
HS11286
KP1768
AR_0143
HS09565
HS10 2438
NTUH-K2044
342
AR_0080
Reference
L. pneumophila
dN/dS
0.1
0.2
0.3
0.4
0.5
Paris
Alcoy
Phil. 1
Lens
Lorraine
Lansing 3
U8W
Reference
N. gonorrhoeae
dN/dS
0.0
0.1
0.2
0.3
0.4
FA 1090
32867
NCTC13798
Reference
P. aeruginosa
dN/dS
0.1
0.2
0.3
0.4
PAO1
M18
UCBPP-PA14
12939
Pa124
PA7
Reference
S. marcescens
dN/dS
0.00
0.25
0.50
0.75
1.00
Db11
UMH9
WW4
CAV1761
Reference

Fig12

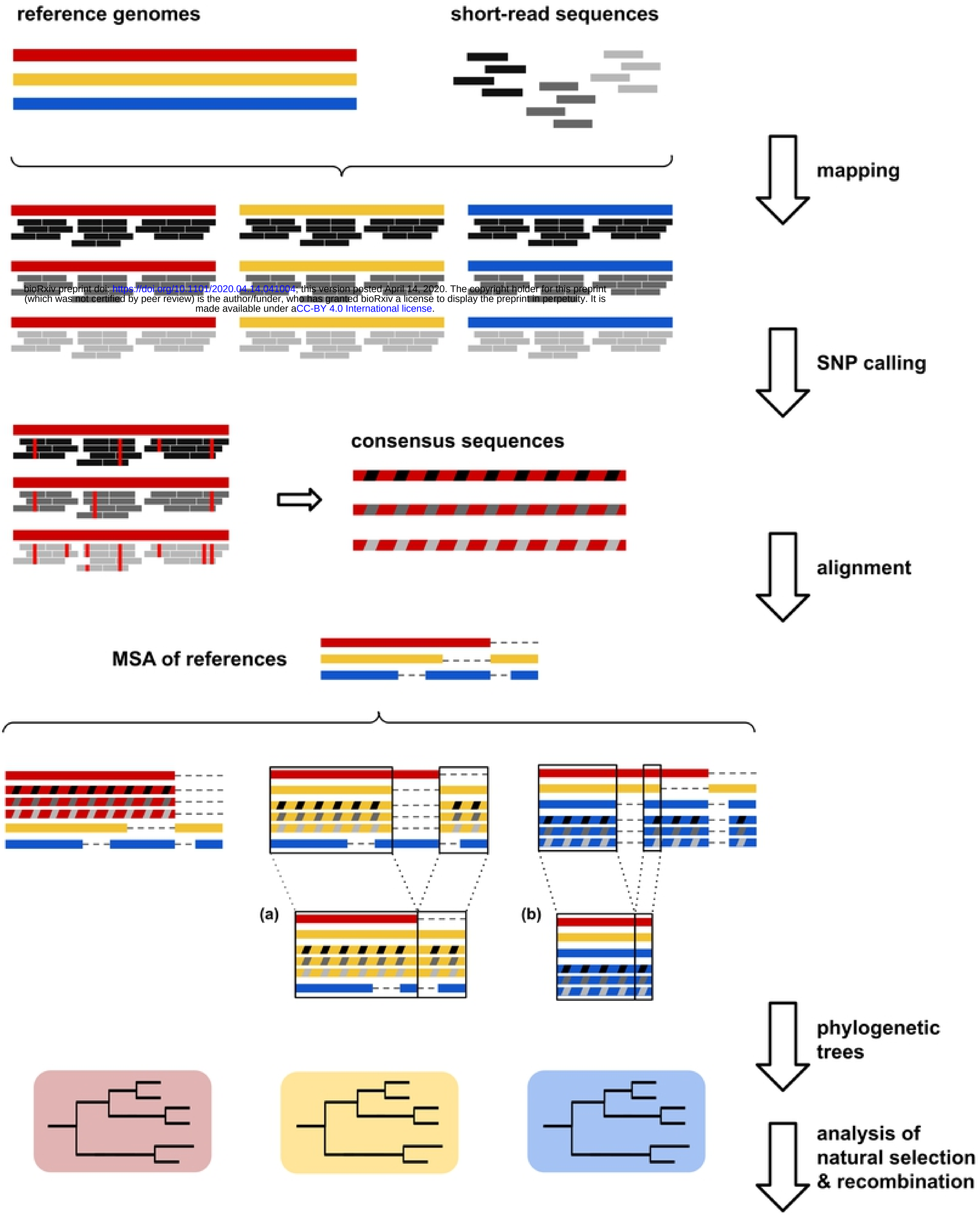

reference genomes
short-read sequences
mapping
SNP calling
consensus sequences
alignment
MSA of references
(a)
(b)
phylogenetic trees
analysis of natural selection & recombination

Fig13