Benchmarking for *Campylobacter coli* phylogeny

Report number	#6
Responsible	Anaïs Painset (PHE) and Timothy Dallman (PHE)
Other partners/institutions involved	APHA (United Kingdom), BfR (Germany), DTU (Denmark), IZSLT (Italy), IZSVe (Italy), NIPH-NIH (Poland), NVRI (Poland)
Benchmarking launched (date)	December 2017
Deliverable due (date)	January 2018

Purpose of the benchmarking exercise

The main purpose of this benchmarking exercise was to evaluate a number of available bioinformatics tools both to detect genomic variants and to build a phylogeny based on the variants detected for *Campylobacter coli* isolates. In this specific exercise, we ask participants to take into account the possibility of recombination between isolates. With the use of Whole Genome Sequencing, phylogenetic is used as a method to characterise microorganisms in outbreak investigations and for surveillance of isolates that are may be genetically related.

Participants

Participants in this benchmarking were institutions from the ENGAGE network.
Eleven sets of results were submitted from the following institutions:
APHA (United Kingdom), BfR (Germany), DTU (Denmark), IZSLT (Italy), IZSVe (3 phylogenies) (Italy), NIPH-NIH (Poland), NVRI (Poland), PHE (2 phylogenies) (United Kingdom).

Results from participating institutes are identified by codes (1-11, see below) and each code is known only by the corresponding laboratory. The full list of laboratory codes is known only by the organizers (PHE). Table 1 (below) described the methods/tools used to produce each phylogeny. Final phylogenies will be referred as Centre XX later in the document where XX correspond to the row number of Table 1.

Tools used by participants

Centre	SNP alignments tools (version) [parameters]	Tree building (version) [parameters]	Recombination detection (version) [parameters]
1	Snippy 3.0 [default]	Gubbins 2.1.0 [default]	Gubbins 2.1.0 [default] → post SNP detection
2	Snippy 3.2 [mapqual 60, basequal 20, mincov 10,minfrac 0.9] vcftools 0.1.15 [thin 100,recode]	FastTree 2.1.7 [-nt, Nucleotide distances: Jukes-Cantor, Joins: balanced, Support: SH-like 1000]	-
3	CSIPhylogeny 1.4 command line [default]	CSIPhylogeny 1.4 command line [default]	-

4	BWA Mem 0.7.12 [-p and default for other parameters] samtools 1.5 • view [-sb] • sort [default] • mpileup [-6 (Illumina +1.3), -C 50 (min quality 50), -v, -u] bcftools 1.5 • call [-O v, --ploidy 1, -v, -m] vcf_fa_extractor (https://github.com/moskalenko/vcf_fa_extractor) [default] clustalW built-in MEGA7 [default]	MEGA 7 [Statistical Method: Maximum likelihood, Model/Method: Jukes and Cantor, Rates among Sites: Uniforms, ML Heuristic method: NNI (Nearest-Neighbor-Interchange)	-
5	FastQC 0.11.2 [default] Kraken 0.10.6 [default] Trimmomatic0.32 [ILLUMINACLIP:Nextera-PE.fa:2:30:10 LEADING:20 TRAILING:20 SLIDINGWINDOW:4:20 MINLEN:100] PHEnix 1.3 – bwa/GATKbuilt-in [sample_ploidy: 1, genotype_likelihoods_model: SNP, rf: BadCigar, out_mode: EMIT_ALL_SITES, nt: 1, ad_ratio: 0.9, min_depth: 15, qual_score: 30, mq_score: 30]	RAxML 7.2.8 [-f a -x 12345 -p 12345 -# autoMRE -m GTRGAMMA]	-
6	FastQC 0.11.2 [default] Kraken 0.10.6 [default] Trimmomatic0.32 [ILLUMINACLIP:Nextera-PE.fa:2:30:10 LEADING:20 TRAILING:20 SLIDINGWINDOW:4:20 MINLEN:100] Snippy 3.2 [mincov 15, minqual 30, types snp]	RAxML 7.2.8 [-f a -x 12345 -p 12345 -# autoMRE -m GTRGAMMA]	-
7	FastQC 0.11.2 [default] Kraken 0.10.6 [default] Trimmomatic0.32 [ILLUMINACLIP:Nextera-PE.fa:2:30:10 LEADING:20 TRAILING:20 SLIDINGWINDOW:4:20 MINLEN:100] bwa-mem 0.7.12 [default] samtools 0.1.19-44428cd • mpileup [-I,-u] bcftools (part of samtools 0.1.19-44428cd) • view [-vcgI] vcfutils.pl varFilter [Q 25, d 30, w 10, W 15] Internal script for format conversion (vcf -> fasta)	RAxML 7.2.8 [-f a -x 12345 -p 12345 -# autoMRE -m GTRGAMMA]	-

8	CGE CSI phylogeny online tools 1.4 [default]	CGE CSI phylogeny online tools 1.4 [default]	-
9	BWA mem 0.7.13 [default] Freebayes 1.1.0 [ploidy 1] VariantAnnotation 1.22.3 [remove monomorpic (AC > 0 & AC<8), set calls below 10x DP to no-call, remove variants with missing GT for >2 samples and no alts, remove variants with MQM < 50] VCF-kit 0.1.2 • pheno fasta	VCF-kit 0.1.2 • pheno tree nj	-
10	PHEnix 1.2 – bwa/GATK built-in [ad_ratio: 0.9, min_depth: 10, qual_score: 30, mq_score: 30] SnapperDB 0.2.4 [a: A80, r: Y, ng: gubbins gff]	RAxML 8.2.8 [-N autoMRE -f a -p 12345 -x 12345 -m GTRCAT]	Gubbins 2.0.0 [c 16, u] recombination detected on WGS SNP alignment
11	PHEnix 1.2 – bwa/GATK built-in [ad_ratio: 0.9, min_depth: 10, qual_score: 30, mq_score: 30] SnapperDB 0.2.4 [a: A80, r: N, ng: gubbins gff]	RAxML 8.2.8 [-N autoMRE -f a -p 12345 -x 12345 -m GTRCAT]	Gubbins 2.0.0 [c 16, u] recombination detected on WGS SNP alignment

Table 1. List of tools/software used to produce each phylogeny submitted. Each row will be referred to as Centre XX

Species/genomes included

Public Health England selected and provided genomes from *Campylobacter coli* and part of the same sequence type complex ST-828 complex. This ST-complex is very diverse as it is one of the most common ST-complex found amongst *Campylobacter* isolates. The genomes were selected because they were part of a suspected outbreak investigated by PHE. The outbreak occurred in 2008 in the North of England where a teacher and students from a primary school were having gastrointestinal symptoms. A suspected tap water isolate was also collected. The isolates were sent for testing, all came back as *Campylobacter coli* and the same phage type PT44 as found. Retrospective WGS analysis shows that all the isolates were part of the same ST-complex: ST-828.

Nine genomes represented by sets of fastq (paired) were included the data set (Additional table with the list of selected genomes). All genomes were generated using an Illumina HiSeq and fastq provided to the partners were trimmed and assessed for quality before sharing. Fastq were trimmed using Trimmomatic 0.32 with the following options: ILLUMINACLIP:NexteraPE-PE.fa:2:30:10:8:true LEADING:30 TRAILING:30 SLIDINGWINDOW:10:20 MINLEN:50. Quality of the sequencing was assessed by running FastQC 0.11.3. The trimmed and quality assessed reads were used for the analysis (see Additional Table 1 under ‘Additional notes’, and Supplementary Table 6 (Annex F)).

In this outbreak, a recombination was suspected to occur; investigation confirmed the existing recombination that changes the topology and the branch length of the tree.

Two reference phylogenies were used; these were constructed by removing recombination region according to the gold standard methods. In the following report, gold standard methods used to generate the reference phylogenies consist of high quality SNPs filter, recombination detected, exclusion of the SNPs included into a recombination region and final phylogeny build with a maximum-likelihood method.
Tools used to build the reference phylogenies are PHEnix 1.2 for variants calling and filtering, Gubbins 2.0.0 for recombination detection, SnapperDB 0.2.4 to extract relevant SNPs and RAxML 8.2.8 to build the phylogeny. In

this case reference phylogenies will be the phylogenies build following Centre 10 and Centre 11 tools/methods. One included the reference genome, the other did not. This choice was made to balance the bias related to the high diversity on the ST-complex.

Overall results

The results were compared using two main approaches:

1. Alignment and distance matrix comparison
2. Topology of the tree: global topology, Robinson-Fould symmetric difference and percentage of edge similarity (number of branches in one tree that are present in another)

Each result was compared to their closest reference phylogeny i.e. with or without the reference genome.

Alignment and distance matrix

All the participants were required to provide a fasta alignment of the SNPs detected by the method they employed to generate the phylogeny. To ensure consistent comparison of the alignments, we generated the distance matrices from the alignment using an in-house python script. Distances from the reference phylogenies were calculate and the graphic generated using a R script.

Table 2. Alignment and statistic metrics. Columns numbers correspond to Centre (ref. Table 1)

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Min distance matrix	11	5	0	21	0	5	2	0	2	0	0
Max distance matrix	9367	157	177	212	9276	8804	6293	3121	230	1976	30
Reference	+	-	-	+	+	+	+	+	-	+	-

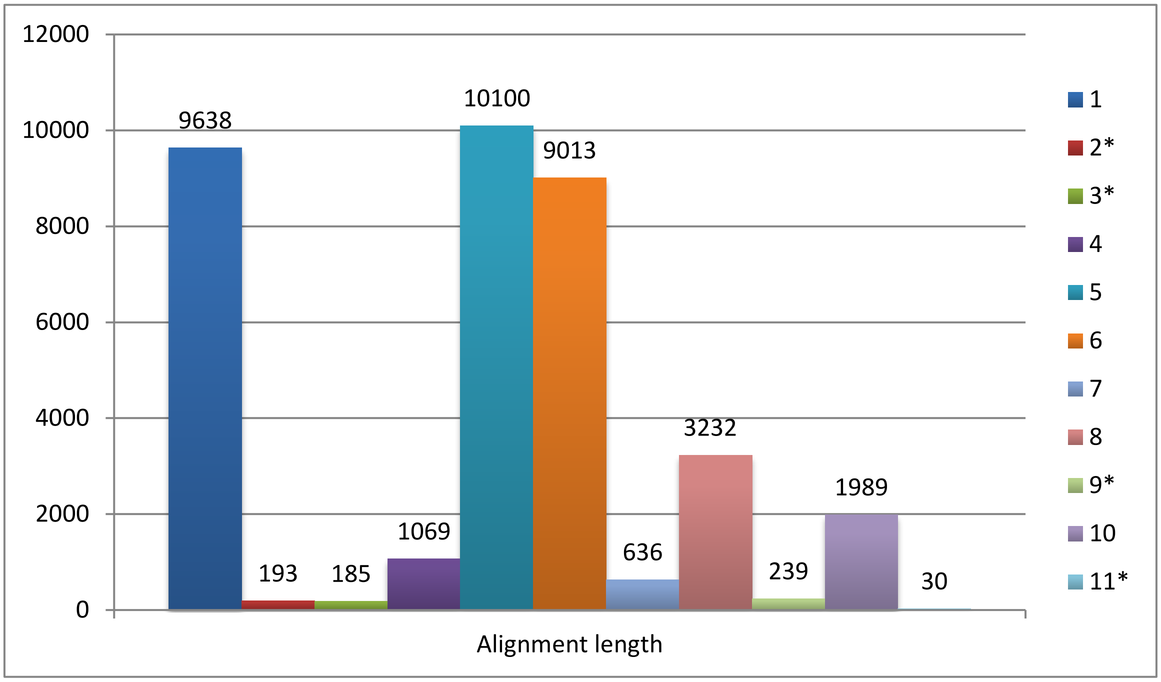
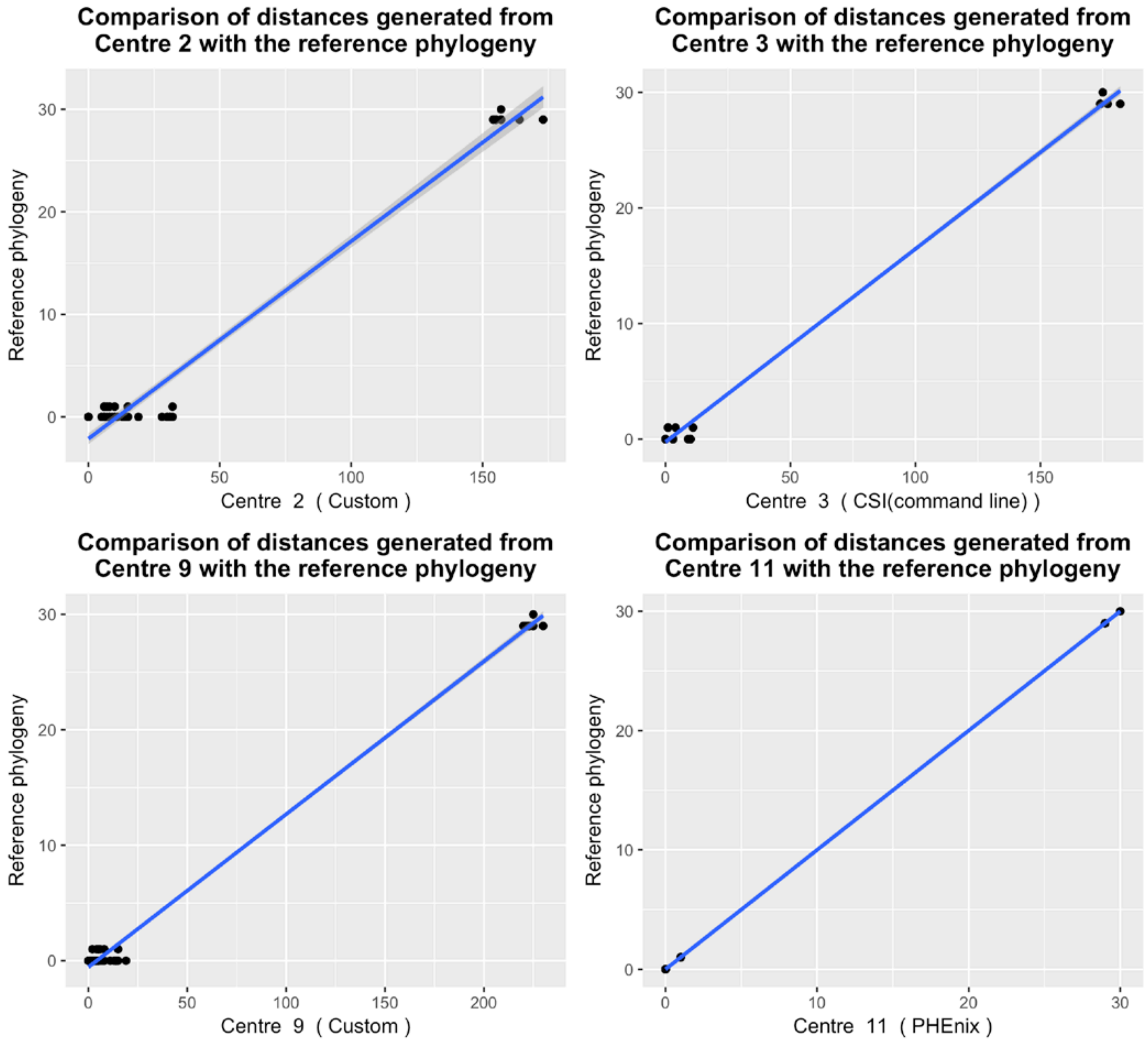


Figure 1. *Size of the SNPs alignments for each results (* indicate no reference in the alignment)*

For this benchmarking, we advised partners about potential recombination in the dataset. Results from Centre 10 and Centre 11 are taking into account recombination at the alignments step.

Results including the reference genome have significant longer alignments due to the number of SNPs present between samples from the dataset and the reference genome. *Campylobacter* is known for having huge diversity inside the same ST-complex. As we can see results that did not include the reference are including only SNPs detected between isolates of the dataset.

The selected isolates were part of a suspected outbreak and therefore the minimal SNP distance in the matrices should reflect the link between isolates. For eight out of eleven results the minimum SNPs distance is < 10 SNPs. This proves that all the methods are able to identify a strong link between some isolates of the dataset. The maximum distance found is highly related to the inclusion of the reference inside the alignment.



Comparison of distances generated from Centre 2 with the reference phylogeny

Comparison of distances generated from Centre 3 with the reference phylogeny

Comparison of distances generated from Centre 9 with the reference phylogeny

Comparison of distances generated from Centre 11 with the reference phylogeny

Figure 2. *Comparisons of distances generated from centre with gold standard <u>without</u> reference genome. Centre numbers correspond to the list of benchmarking tools and participants for variants calling.*

The method used by Centre 2 and Centre 9 to generate the SNP alignment seems to show discrepancies for the closest isolates. This can be related to a recombination between closely related isolates.

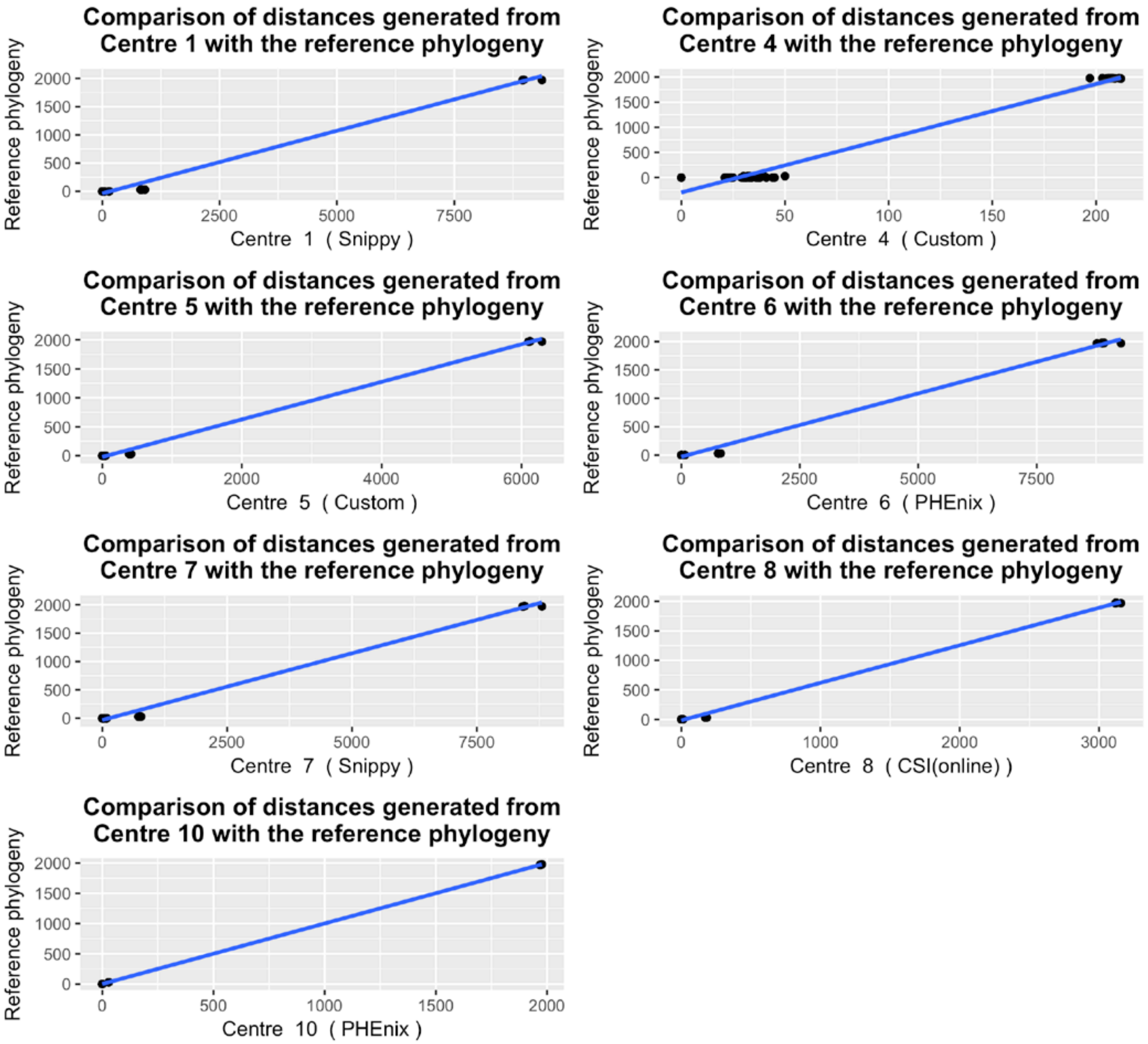


Figure 3. *Comparisons of distances generated from centre with gold standard with reference genome. Centre numbers correspond to the list of benchmarking tools and participants for variants calling.*

The methods used to generate the SNP alignment by the different partner show similar results except for Centre 4 where the comparisons of the distance matrix shows discrepancies. There were also slight differences inside the closest isolates for all the methods used by the partners.

Topology of the tree

All the phylogenies are presented on the additional figure. They are labelled according to row number on the following table. The phylogenetic distance metrics were generated by using the ete toolktit (http://etetoolkit.org/) ete3 v.3.0.0 with his module compare and the additional phangorn R package v2.0.0.

Table 3. *Phylogenetic distance metrics[1]. Row numbers correspond to Centre (ref. Table 1)*

	REF*	E.SIZE	NRF	RF	MAXRF	%SRC-BR+	%REF-BR+	KF DIST
1	+	10	0.86	12.00	14.00	0.62	0.62	25976.63
2	-	9	1.00	12.00	12.00	0.57	0.57	5.979280
3	-	9	0.78	7.00	9.00	0.82	0.64	5.037388
4	+	10	1.00	14.00	14.00	0.56	0.56	3.808749
5	+	10	1.00	14.00	14.00	0.56	0.56	1.428548
6	+	10	1.00	14.00	14.00	0.56	0.56	1.373172
7	+	10	1.00	14.00	14.00	0.56	0.56	1.174685
8	+	10	0.82	9.00	11.00	0.77	0.62	0.984009
9	-	9	1.00	12.00	12.00	0.57	0.57	9.104187
10	+	10	0.00	0.00	14.00	1.00	1.00	0
11	-	9	0.00	0.00	12.00	1.00	1.00	0

* +/- indicated presence/absence of the reference in the final phylogeny

The closer the normalized Robinson-Foulds (nRF) value is to 0, the better the match of the topology to the reference phylogeny. The results shows that most of the trees are very different to the reference in terms of topology. Also, they seem to be consistently different from the reference phylogeny regardless of including the reference genome.

The closest phylogenies are from Centre 1, Centre 3 and Centre 8. Centre 3 and 8 used the CSI-phylogeny (CGE tools) and got better results. The phylogeny provided by Centre 1 is using a recombination detection software similar to the reference phylogeny explaining the similar results.

The KF distance measures the difference in term of branch length. As we can see most of the trees have somewhat a similar branch length. The Centre 1 branch length in the newick file has been derived from the recombination software used to build the phylogeny, it is not based on the SNP explaining why the KF distance (difference in term of branch length) is really high compared to the others methods shown.

Most of the partners have not used a specific tool or method to remove the recombination. The following table shows the matrix of KF distances comparing results between each other.

[1] Information on the metrics are available on the additional notes.

Table 4a. KF distance between all centre phylogenies – reference genome included

	1	4	5	6	7	8	10
1	0.00						
4	25980.44	0.00					
5	25978.06	2.38	0.00				
6	25978.00	2.44	0.06	0.00			
7	25977.81	2.64	0.25	0.20	0.00		
8	25975.65	4.79	2.41	2.35	2.16	0.00	
10	25976.63	3.81	1.43	1.37	1.17	0.98	0.00

Table 4b. KF distance between all centre phylogenies – reference genome not included

	2	3	9	11
2	0			
3	0.9487111	0.00		
9	3.2711648	4.16	0	
11	5.9792803	5.04	9.1041865	0

These matrices pointed out that some of the phylogenies are more similar in term of branch length between each other than they are with the reference phylogenies. For example phylogenies 5 and 6 seems highly similar.

Conclusion

The methods used to generate the SNP alignment by the different partners show similar results except for three (Centre 2, Centre 9 and Centre 4) where the comparisons of the distance matrix shows discrepancies between those isolates that are closely related. The distances matrices are informative to assess the relation between isolates and the phylogeny.

The overall topology of the trees compared to the gold standard reference is respected with all the methods able to pool together the isolates related to the outbreak and detach the suspected source from the main cluster. The main discrepancy linked to the presence of a recombination appears on the branch length and on the topology inside the closely related cluster.

The scores based on the topology demonstrate that most of the methods give different branch length and topology when a recombination occurs within the dataset, this could lead to biased distance matrices and an over-detection of SNPs. Removing the recombinant regions in this case shows strongest evidence of a cluster inside the dataset.

During this benchmarking we have identified that a key point in building a phylogeny where a recombination can occur is to link distance matrices and the phylogeny. The subtle topology of a closely related cluster is highly correlated to the presence of a recombination. We can also confirm with this benchmarking that if the organism is likely to contain recombination and discrepancies occurs between phylogeny and epidemiology information it is recommended to carry out a detection of recombination.

This benchmarking shows that partners have used “gold standard” methods both for SNP detection and tree building[2]. Filtering of SNPs has been properly carried out and most of the participants have used a maximum likelihood method to generate the phylogeny. It also demonstrated that despite knowing the good practise to derive a phylogeny from WGS, phylogeny need to be used with caution and can be only fully explained given support from other data, especially if in relation to outbreak investigations (i.e. for outbreak case definitions).

Additional notes

Meaning of the metrics (ete-compare):

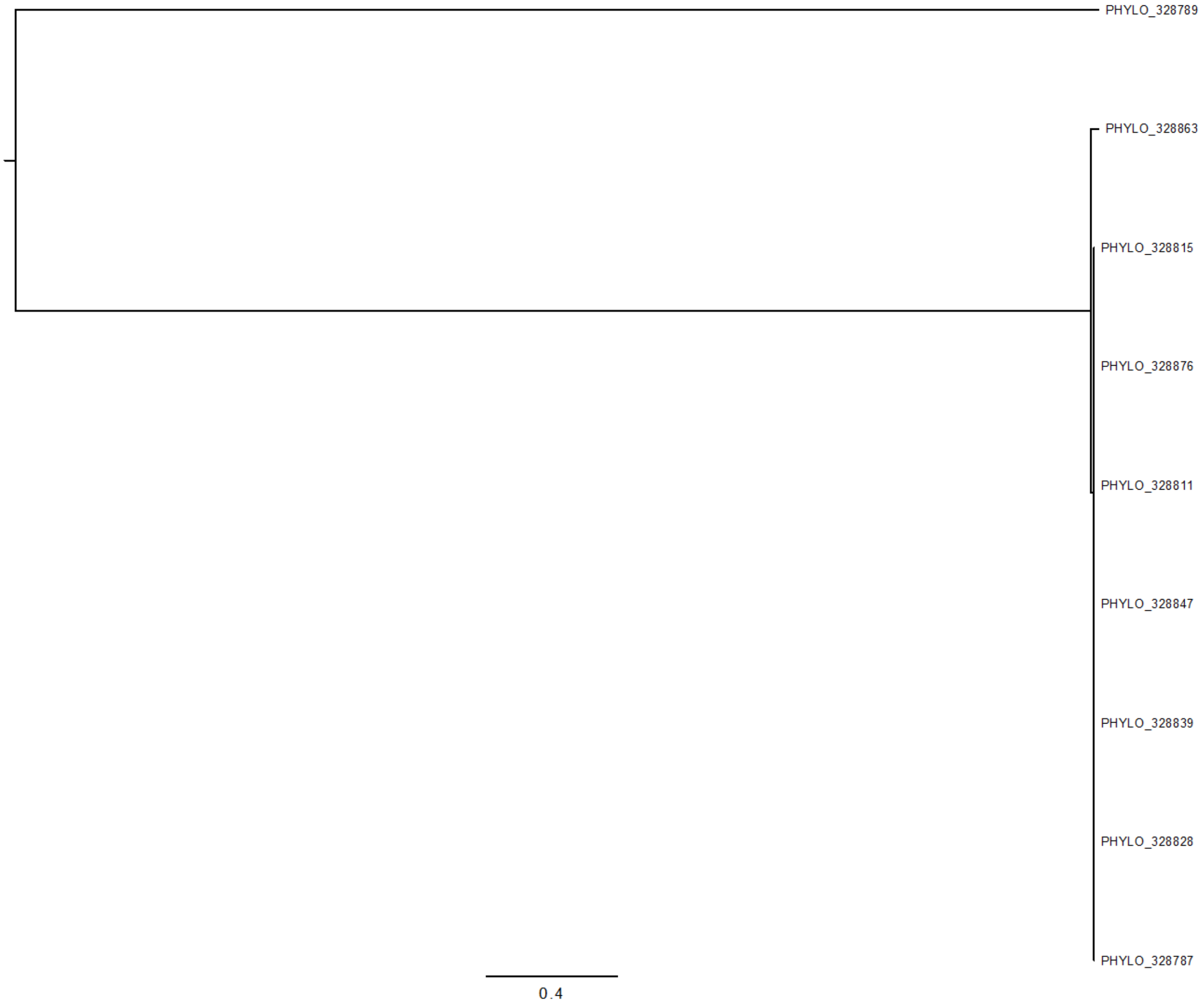
- E.SIZE: effective size of the dataset used to calculate metrics
- nRF: Normalized Robinson-Foulds distance (RF/maxRF)
- RF: Robinson-Foulds symmetric distance
- maxRF maximum Robinson-Foulds value for this comparison
- %src_br (percent source branch): frequency of edges in target tree found in the reference (1.00 = 100% of branches are found)
- %ref_br (percent reference branch): frequency of edges in the reference tree found in target (1.00 = 100% of branches are found)
- KF.dist (Kuhner-Felsenstein distance): branch score distance (Kuhner & Felsenstein 1994) [compute with Phargorn]

Additional Table 1. Genomes selected for the benchmarking (further info in Supplementary Table 6 (Annex F))

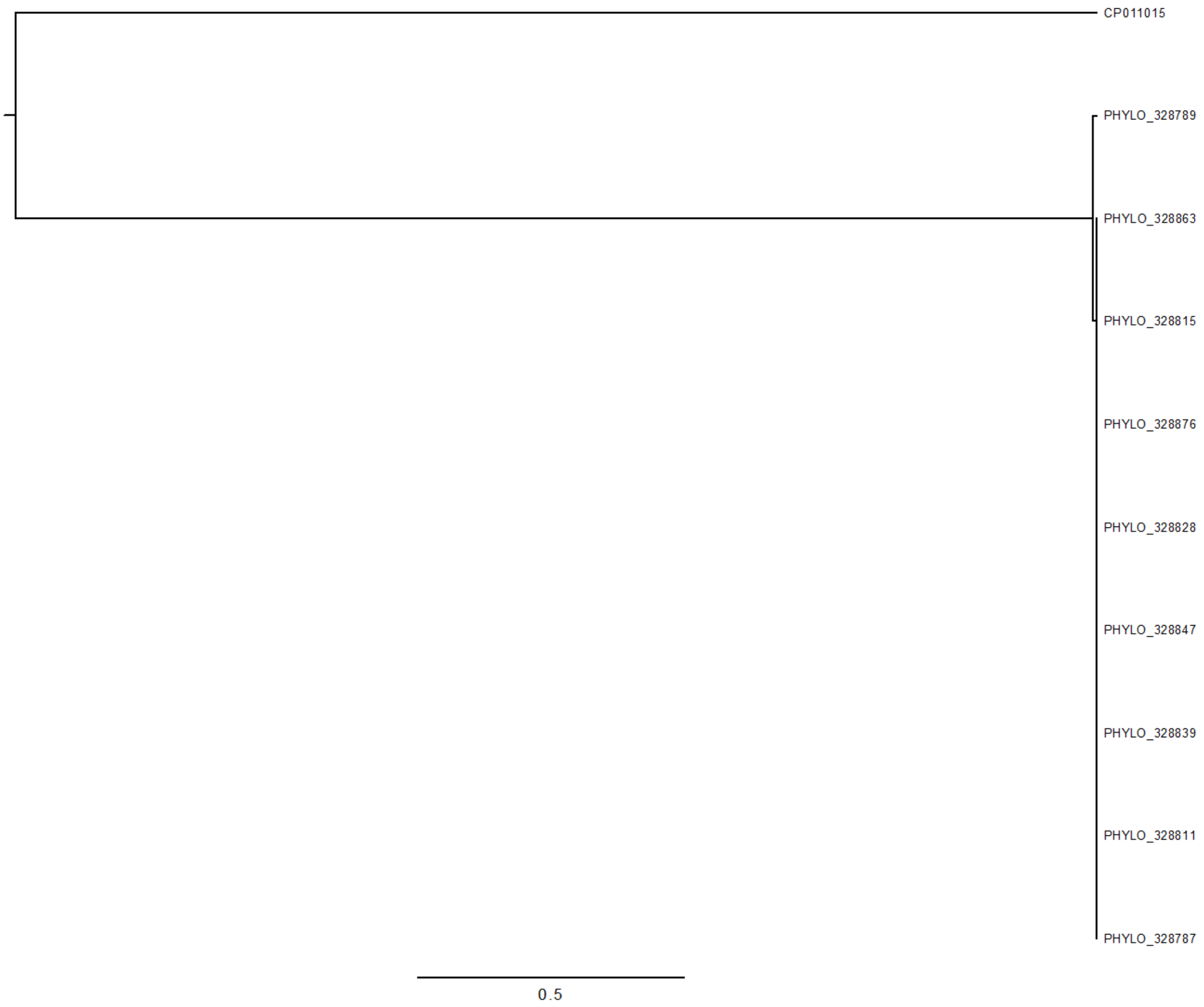
Sequence name
Reference: CP011015
PHYLO_CAMPY_328787
PHYLO_CAMPY_328789
PHYLO_CAMPY_328811
PHYLO_CAMPY_328815
PHYLO_CAMPY_328828
PHYLO_CAMPY_328839
PHYLO_CAMPY_328847
PHYLO_CAMPY_328863
PHYLO_CAMPY_328876

[2] Gold standard methods here referred to the filtering apply to detect SNPs and build the phylogeny with a maximum likelihood method.

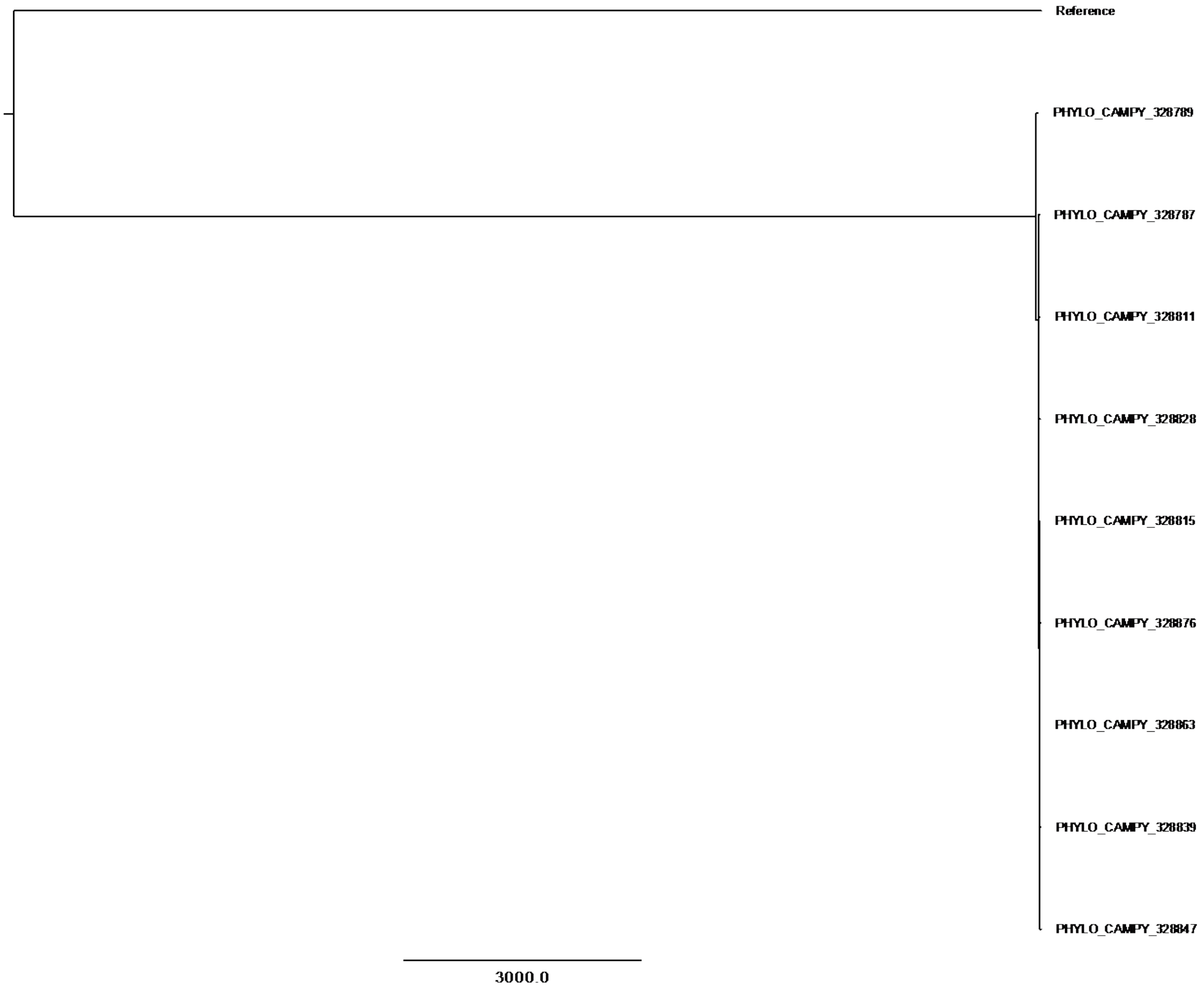
Additional figures



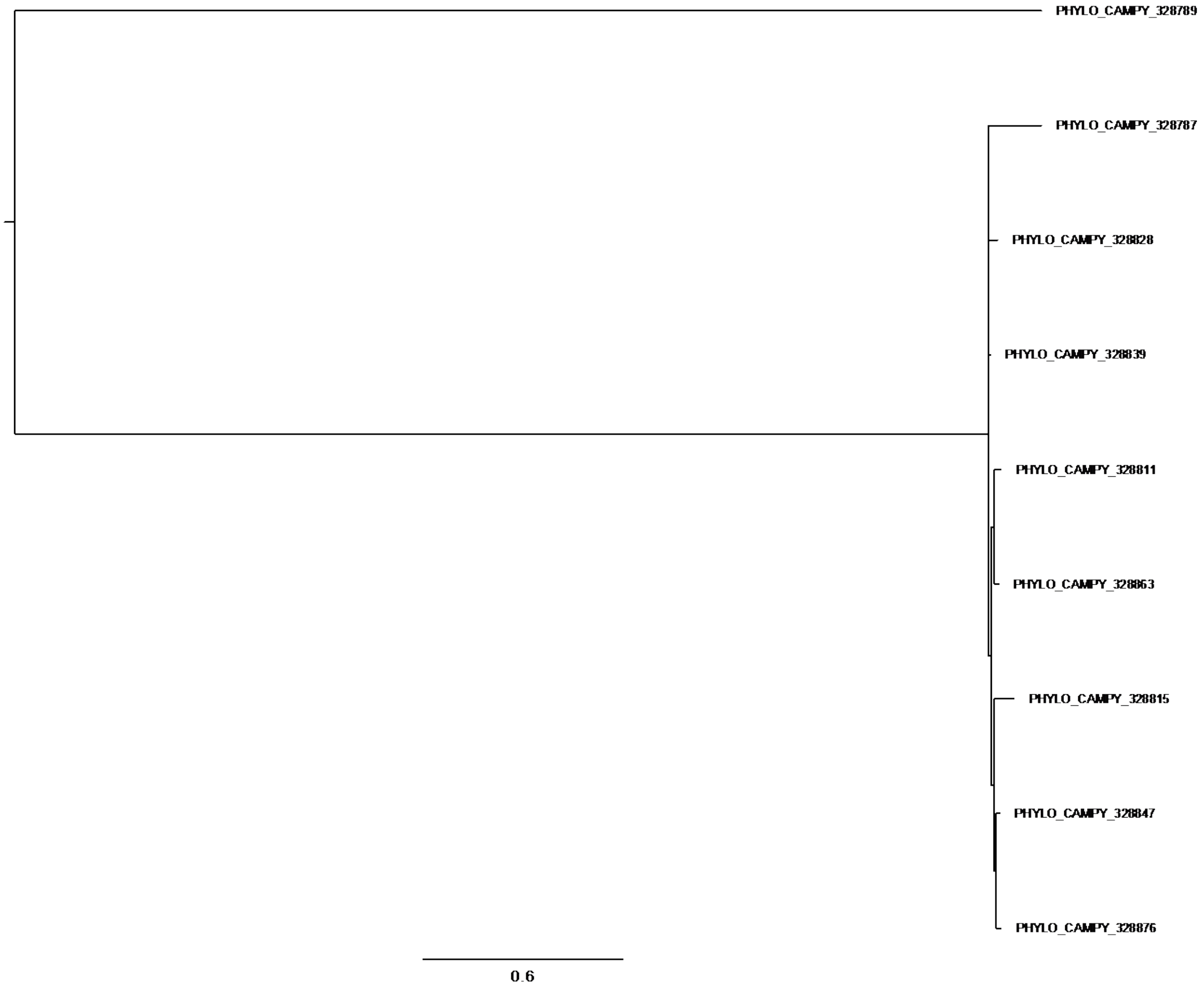
Additional Figure 1. Reference phylogeny without reference genome and recombination removed (scale represents the branch length stipulated into the newick file)



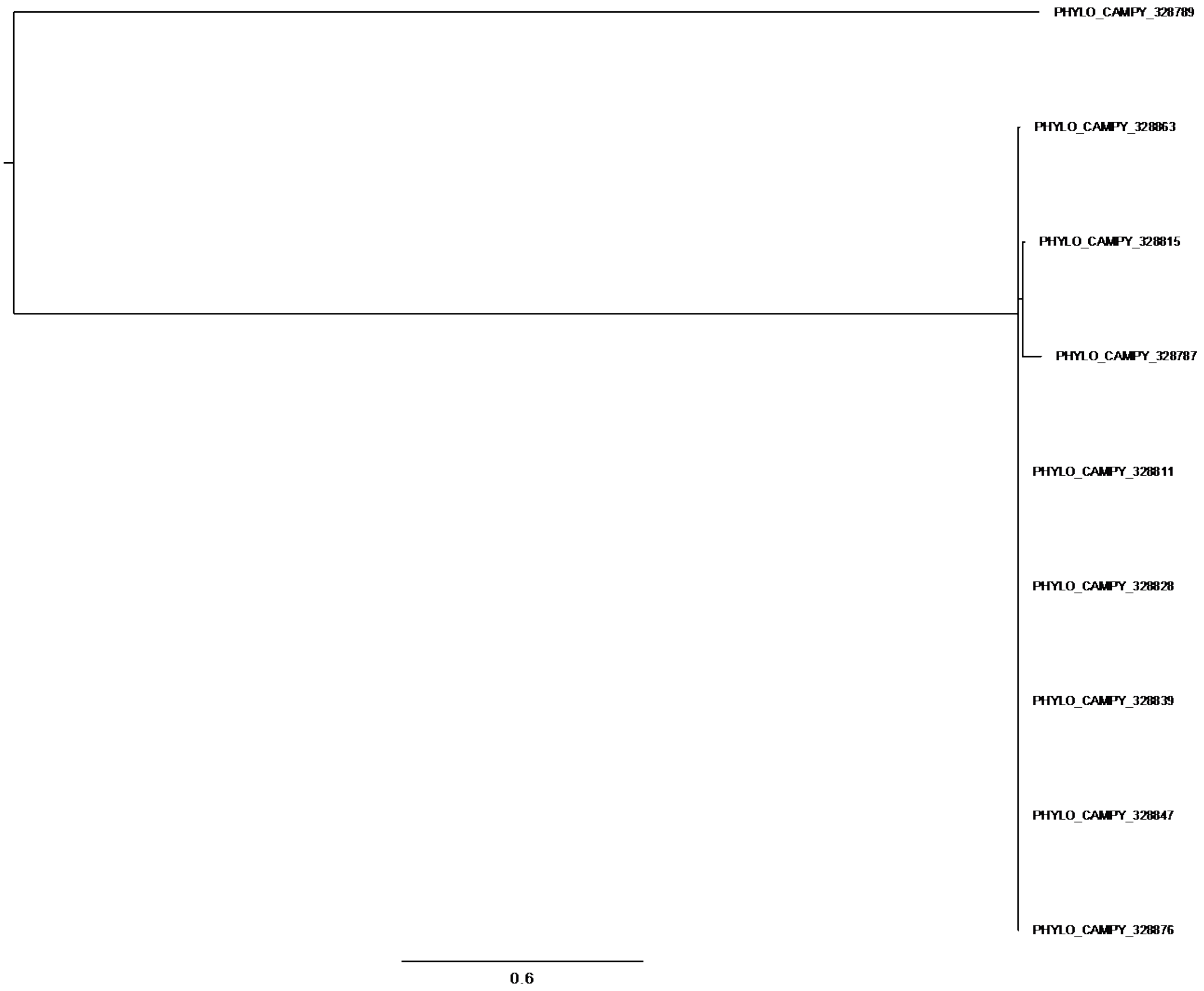
Additional Figure 2. Reference phylogeny with reference genome and recombination removed (scale represents the branch length stipulated into the newick file)



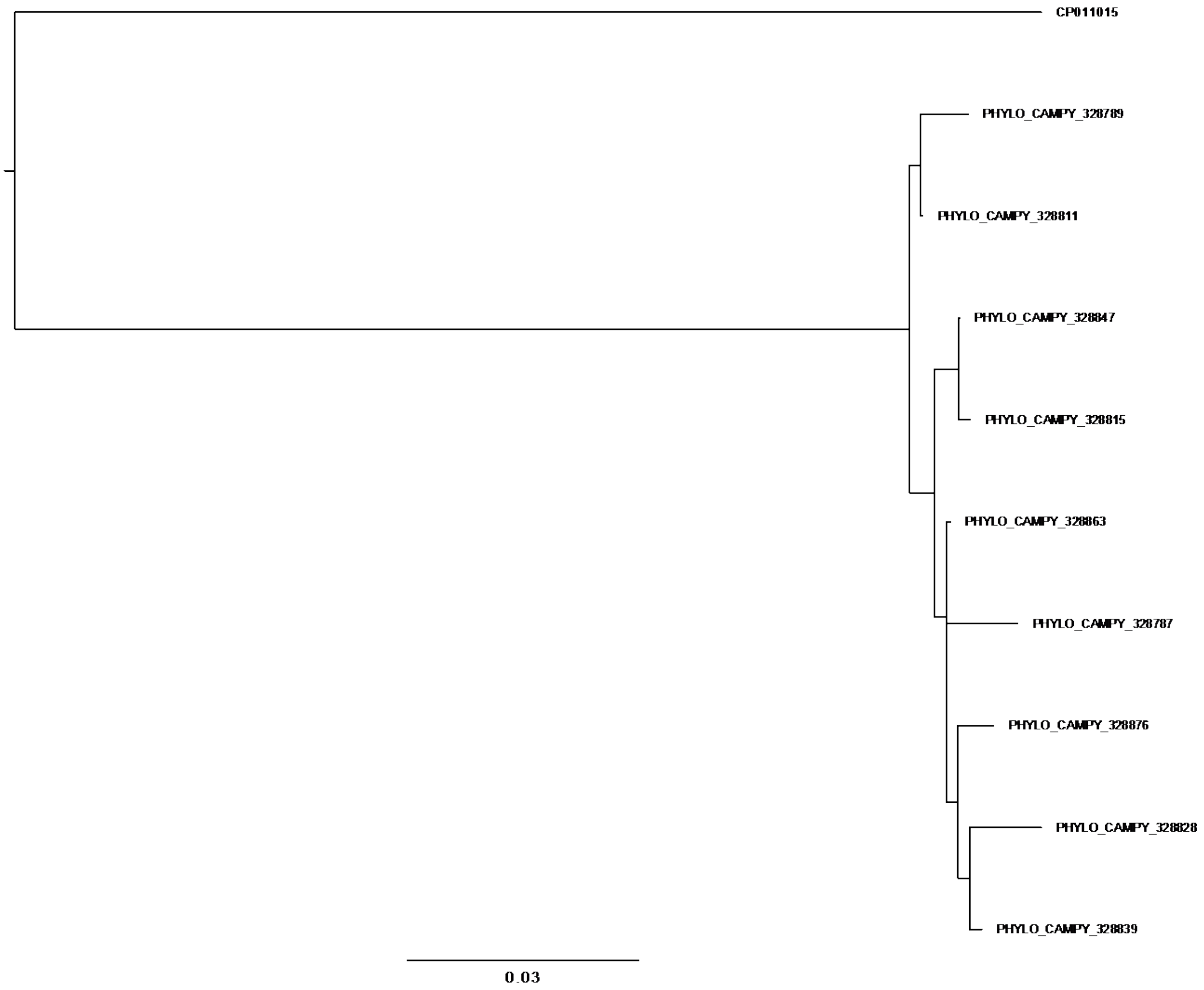
Additional Figure 3. Phylogeny results Centre 1 (scale represents the branch length stipulated into the newick file)



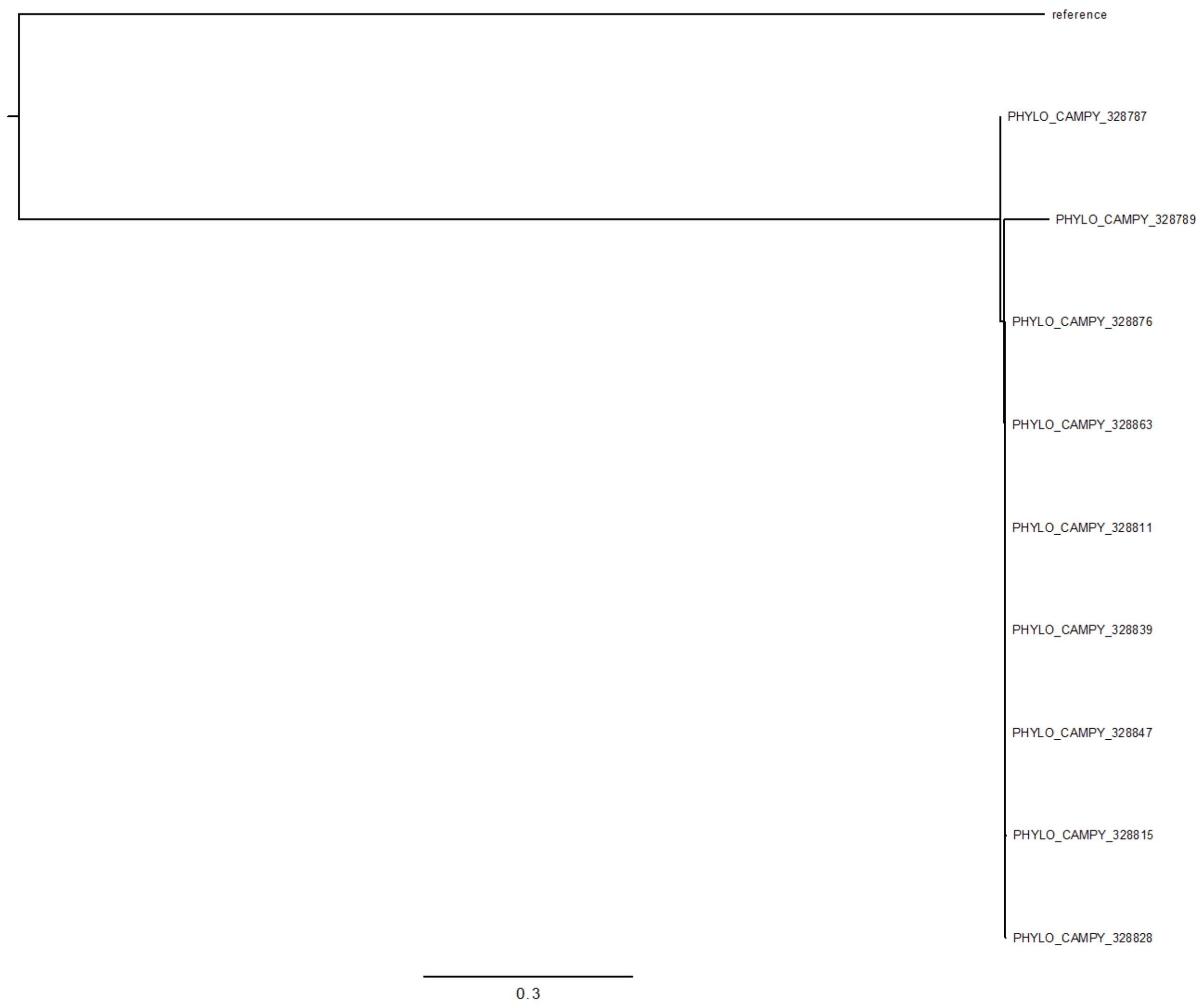
Additional Figure 4. Phylogeny results Centre 2 (scale represents the branch length stipulated into the newick file)



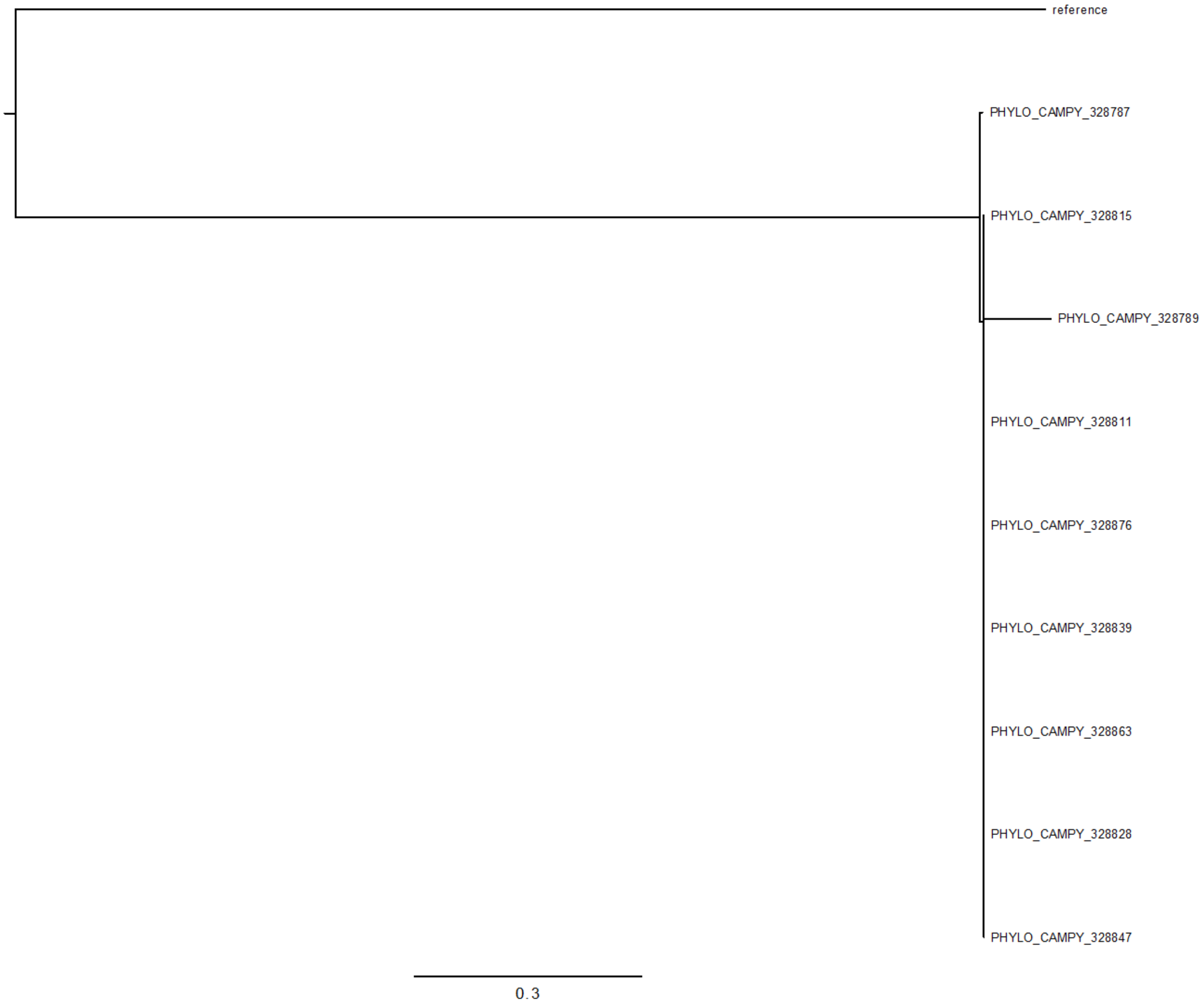
Additional Figure 5. Phylogeny results Centre 3 (scale represents the branch length stipulated into the newick file)



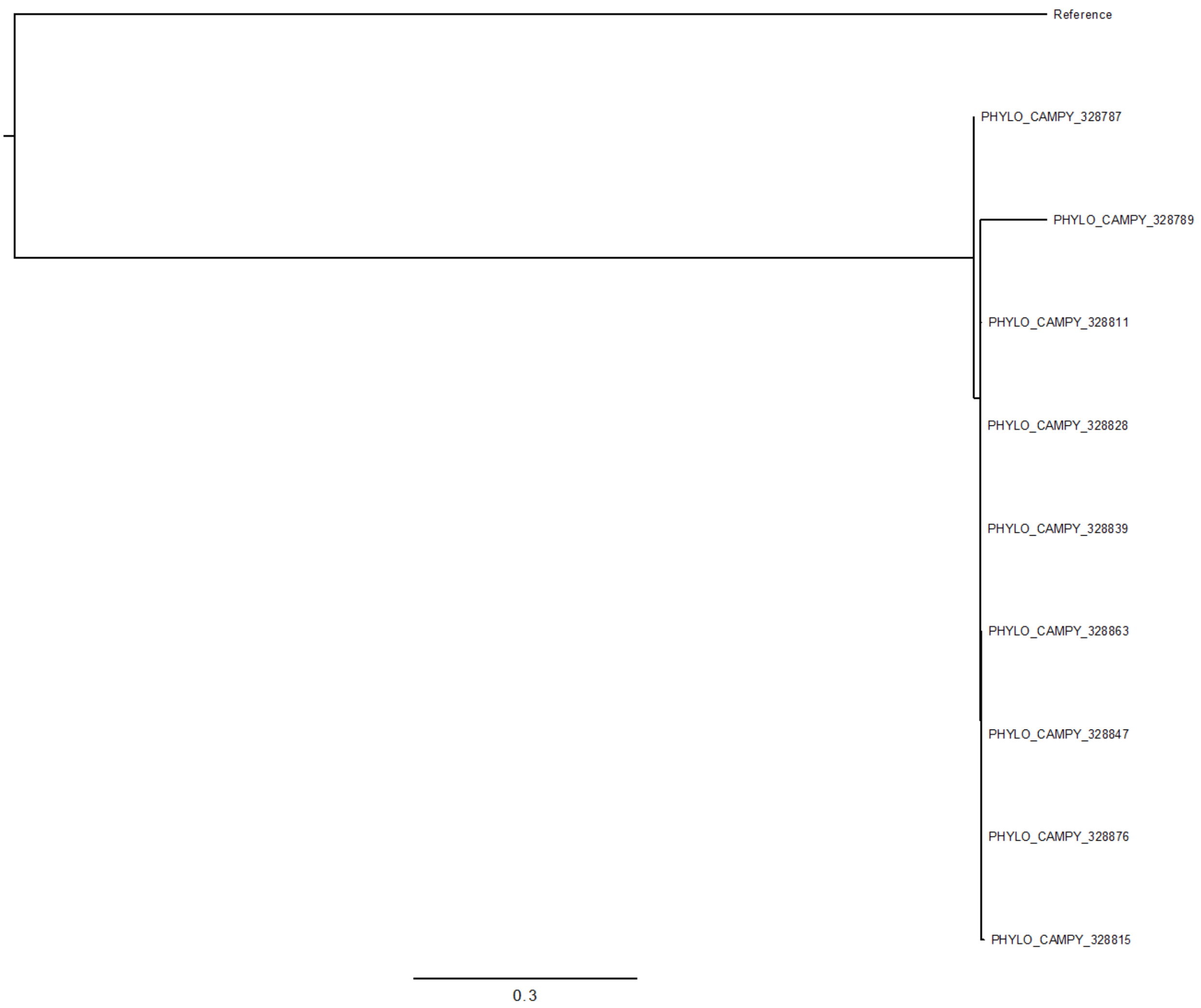
Additional Figure 6. Phylogeny results Centre 4 (scale represents the branch length stipulated into the newick file)



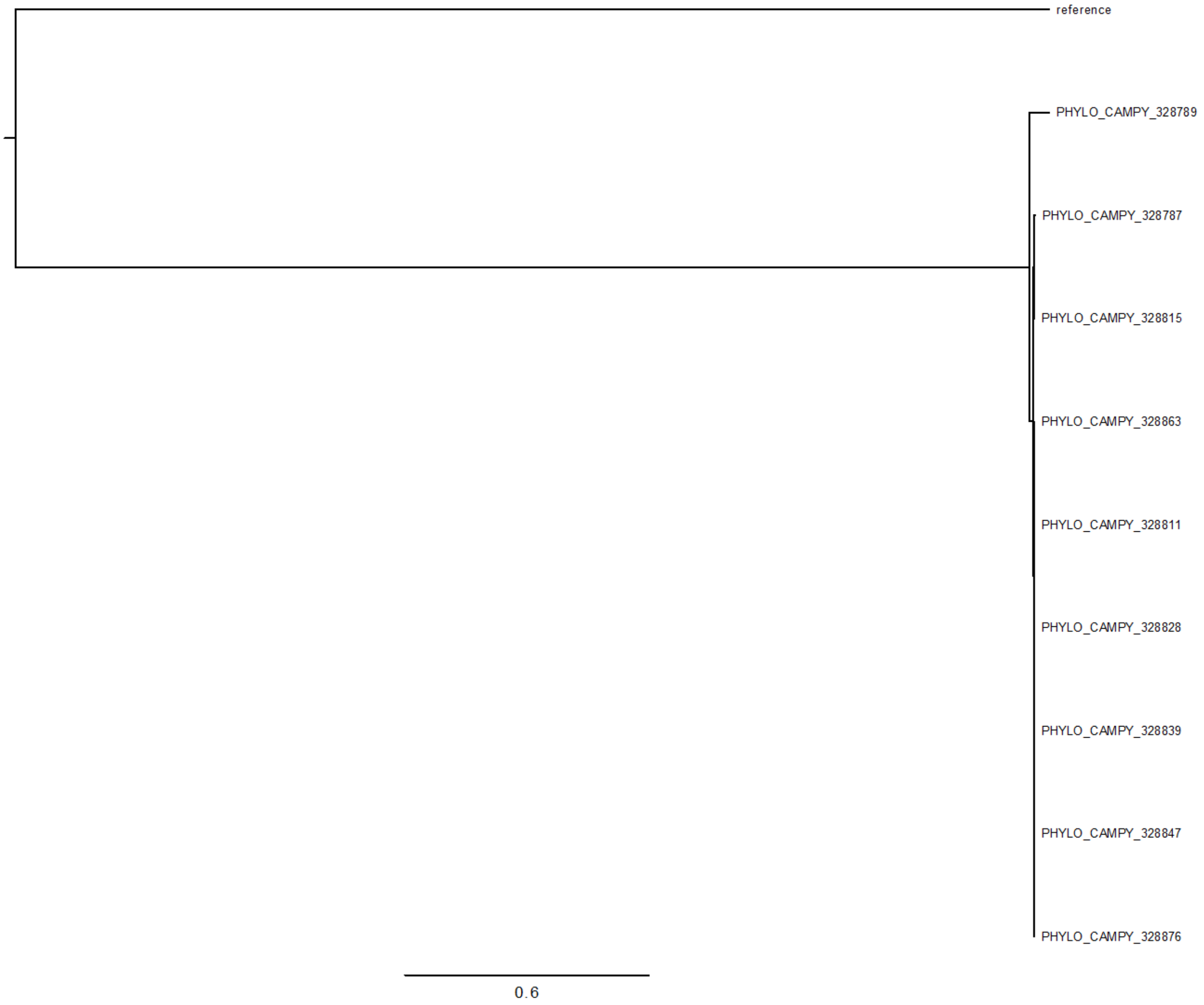
Additional Figure 7. Phylogeny results Centre 5 (scale represents the branch length stipulated into the newick file)



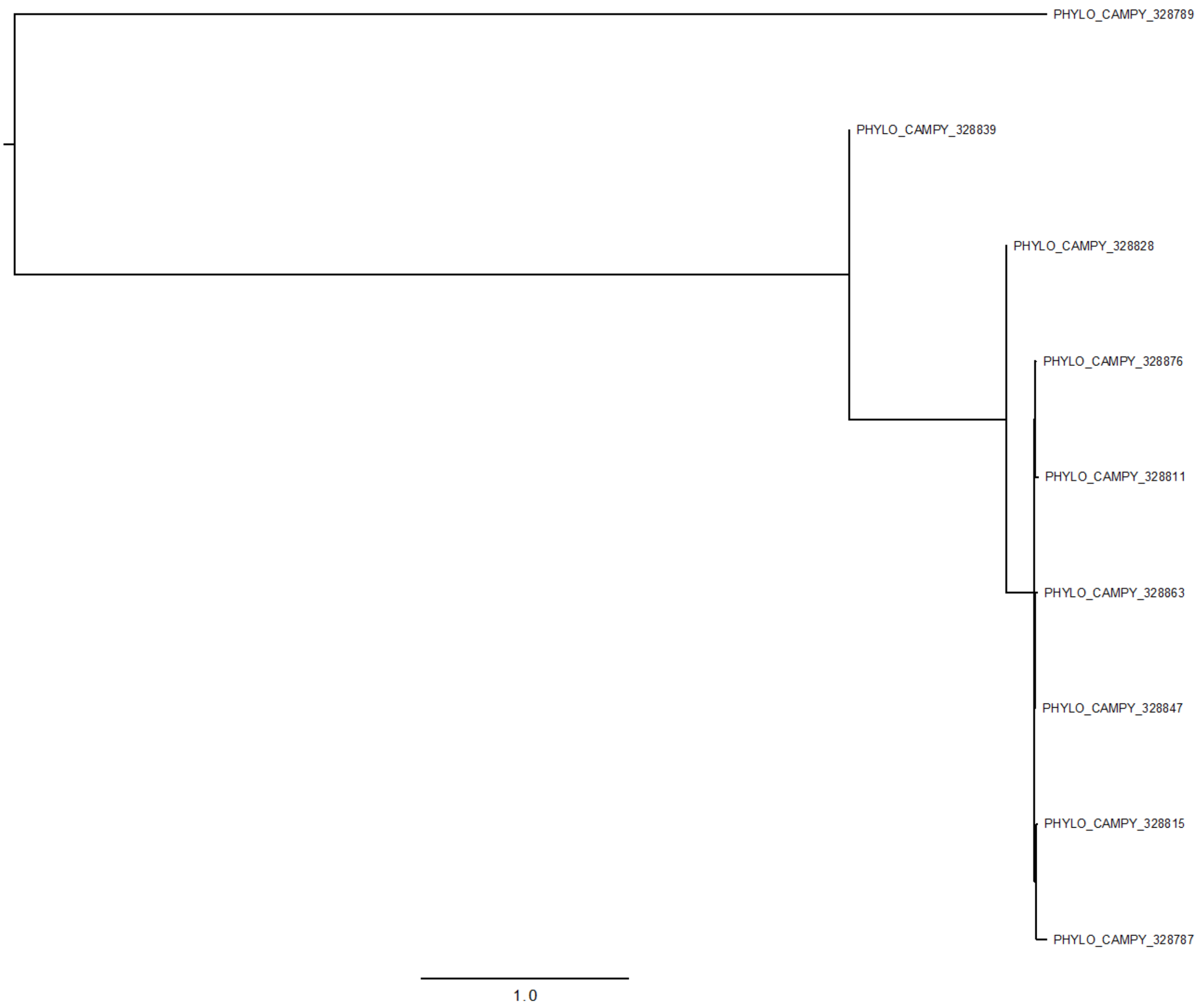
Additional Figure 8. Phylogeny results Centre 6 (scale represents the branch length stipulated into the newick file)



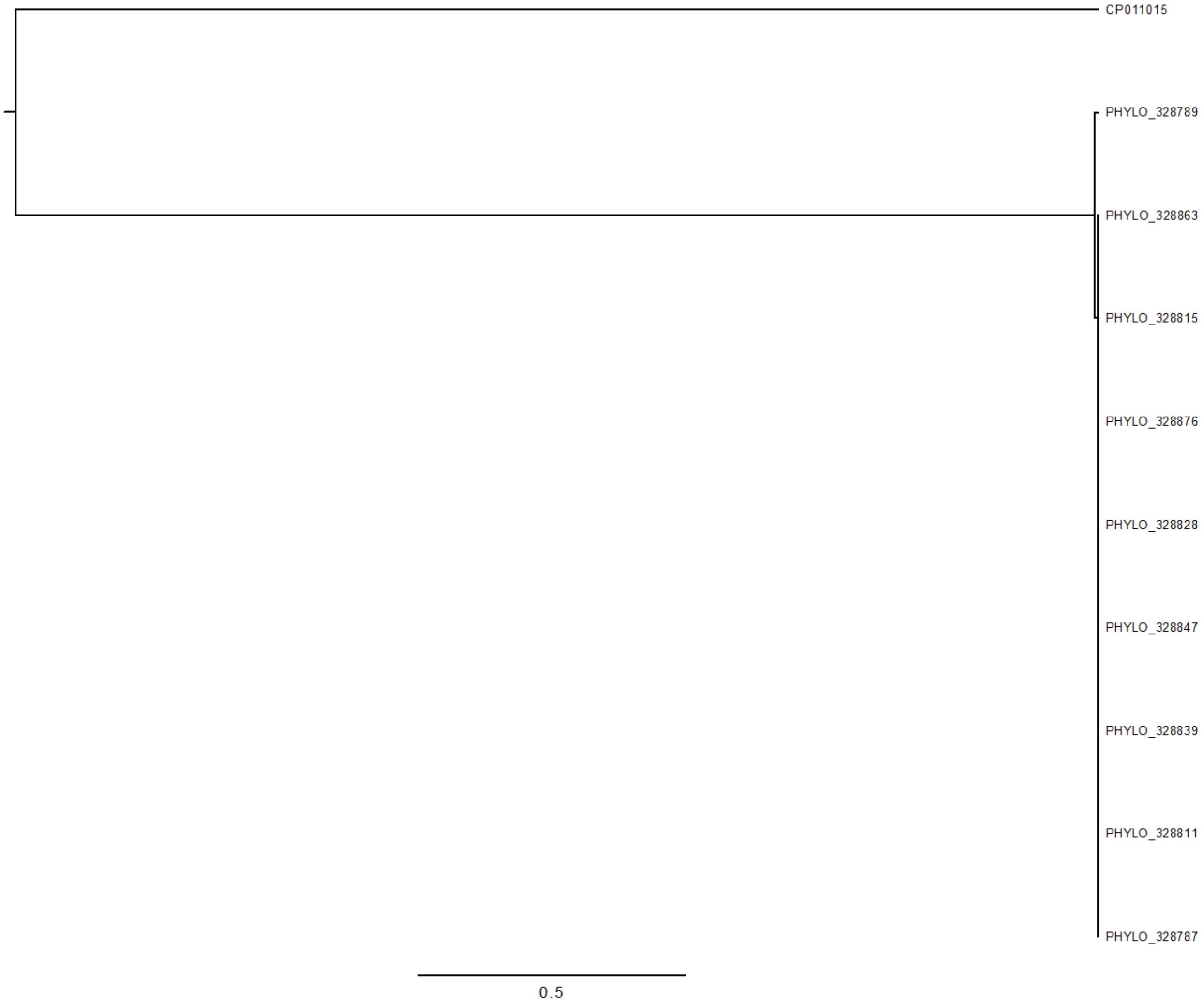
Additional Figure 9. Phylogeny results Centre 7 (scale represents the branch length stipulated into the newick file)



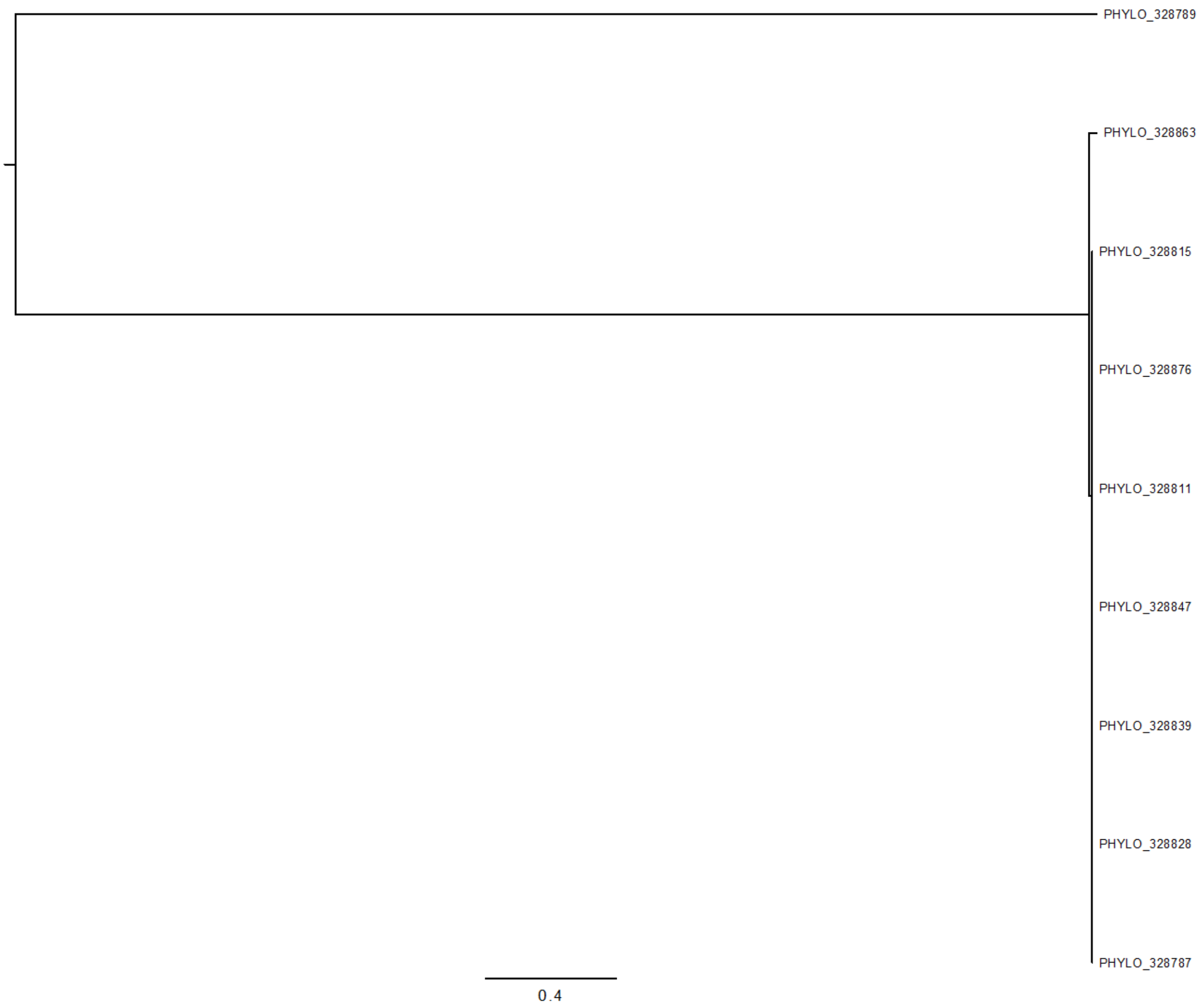
Additional Figure 10. Phylogeny results Centre 8 (scale represents the branch length stipulated into the newick file)



Additional Figure 11. Phylogeny results Centre 9 (scale represents the branch length stipulated into the newick file)



Additional Figure 12. Phylogeny results Centre 10 (scale represents the branch length stipulated into the newick file)



Additional Figure 13. Phylogeny results Centre 11 (scale represents the branch length stipulated into the newick file)

--- --- ---