Supplementary Materials/Methods

SNVPhyl Parameters

Unless otherwise specified, SNVPhyl analyses for this study were run using the SNVPhyl workflow version 1.0, with paired-end sequence reads, and with parameters:

- Variant quality parameters
  - Minimum coverage = 10X
  - Relative SNV abundance = 0.75
  - Minimum mean mapping quality = 30
- SNV density filtering
  - Window size = 20
  - SNV threshold = 2
- Repeat identification
  - Minimum length = 150
  - Minimum percent identity = 90%

All other parameters were left at the default settings as recorded in the SNVPhyl Galaxy workflow.

Supplementary Figures and Tables

Figure S1 was constructed using the phylogenetic trees produced in the "SNV density filtering evaluation" section of the main manuscript. The tree produced from the "truth" set of SNVs identified using Gubbins is used as a reference tree and is annotated as "Original alignment" in the figure. This reference tree is compared to all other phylogenetic trees from each SNV-density filtering scenario. The comparisons were made using the phytools [34] package in R.

Figure S2 was constructed using the phylogenetic trees produced in the "Parameter optimization" section of the manuscript. The phylogenetic trees were evaluated for concordance with the epidemiological data according to the conditions: 1) all outbreak isolates group monophyletically, and 2) the maximum SNV distance between any two isolates within an outbreak clade must be less than 5 SNVs. These conditions were tested and the figure constructed using the APE [37] package within R.

Table S1 was constructed from previously published strain and accession identifiers [32]. These were mapped to sequencing run accession identifiers using SRAdb [35].

Table S2 was constructed using information obtained from a previous study [36] along with information from the project SRP067504 under the NCBI Sequence Read Archive.

Table S3 was constructed using the SNVs identified by SNVPhyl from the simulated dataset. The copy numbers covering each position were determined by aligning each variant genome to the reference genome using the MUMmer [25] software package. In particular, the command “show-snps” was used to extract the number of alignments, and so copies, covering each position.

Supplementary Figures and Tables

Figure S1.

Figure S2.

Table S1.

Separate file **Table_S1.xlsx**

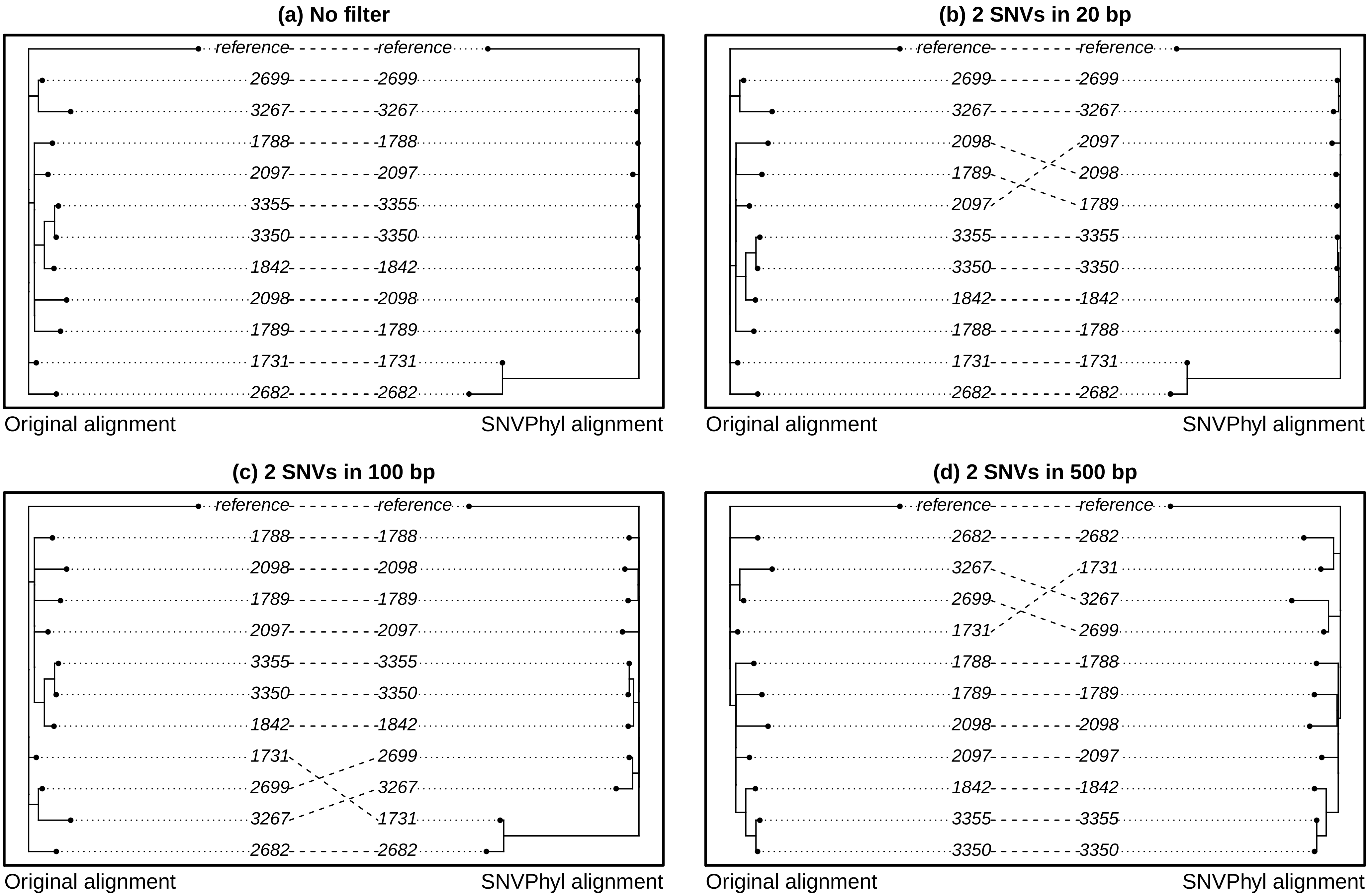
Table S2.

Separate file **Table_S2.xlsx**

Table S3.

Separate file **Table_S3.xlsx**

Figure S1



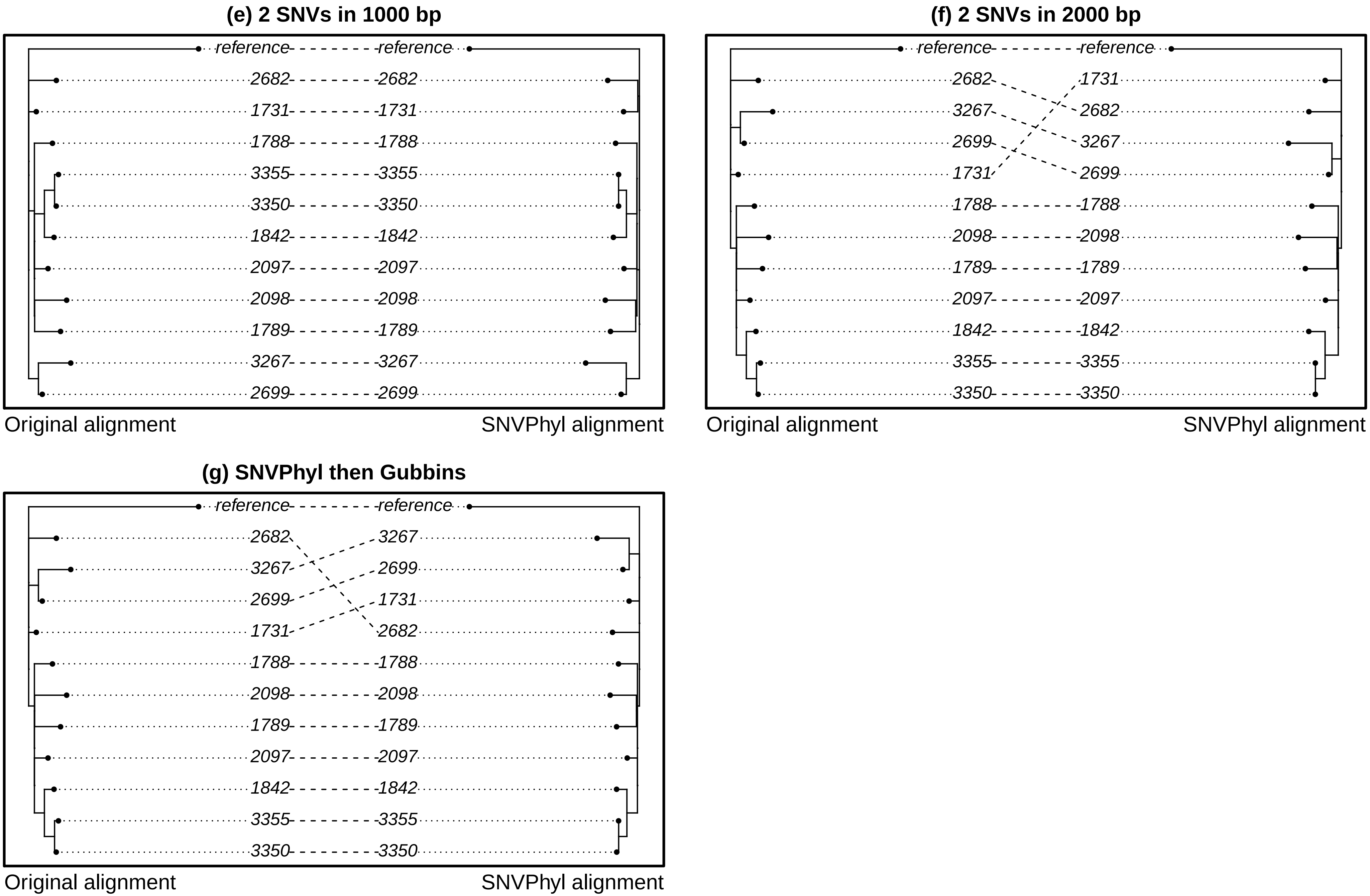
(a) No filter

(b) 2 SNVs in 20 bp

(c) 2 SNVs in 100 bp

(d) 2 SNVs in 500 bp

Figure S1



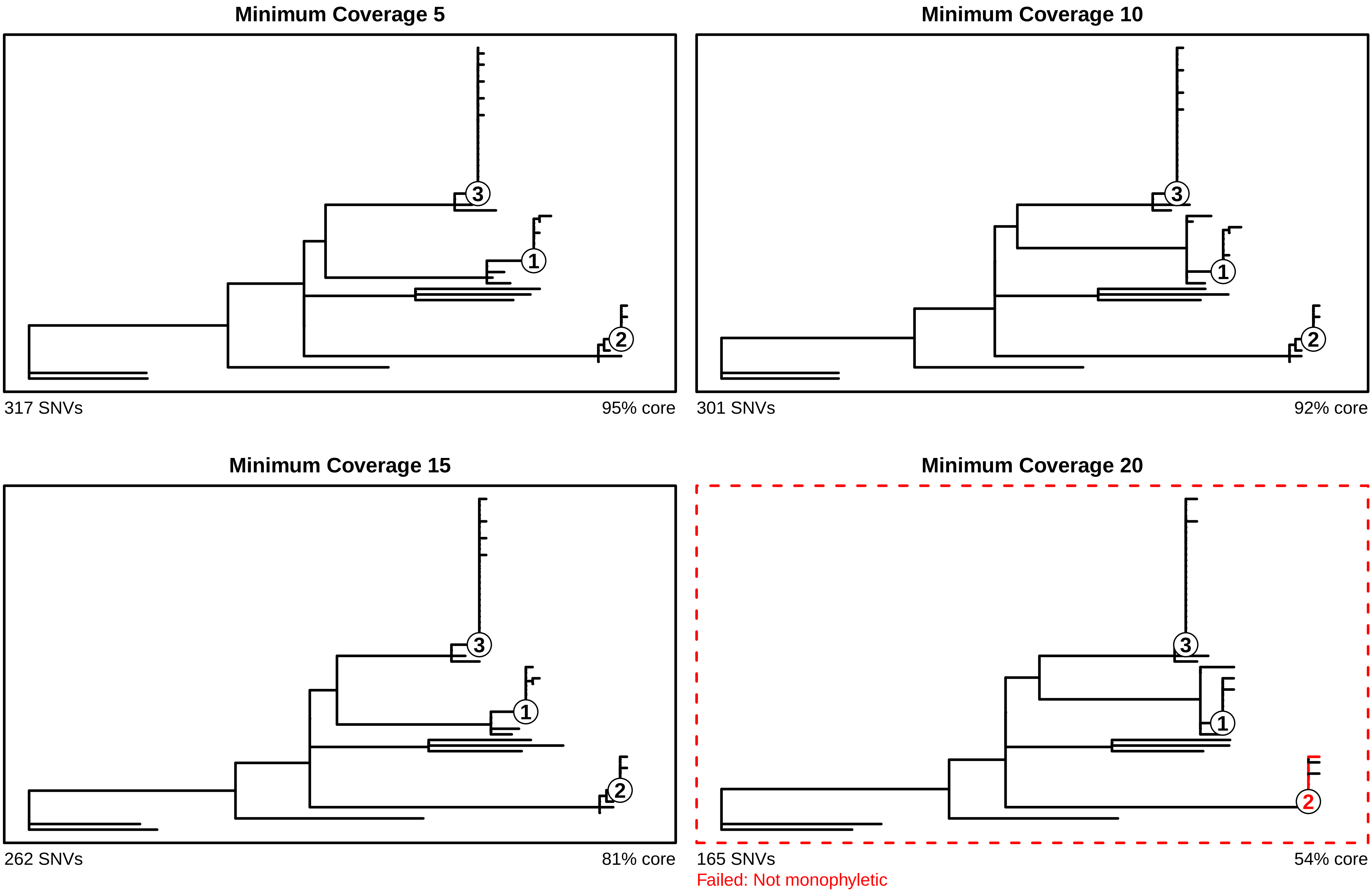
(e) 2 SNVs in 1000 bp

(f) 2 SNVs in 2000 bp

(g) SNVPhyl then Gubbins

Figure S2

(a) Minimum Coverage



Minimum Coverage 5

317 SNVs

95% core

Minimum Coverage 10

301 SNVs

92% core

Minimum Coverage 15

262 SNVs

81% core

Minimum Coverage 20

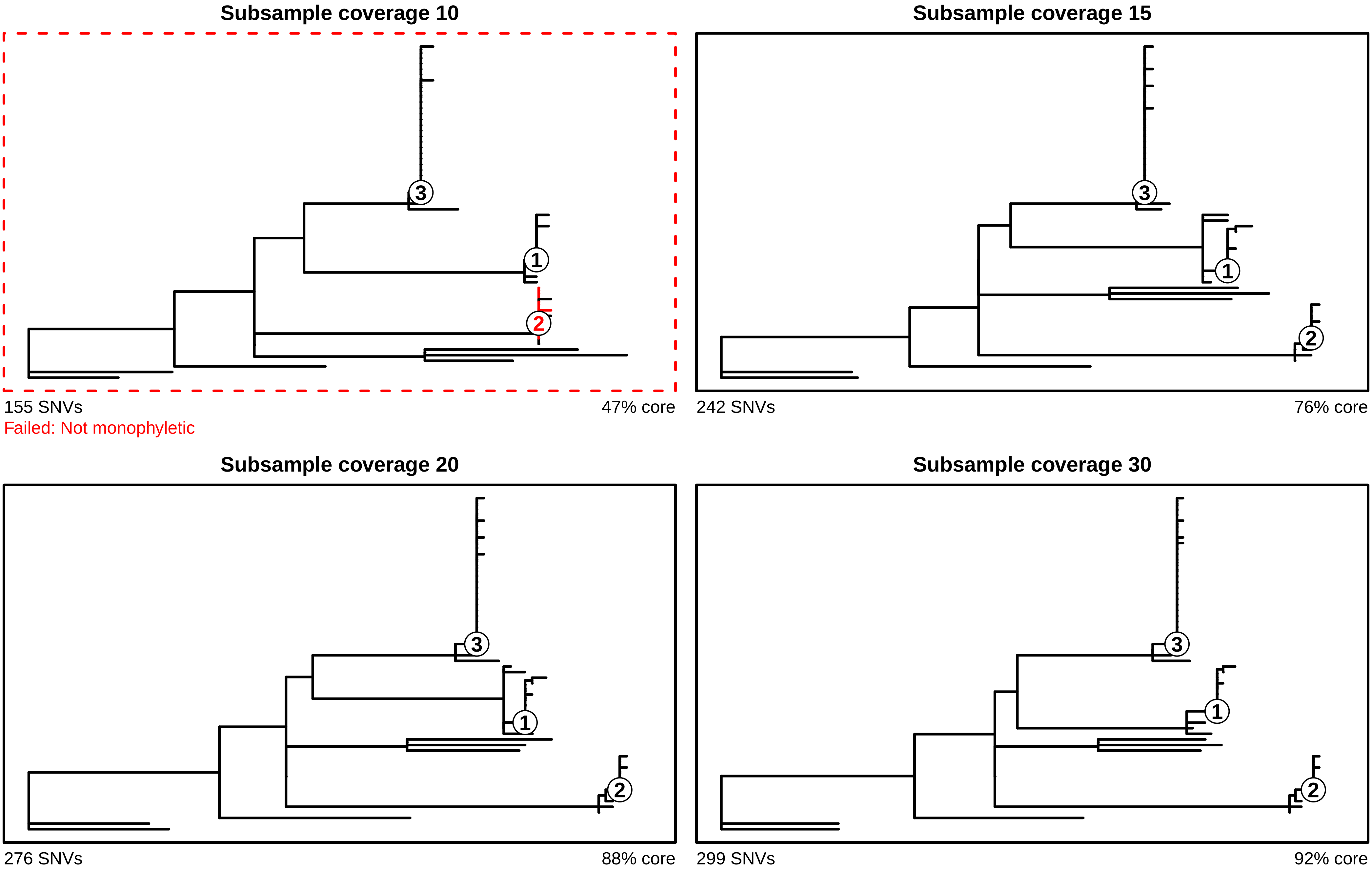
165 SNVs

54% core

Failed: Not monophyletic

Figure S2

(b) Subsample coverage level



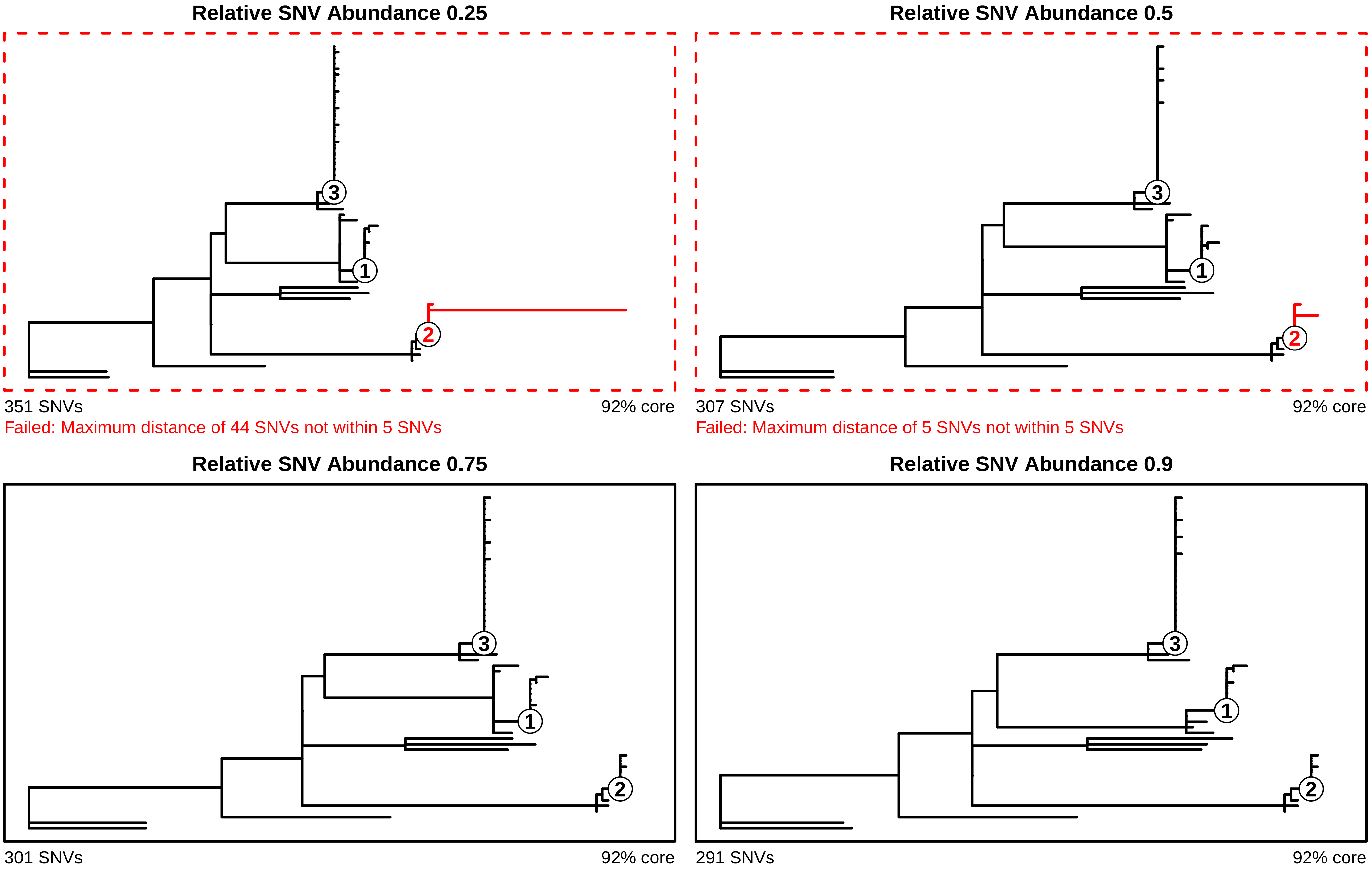
Subsample coverage 10

Subsample coverage 15

Subsample coverage 20

Subsample coverage 30

Figure S2
(c) Relative SNV Abundance



Relative SNV Abundance 0.25

351 SNVs 92% core

Failed: Maximum distance of 44 SNVs not within 5 SNVs

Relative SNV Abundance 0.5

307 SNVs 92% core

Failed: Maximum distance of 5 SNVs not within 5 SNVs

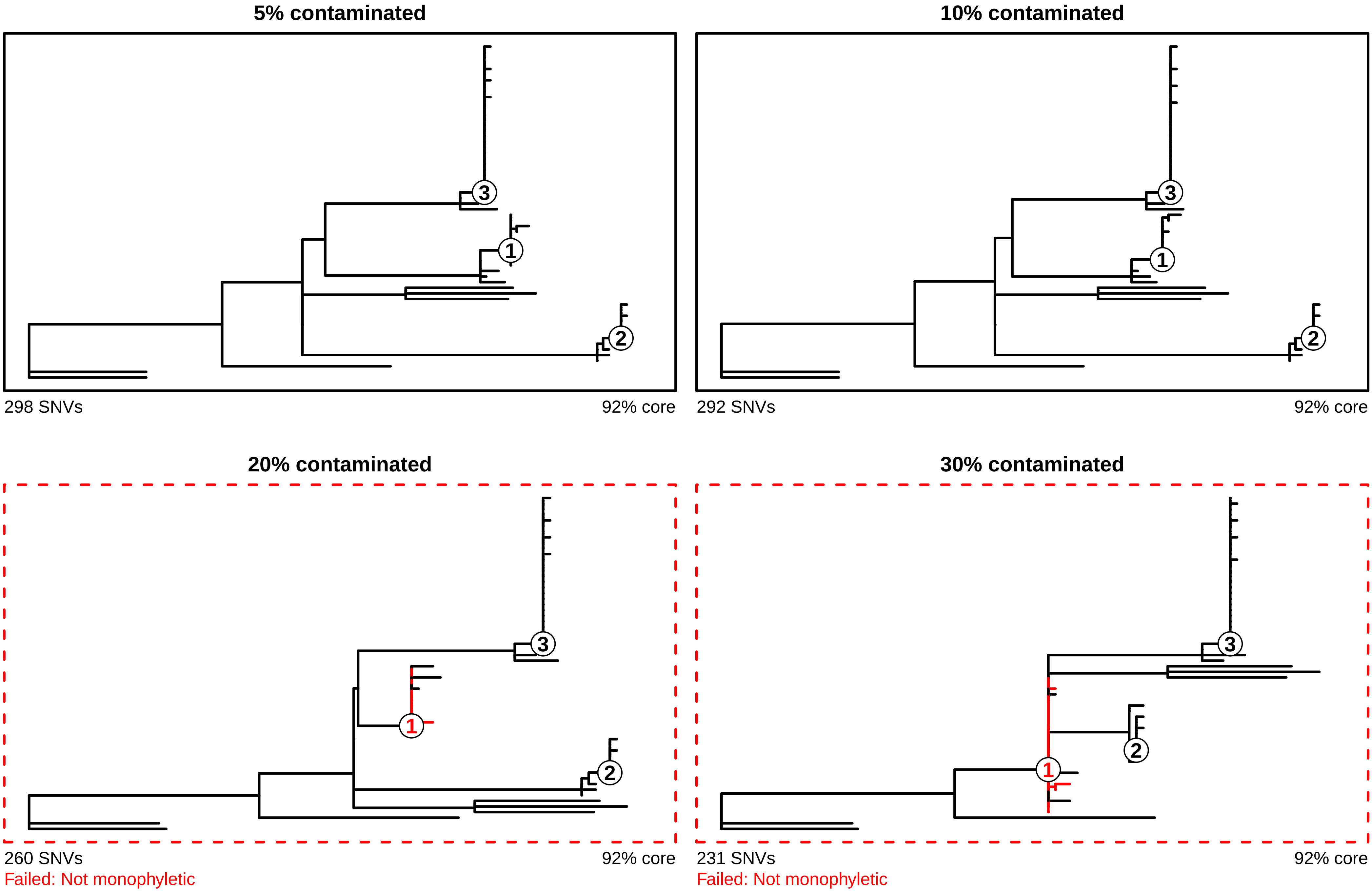
Relative SNV Abundance 0.75

301 SNVs 92% core

Relative SNV Abundance 0.9

291 SNVs 92% core

Figure S2
(d) Contamination



5% contaminated

10% contaminated

20% contaminated

30% contaminated