Benchmarking for *Salmonella Enteritidis* phylogeny

Report number	#5
Responsible	Anaïs Painset (PHE) and Anthony Underwood (PHE)
Other partners/institutions involved	APHA (United Kingdom), BfR (Germany), EFSA (Italy), DTU (Denmark), IZSLT (Italy), IZSVe (Italy), NIPH-NIH (Poland), NVRI (Poland)
Benchmarking launched (date)	September 2017
Deliverable due (date)	October 2017

Purpose of the benchmarking exercise

The main purpose of this benchmarking exercise was to evaluate a number of available bioinformatics tools both to detect variants and to build a phylogeny based on the variants alignment detected for *Salmonella* Enteritidis isolates. With the use of Whole Genome Sequencing, phylogeny is used as a method to characterize microorganisms in outbreak investigations and for surveillance of isolates that are genetically related.

Participants

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

Results from participating institutes are identified by codes (1-12 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).

Tools benchmarked

Benchmarking by variants calling and generating SNPs alignment using the following tools and setup:

1. Snippy v3.0 [default setting: min depth 10, 90% difference from ref]
2. BioNumerics 7.6 (- Mapping /SNP Filtering (relative coverage: total: 5, forward: 1, reverse: 1, unreliable bases, ambiguous bases, gaps, non-informative SNPs))
3. CGE Tools (command line version) – CSIPhylogeny v1.4
4. CGE Tools – CSIPhylogeny v1.4 online version default parameters (BWA v. 0.7.2 + BEDTools v. 2.16.2 + SAMTools v. 0.1.18)
5. CGE Tools CSIPhylogeny v1.4 online version default parameters (BWA v. 0.7.2 + BEDTools v. 2.16.2 + SAMTools v. 0.1.18)
6. CGE Tools – CSIPhylogeny v1.4 online version default parameters (BWA v. 0.7.2 + BEDTools v. 2.16.2 + SAMTools v. 0.1.18)
7. CGE Tools – CSIPhylogeny v1.4 online version default parameters and reference include in the final phylogeny (BWA v. 0.7.2 + BEDTools v. 2.16.2 + SAMTools v. 0.1.18)
8. CGE Tools – CSIPhylogeny v1.4 online version default parameters and reference include in the final phylogeny (BWA v. 0.7.2 + BEDTools v. 2.16.2 + SAMTools v. 0.1.18)
9. CGE Tools – CSIPhylogeny v1.4 online version, reference include in the final phylogeny (BWA v. 0.7.2 + BEDTools v. 2.16.2 + SAMTools v. 0.1.18)
10. Custom pipeline

- Trimmomatic v.0.36 and Nextera-PE adapters to trim the reads. Following parameters were set: ILLUMINACLIP:Nextera-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
- BWA MEM v.0.7.13 with default settings for mapping paired and unpaired reads (after trimming)
- Freebayes v.1.1.0 (d784cf8) with "--ploidy 1" for joint variant calling on all samples
- R package VariantAnnotation v1.22.3 to filter variants: variant calls with genotype-likelihood (GL) > -30 (likelihood > 10e-3) were set to unknown genotype.
- VCF-kit v.0.1.2 "pheno fasta" and "pheno tree nj" to generate alignment and newick tree

11. PHEnix 1.2 (BWA mapping + GATK variant calling) + SnapperDB 0.2.4 (get the snps A:80)
12. CGE Tools – CSIPhylogeny v1.4 online version default parameters (BWA v. 0.7.2 + BEDTools v. 2.16.2 + SAMTools v. 0.1.18)

Benchmarking by building trees using the following tools and setup:

1. RAxML v.8.2.9
2. Bionumerics v.7.6: Neighbor joining tree
3. CGE Tools – CSIPhylogeny 1.4 command line (FastTree built-in)
4. CGE Tools – CSIPhylogeny 1.4 online version (FastTree built-in)
5. MEGA-CC 7 Minimum Evolution Methods
6. MEGA-CC 7 Maximum Parsimony
7. CGE Tools – CSIPhylogeny 1.4 online version (FastTree built-in)
8. CGE Tools – CSIPhylogeny 1.4 online version (FastTree built-in)
9. CGE Tools – CSIPhylogeny 1.4 online version (FastTree built-in)
10. VCF-kit v0.1.2 with pheno tree nj
11. RAxML v8.2.8-multithread (-N autoMRE -f a -p 12345 -x 12345 -m GTRCAT)
12. CGE Tools – CSIPhylogeny 1.4 online version (FastTree built-in)

Species/genomes included

Public Heath England selected and provided genomes from *Salmonella enterica* serotype Enteriditis and part of the same eburst group EGB4. The genomes have been selected because they were part of an outbreak investigated by PHE. This outbreak has been well-studied (Dallman et al., 2016) and epidemiological information support the phylogeny associated with the selection.

Thirty genomes represented by sets of fastq (paired) were included in the data set (Annex F). All genomes originated from sequencing using an Illumina HiSeq. 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, then the quality of the sequencing was assessed by running FastQC 0.11.3. The trimmed and quality assessed reads were used for the analysis (Supplementary Table 6, Annex F).

The gold standard phylogeny use to perform comparisons was constructed following the methods employed by Centre 11. Tools used to build the gold standard phylogenies are PHEnix 1.2 for variants calling and filtering, follow by SnapperDB 0.2.4 to extract relevant SNPs and RAxML 8.2.8 to build the phylogeny. In this benchmarking, gold standard will be the phylogeny build following Centre 11 tools/methods.

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)

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 script and build the graphic with an in-house R script.

Eight out of the twelve set of results provided by the partners were generated by using the CGE CSI Phylogeny tools. We've decided to regroup results in Table 1 where the parameters were the same.

Table 1. Alignment and statistic metrics. Columns numbers correspond to the results submit by the partners. The list of benchmarking tools and participants for variants calling.

Results	1	2	3/4/5/6/12	7/8	9	10	11
Alignment length	779	698	633	644	636	1465	786
Min distance matrix	0	0	0	0	0	0	0
Max distance matrix	558	489	428	428	422	713	527
Reference	+	-	-	+	+	-	+

The longest the alignment is the more SNPs have been detected in the dataset. Min and max distance matrix represent the number of SNP different between strains in the dataset. Strains supposedly part of an outbreak or closely related are expected to have a low number of SNP difference. The minimum distance captures the minimum number of SNPs between two strains in the dataset i.e. the two closest strains in the dataset. The maximum distance reflects the maximum number of SNPs between two strains in the dataset i.e. the more distant strain in the dataset.

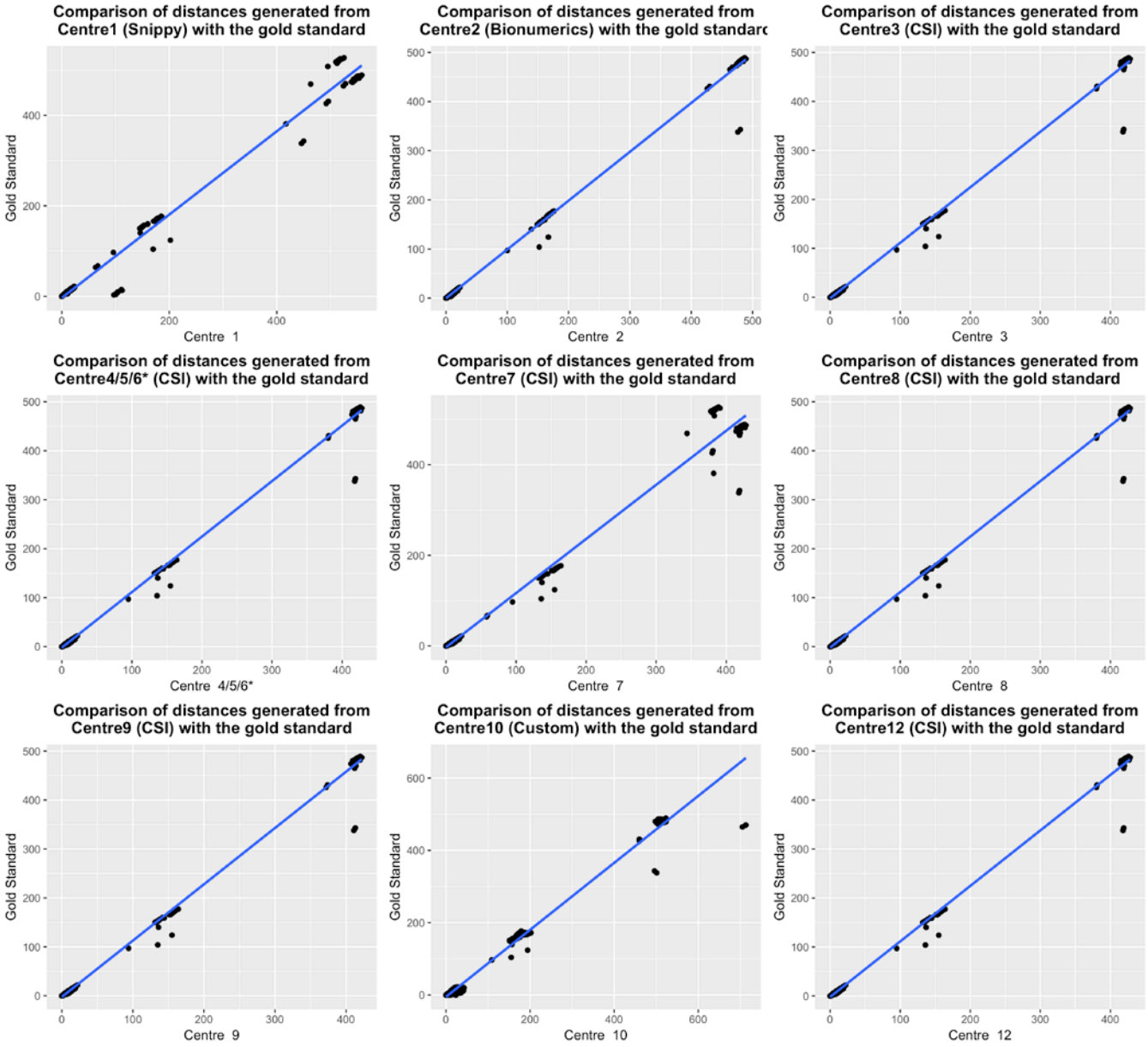


Figure 1. Comparisons of distances generated from centre with gold standard. Centre numbers correspond to the list of benchmarking tools and participants for variants calling.[1]

[1] Method use by centre 11 was used as the gold standard and is not represented on the comparisons.
* Phylogenies 4/5/6 were based on the same alignment, therefore only one graph can be produce.

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 2. Phylogenetic distance metrics. Columns numbers correspond to the list of benchmarking tools and participants for the tree building[2].

	1	2	3	4	5	6	7	8	9	10	11	12
E.size	31	30	30	30	30	30	31	31	31	30	31	30
Ref	+	-	-	-	-	-	+	+	+	-	+	-
nRF	0.46	0.32	0.2	0.2	0.37	0.37	0.19	0.19	0.19	0.78	0	0.2
RF	26	13	10	10	20	20	10	10	10	42	0	10
maxRF	56	41	50	50	54	54	52	52	52	54	56	50
src-br+	0.78	1	0.94	0.94	0.82	0.82	0.94	0.94	0.94	0.62	1	0.94
ref-br+	0.78	0.77	0.88	0.88	0.82	0.82	0.88	0.88	0.88	0.62	1	0.88
KF.dist	0.198	356.692	0.073	0.073	313.742	-	0.151	0.151	0.166	0.395	0	0.073

The closer the normalized Robinson-Foulds (nRF) value is to 0, the better the match of the topology to the 'gold standard' phylogeny. As we can see most of the trees are close to the reference one. One tree (Centre 10) is significantly different in terms of topology compared to the gold standard.
The KF distance (KF.dist) measures the difference in term of branch length. As we can see most of the trees have really similar branch length. Tree of centre 2 and tree of centre 5 are not using the SNPs for the alignment as a branch length and this would explain why the difference in term of branch length is really high.
The tree of centre 6 does not provide branch length in the newick file and therefore was exclude from this metric (Table 2).

Table 3. Clade retrieval from gold standard compared to others methods. Columns numbers correspond to the list of benchmarking tools and participants for the tree building.*

	1	2	3	4	5	6	7	8	9	10	11	12
Reference	+	-	-	-	-	-	+	+	+	-	+	-
Outliers n=5	N	Y	Y	Y	Y	Y	Y	Y	Y	N	Y	Y
Clade I n=3	N	Y	Y	Y	Y	Y	Y	Y	Y	N	Y	Y
Clade II n=8	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y	Y
Clade III n=14	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y	Y

*n = number of isolates per clade, N if all isolates from the gold standard clade are not retrieve on the same clade in the phylogeny, Y if all isolates from the gold standard clade are retrieve on the same clade. +/- indicated presence/absence of the reference in the phylogeny

This is a topological assumption: isolates from a clade are considered correct if they are on the same

[2] See additional notes for information regarding the metrics

monophyletic branch. The three clades should be separated from the outliers by a long branch. The assumption is that isolates group in clades accordingly to the gold standard (Table 3).

The following tanglegrams illustrate the difference/similarity between the gold standard and the phylogeny where we found clade discrepancies. The lines in the middle reflect inversions in the position of isolates between the two phylogenies; it is used to illustrate the most different trees in terms of clade retrieval compare to the gold standard.

Figure 2. Tanglegram of the gold standard (left) versus the most different topology produce by Centre 10.

Figure 3. Tanglegram of the gold standard (left) versus the topology produce by Centre 1.

Conclusion

The methods used to generate the SNP alignment by the different partners showed similar results except for three Centres, Centre 1 (Snippy), Centre 7 (CSI without heterozygote removal) and Centre 10 (Custom pipeline) where the comparisons of the distance matrix show discrepancies between those isolates that are distantly related. Also, the topology produced shows great similarity, the number of SNPs difference between isolates can vary based on the tools and parameters.

The scores based on the topology demonstrate that most of the methods tested are able to retrieve the topology derived from the gold standard. Only one method seems to give a markedly different topology (Centre 10, custom pipeline).

During this benchmarking we have identified that a key point in building a phylogeny based on the SNP differences between isolates is the detection and filtering of the SNPs. Based on this benchmarking we can recommend a minimum depth coverage for the SNPs detection > 10, a minimum mapping read quality of 30, and 90% consensus for the reads mapped at a position that differs from the reference.

The best tools to build tree from an alignment are maximum likelihood methods. Topology obtained using these methods produce trees with the best correlation between gold standard and the obtained phylogeny.

Additional notes

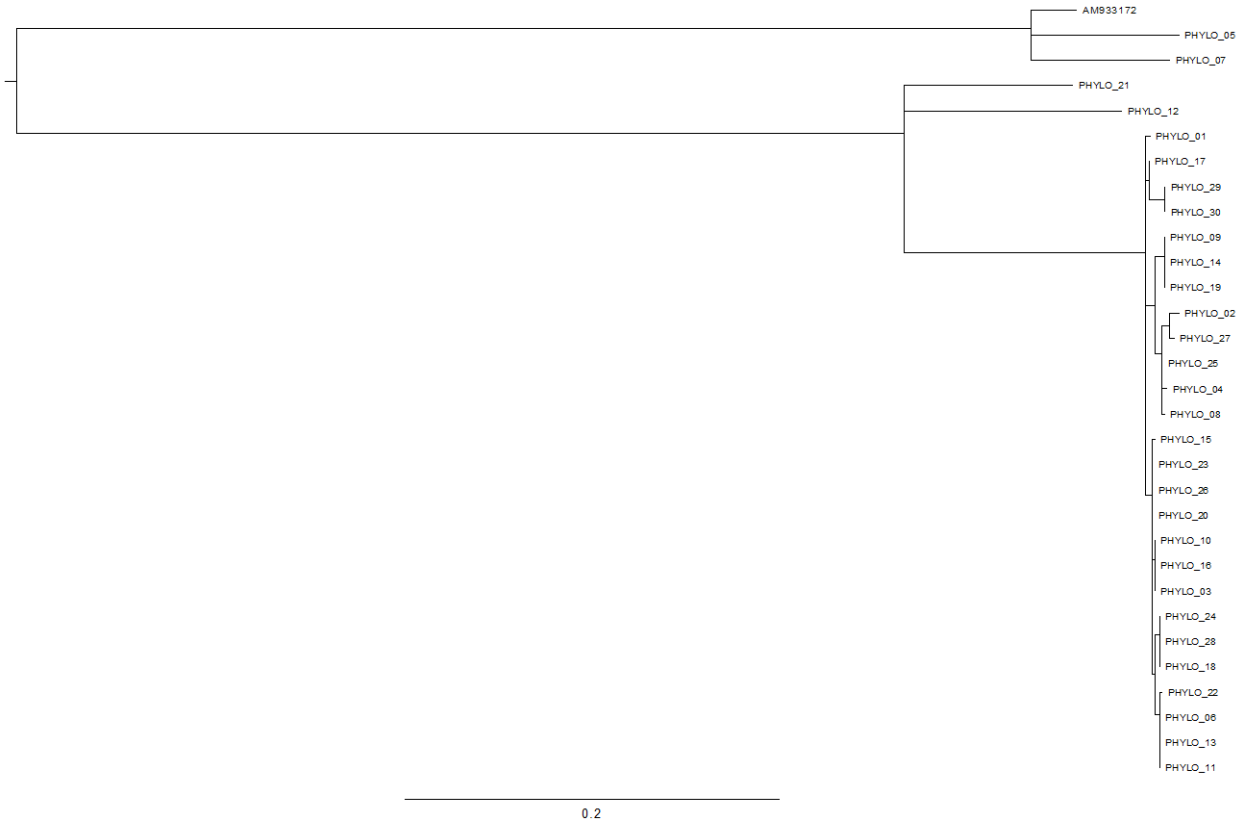
Table 2 meaning of the metrics (ete-compare):

- 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]

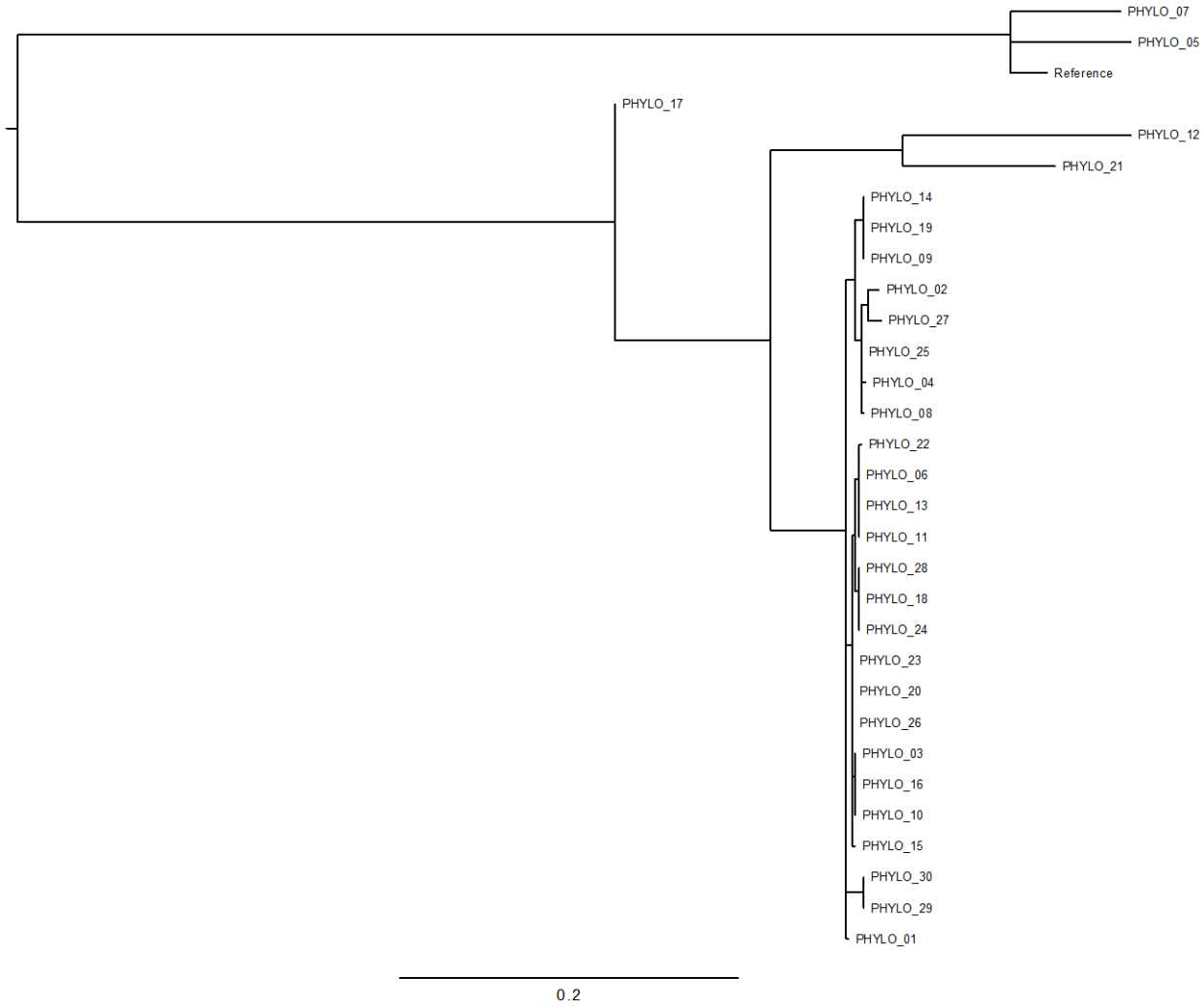
References

Dallman, T., Inns, T., Jombart, T., Ashton, P., Loman, N., Chatt, C., Messelhaeusser, U., Rabsch, W., Simon, S., Nikisins, S., et al. (2016). Phylogenetic structure of European Salmonella Enteritidis outbreak correlates with national and international egg distribution network. Microb. Genomics 2.

Additional Figures



Additional Figure 1. Gold standard phylogeny with reference (scale represents the branch length stipulated into the newick file)



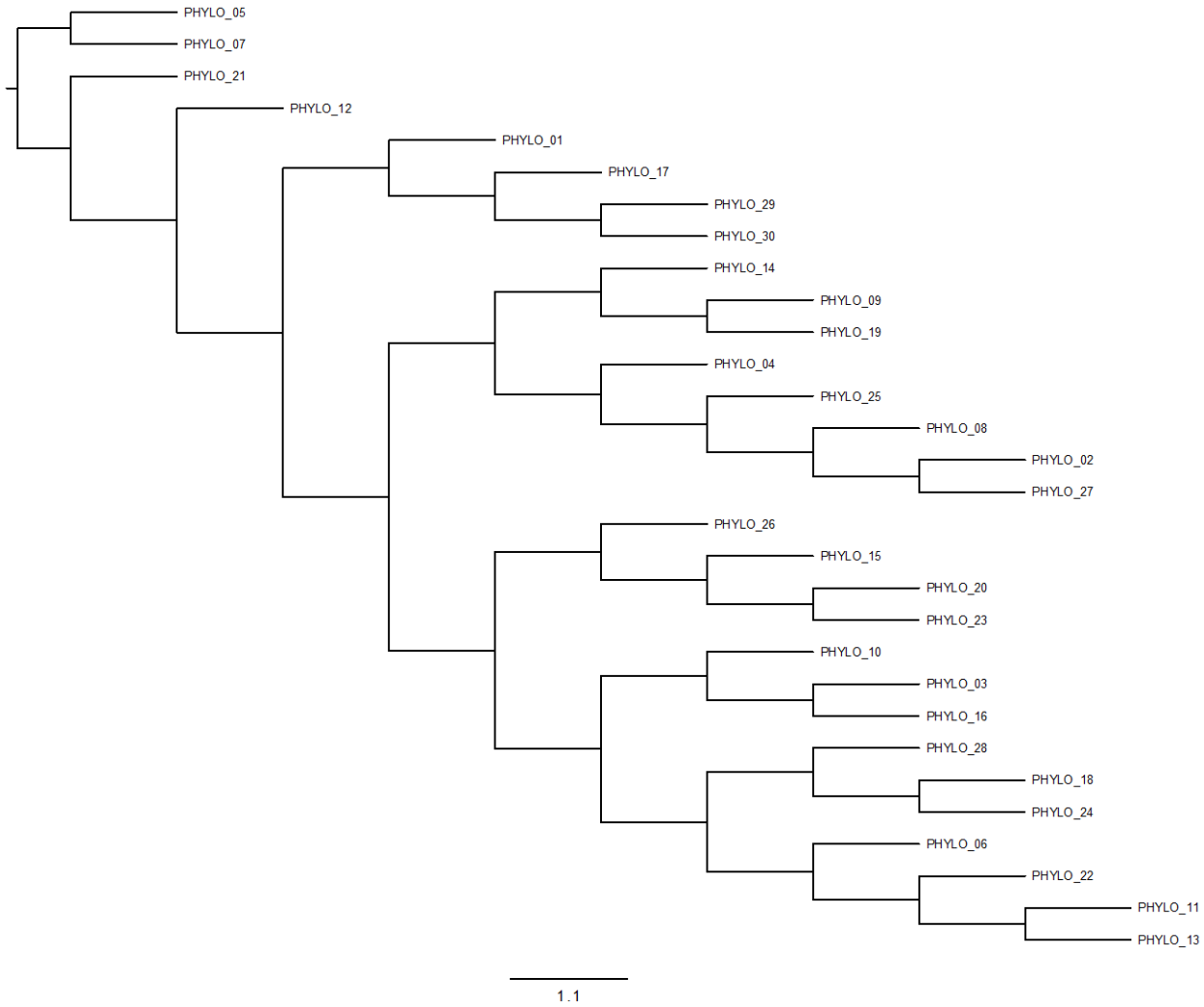
Additional Figure 2. Phylogeny Centre 1 obtained with Snippy tool and RAxML (scale represents the branch length stipulated into the newick file)

Additional Figure 3. Phylogeny Centre 2 obtained with BioNumerics and a Neighbor joining tree reconstruction (scale represents the branch length stipulated into the newick file)

Additional Figure 4. Phylogeny Centre 3 obtained with CSI Phylogeny (CGE tools, command line version) (scale represents the branch length stipulated into the newick file)

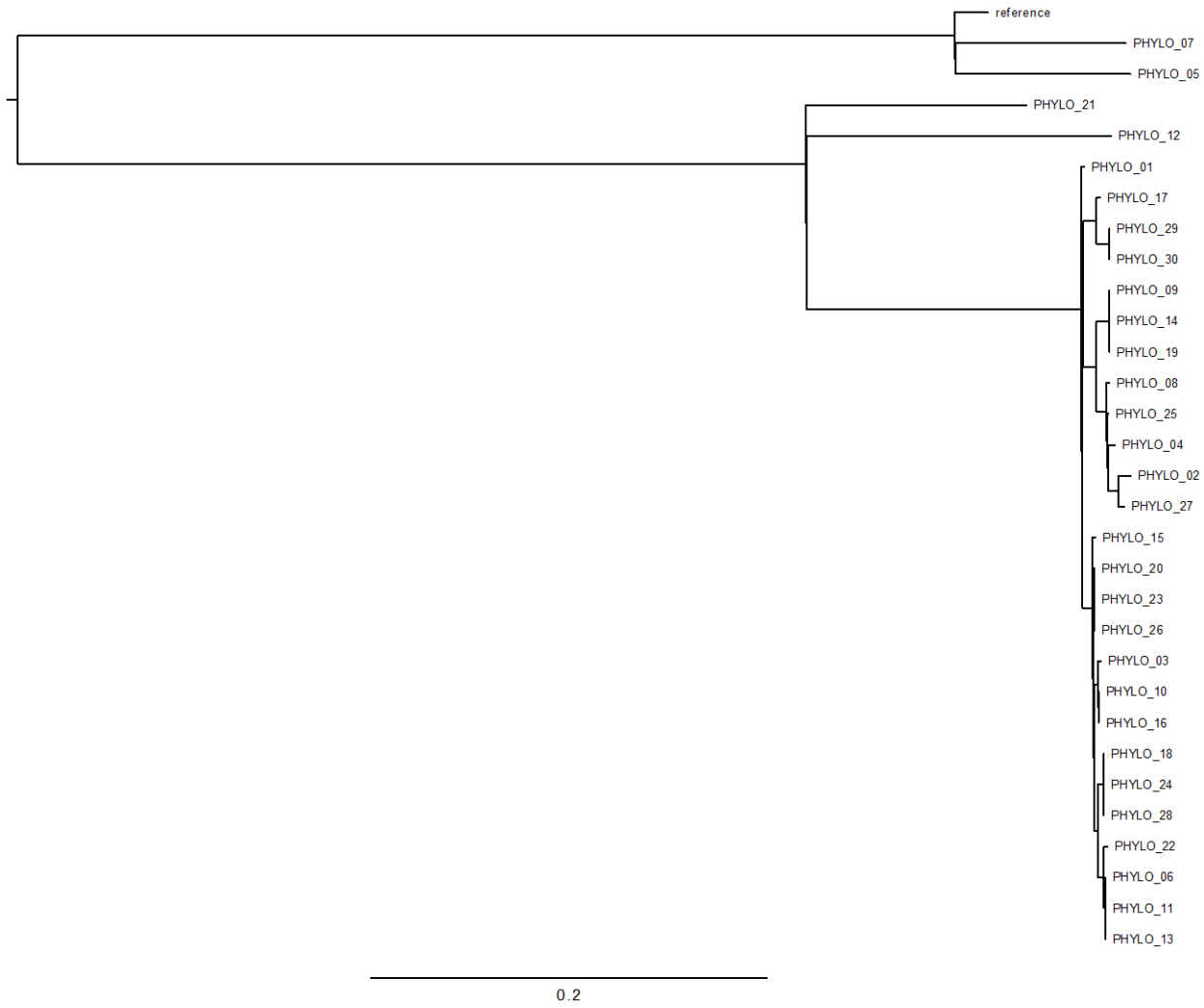
Additional Figure 5. Phylogeny Centre 4 obtained with CSI Phylogeny (CGE tools, online version) (scale represents the branch length stipulated into the newick file)

Additional Figure 6. Phylogeny Centre 5 obtained with CSI tools alignment (command line version) and a minimum evolutionary model for tree reconstruction (scale represents the branch length stipulated into the newick file)

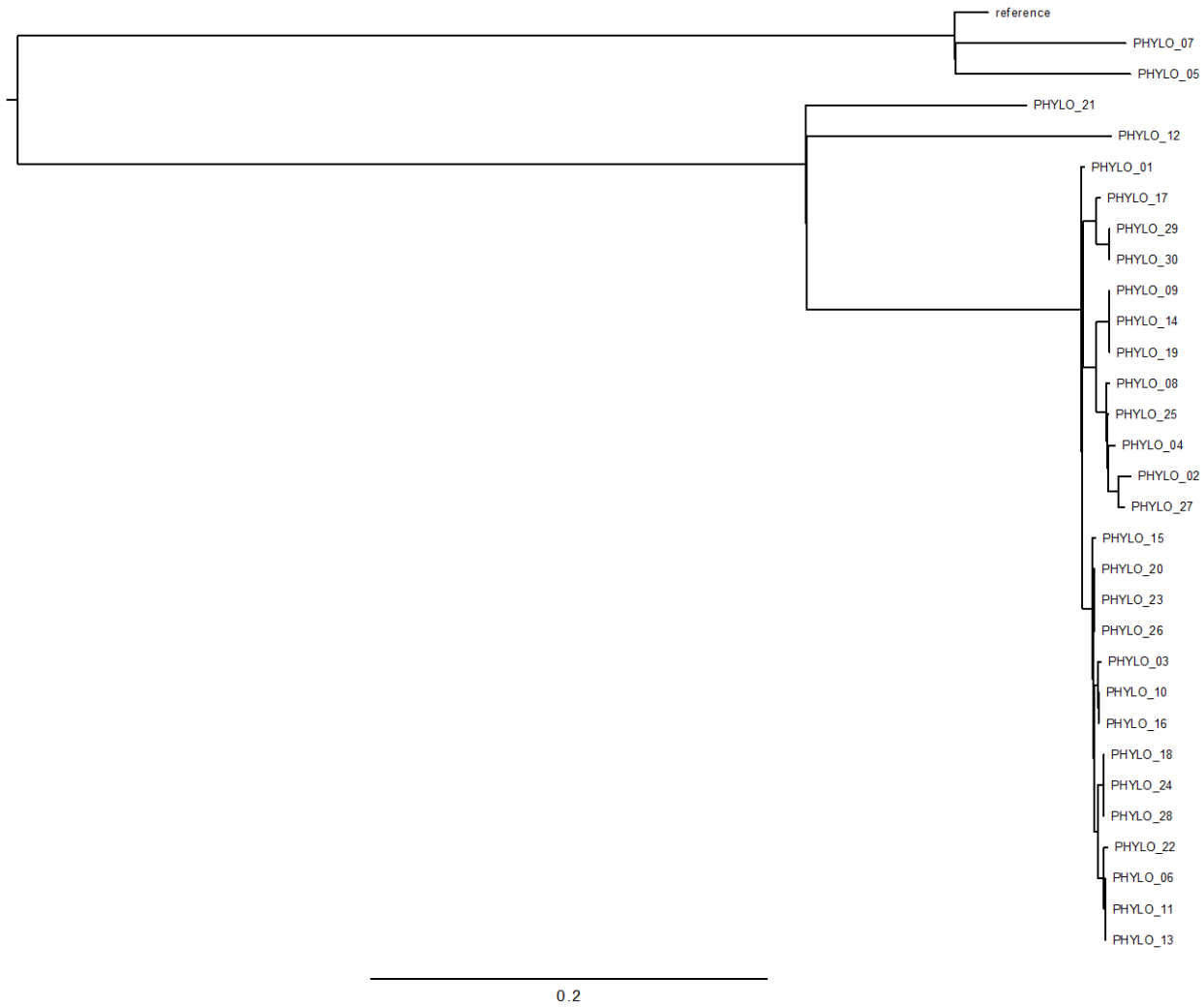


Additional Figure 7. Phylogeny Centre 6 obtained with CSI Phylogeny (CGE tools, command line version) and a maximum parsimony tree reconstruction[3] (scale represents the branch length stipulated into the newick file)

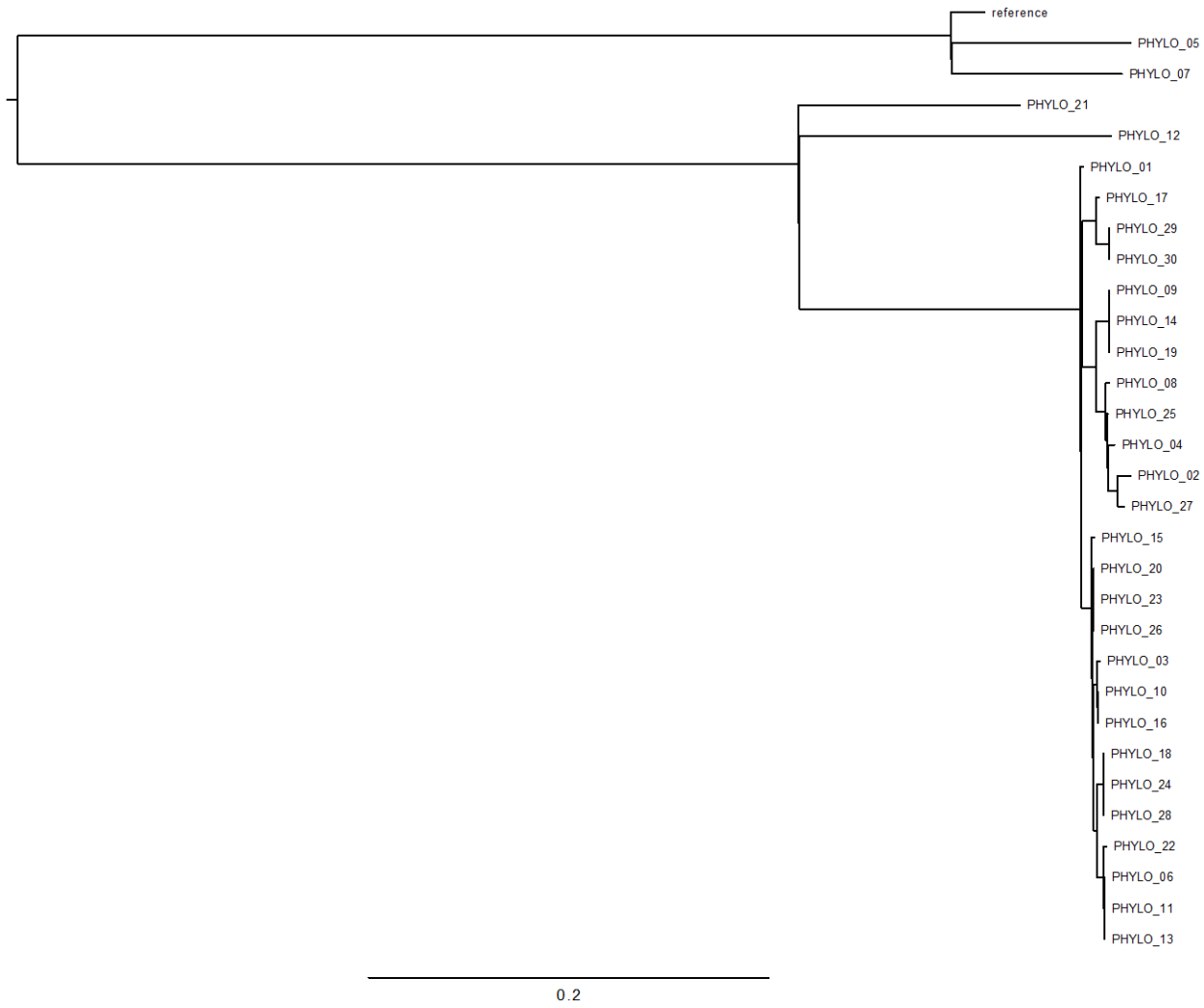
[3] Due to some missing branch lengths in the newick format all the branches appear with the same length.



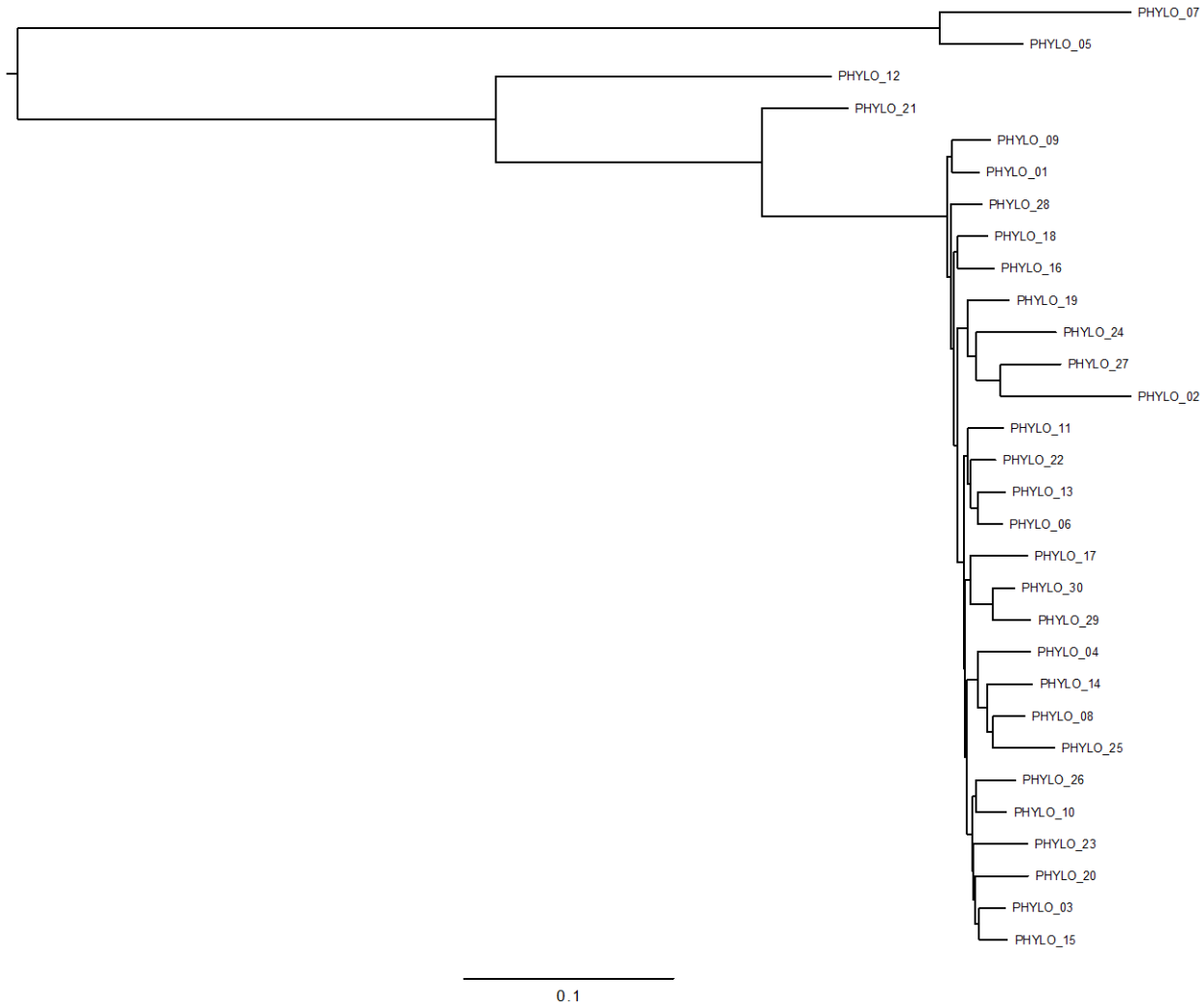
Additional Figure 8. Phylogeny Centre 7 obtained with CSI Phylogeny (CGE tools, online version) (scale represents the branch length stipulated into the newick file)



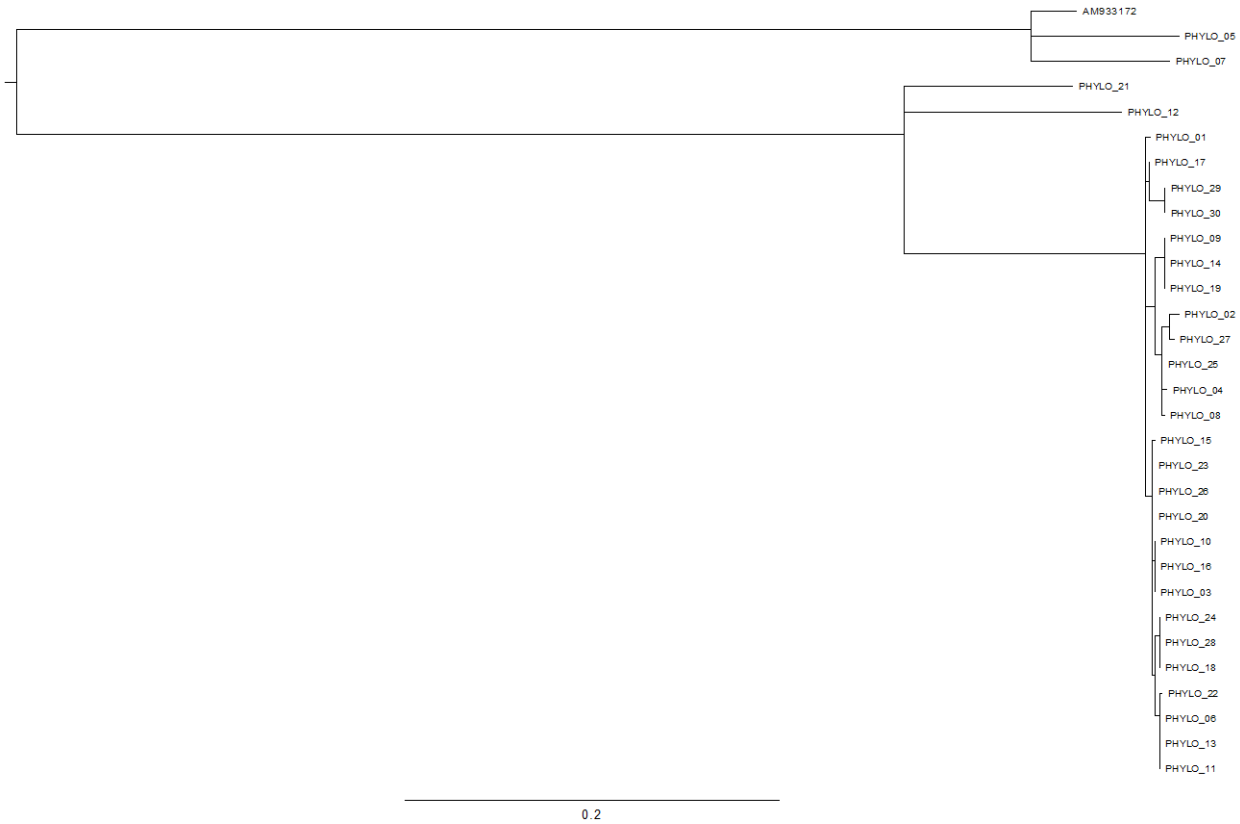
Additional Figure 9. Phylogeny Centre 8 obtained with CSI Phylogeny (CGE tools, online version) with heterozygous SNPs ignored (scale represents the branch length stipulated into the newick file)



Additional Figure 10. Phylogeny Centre 9 obtained with CSI Phylogeny (CGE tools, online version) (scale represents the branch length stipulated into the newick file)



Additional Figure 11. Phylogeny Centre 10 obtained with BWA-Mem mapping, Freebayes, VariantAnnotation for detection/filter of SNPs and VCF-kit to generate the final alignment and the Neighbor Joining Tree (scale represents the branch length stipulated into the newick file)



Additional Figure 12. Phylogeny Centre 11 obtained with PHEnix/SnapperDB and for variants detection and RAxML for tree building (scale represents the branch length stipulated into the newick file)

Additional Figure 13. Partner Centre 12 phylogeny obtained with CSI Phylogeny (CGE tool, online version) (scale represents the branch length stipulated into the newick file)

--- --- ---