Sequence analysis and phylogentic context of Firmicutes and Gammaprotobacteria

Kory Melton

10/21/2019

# Background

I was given 40 sequences of 16S ranging from 5-1,310 bases long in ab1 format derived via Sanger sanquencing. The primary aims of this analysis were to assess sequence usability, quality, and length for the 16S sequences, and then place them in phylogenetic context. First I cleaned the sequences in Geneious Prime to remove incorrect or low quality calls from Sanger sequencing. The sequences were cleaned conservatively, especially for phylogentic inference, to ensure the integrity of the data. After cleaning the data, I sorted them into groups based on their quality for further analysis.

I found BLAST matches for the cleaned, acceptable sequences against NCBI’s GenBank database and developed phylogenetic trees through Maximum Likelihood and Bayesian approaches. The Maximum Likelihood approach finds the most probable tree by maximizing a likelihood function. In contrast, the Bayesian approach creates many trees and uses the previous results to make the new tree. I used two contrarian approaches to provide different insights on the sequences. BLAST matches were found for 22 of the 40 sequences. All of the sequences I analyzed were likely to come from the domain Bacteria. I inferred a monophyletic clade of *Staphylococcus* sequences and possibly identified taxa of at least six sequences from the dataset with strong confidence. The six strongest inferences were *Firmicutes* and *Gammaprotobacteria*.

# Methods

I sorted each of the 40 sequences into two categories: useable and unusable. The useable sequences were trimmed and cleaned and sorted further into very good and acceptable. The very good sequences were aligned and used to create two phylogenies. Both the very good and acceptable sequences were matched to the BLAST nucleotide database.

All data files are in the data folder of this project. The raw data files are contained in raw\_data and were not altered during this analysis. Once sequences are analyzed and changed the new files are in sequence\_data. The multiple sequence alignment and the geneious files for PhyML and MrBayes alignments are contained in the alignments folder. Both trees are saved as png images in the images folder.

## Sequence Cleaning

I used Geneious Prime (v 2019.2.1) (Kearse *et al.*, 2012) to trim, clean, and assess the quality of sequences. I trimmed both ends of the sequences to reduce low quality, common to the ends of Sanger sequences. I conducted the trimming conservatively to have higher confidence in downstream analyses. I trimmed bases if they contained any baseline contamination, low peaks, messy peaks, or if the peaks began to span multiple nucleotide positions with only one clear peak. After trimming, I assessed the usability of the sequences, separating them into three groups: very good, acceptable, and unusable. Very good sequences had long stretches (more than 100 base pairs) of high quality reads. I inspected the entirety of the very good sequences to catch any internally miscalled bases and add associated ambiguity codes. Acceptable sequences had mostly high quality reads without long stretches. Poor sequences had some high quality reads, but overall looked messy. Unuseable sequences were either very short (5 base pairs), or had multiple peaks throughout the sequence.

## Sequence Alignment

I used the MAFFT (Katoh *et al.*, 2002) multiple sequence aligner to align the trimmed, very good 16S sequences. The MAFFT settings used were scoring matrix of 200PAM / k=12, gap open penalty of 1.53, and offset value of 0.123. I aligned the six very good sequences with a 16S sequence from *Thermus aquaticus* with accension code NR\_025900.1 (Saul *et al.*, 1993).

## Phylogeny

I used two approaches (Maximum Likelihood and Bayesian) to make phylogenetic inferences of the six very good sequences.

### Maximum Likelihood

I created a Maximum Likelihood phylogeny using the PhyML algorithm (Guindon *et al.*, 2009). This analysis was ran with General Time Reversible (GTR) substitution model, 100 bootstraps, Gamma and Invariants model, four substitution rate categories, and optimized for topology/length/rate. I used the 16S sequence from *Thermus aquaticus* as an outgroup.

### Bayesian

I built a Bayesian phylogeny using the MrBayes 3.6.2 algorithm (Huelsenbeck and Ronquist, 2001). This analysis was ran with General Time Reversible (GTR) model, invgamma rate variation, chain length of 1,100,000, 4 heated chains, heated chain temperature of 0.2, subsampling frequency of 200, burn-in length of 100,000, random seed of 23,107, and unconstrained branch lengths with GammaDir(1, 0.1, 1, 1). I used the 16S sequence from *Thermus aquaticus* as an outgroup.

## BLAST

All sequences rated very good, average, or poor were used to search for BLAST matches in NCBI’s GenBank database (Benson *et al.*, 2000). More specifically, I searched the (nr/nt) database in GenBank. All BLAST searches were done using the highly similar sequences (megablast) algorithm. All matches were selected by choosing which species was most represented at the top of the results list.

# Results

## Sequences and Alignments

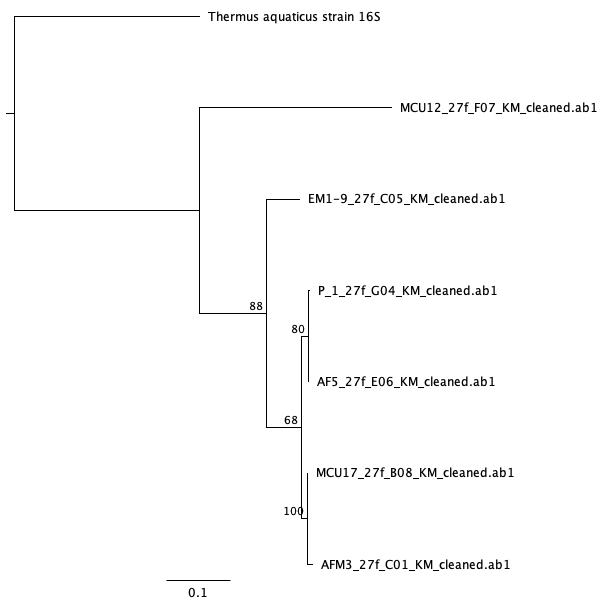
Of the 40 sequences provided 22 were useable with 16 acceptable for BLAST analysis and 6 good enough to place into a phylogeny. The longest sequence was 907 base pairs after trimming while the shortest was 147. Seven sequences were total fails producing only 5 N’s after Sanger sequencing. Another 11 had messy sequences throughout with frequent unidentifiable bases scattered throughout the middle of the sequence. All of the 40 sequences were located in 16S. Table 1 below summarizes the sequence files including original filename, corrected filename, usability, length after trimming, and the number of corrected bases after trimming.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Locus | Original Filename | Corrected Filename | Usability | Length after Trimming | Number of Corrected Bases |
| 16S | AFM3\_27f\_C01.ab1 | AFM3\_27f\_C01\_KM\_cleaned.ab1 | useable | 613 | 3 |
| 16S | P\_1\_27f\_G04.ab1 | P\_1\_27f\_G04\_KM\_cleaned.ab1 | useable | 797 | 0 |
| 16S | MCU17\_27f\_B08.ab1 | MCU17\_27f\_B08\_KM\_cleaned.ab1 | useable | 653 | 1 |
| 16S | AF5\_27f\_E06.ab1 | AF5\_27f\_E06\_KM\_cleaned.ab1 | useable | 404 | 1 |
| 16S | EM1-9\_27F\_C05.ab1 | EM1-9\_27f\_C05\_KM\_cleaned.ab1 | useable | 608 | 5 |
| 16S | MCU12\_27f\_F07.ab1 | MCU12\_27f\_F07\_KM\_cleaned.ab1 | useable | 695 | 0 |
| 16S | PC4A\_27f\_C10.ab1 | PC4A\_27f\_C10\_KM\_trimmed.ab1 | useable | 562 | 1 |
| 16S | PC3A\_27f\_A10.ab1 | PC3A\_27f\_A10\_KM\_trimmed.ab1 | useable | 241 | 0 |
| 16S | CVC1B\_27f\_B03.ab1 | CVC1B\_27f\_B03\_KM\_trimmed.ab1 | useable | 461 | 0 |
| 16S | SB3\_27f\_C06.ab1 | SB3\_27f\_C06\_KM\_trimmed.ab1 | useable | 147 | 0 |
| 16S | P3\_27f\_C08.ab1 | P3\_27f\_C08\_KM\_trimmed.ab1 | useable | 467 | 1 |
| 16S | CVC7A\_27f\_E02.ab1 | CVC7A\_27f\_E02\_KM\_trimmed.ab1 | useable | 621 | 1 |
| 16S | CSB3B\_27f\_F10.ab1 | CSB3B\_27f\_F10\_KM\_trimmed.ab1 | useable | 239 | 0 |
| 16S | CLS6\_27f\_F04.ab1 | CLS6\_27f\_F04\_KM\_trimmed.ab1 | useable | 516 | 0 |
| 16S | FL2A\_27f\_A09.ab1 | FL2A\_27f\_A09\_KM\_trimmed.ab1 | useable | 383 | 0 |
| 16S | Cntrl\_2\_27f\_D05.ab1 | Cntrl\_2\_27f\_D05\_KM\_trimmed.ab1 | useable | 410 | 0 |
| 16S | I5\_27f\_D10.ab1 | I5\_27f\_D10\_KM\_trimmed.ab1 | useable | 268 | 0 |
| 16S | 16A\_C\_27f\_F03.ab1 | 16A\_C\_27f\_F03\_KM\_trimmed.ab1 | useable | 672 | 0 |
| 16S | M3\_2\_27f\_B01.ab1 | M3\_2\_27f\_B01\_KM\_trimmed.ab1 | useable | 258 | 0 |
| 16S | 3A-B\_27f\_B04.ab1 | 3A-B\_27f\_B04\_KM\_trimmed.ab1 | useable | 836 | 1 |
| 16S | CLS2\_27f\_B04.ab1 | CLS2\_27f\_B04\_KM\_trimmed.ab1 | useable | 246 | 0 |
| 16S | PC1B\_27f\_A06.ab1 | PC1B\_27f\_A06\_KM\_trimmed.ab1 | useable | 907 | 0 |
| 16S | CLS1\_27f\_E04.ab1 | CLS1\_27f\_E04\_KM\_unusable.ab1 | unusable | 271 | 0 |
| 16S | UE3B\_27f\_G09.ab1 | UE3B\_27f\_G09\_KM\_unusable.ab1 | unusable | 685 | 0 |
| 16S | EN\_27f\_F02.ab1 | EN\_27f\_F02\_KM\_unusable.ab1 | unusable | 492 | 0 |
| 16S | I6\_27f\_C10.ab1 | I6\_27f\_C10\_KM\_unusable.ab1 | unusable | 746 | 0 |
| 16S | CSB3A\_27f\_E10.ab1 | CSB3A\_27f\_E10\_KM\_unusable.ab1 | unusable | 733 | 0 |
| 16S | M2\_2\_27f\_D01.ab1 | M2\_2\_27f\_D01\_KM\_unusable.ab1 | unusable | 269 | 0 |
| 16S | EVM3\_27f\_C08.ab1 | EVM3\_27f\_C08\_KM\_unusable.ab1 | unusable | 455 | 0 |
| 16S | D6\_27f\_B09.ab1 | D6\_27f\_B09\_KM\_unusable.ab1 | unusable | 317 | 0 |
| 16S | D3\_27f\_G08.ab1 | D3\_27f\_G08\_KM\_unusable.ab1 | unusable | 608 | 0 |
| 16S | CVCWA\_27f\_A02.ab1 | CVCWA\_27f\_A02\_KM\_unusable.ab1 | unusable | 386 | 0 |
| 16S | b1b\_27f\_B02.ab1 | b1b\_27f\_B02\_KM\_unusable.ab1 | unusable | 267 | 0 |
| 16S | IF3\_27f\_E07.ab1 | IF3\_27f\_E07\_KM\_unusable.ab1 | unusable | 5 | 0 |
| 16S | EVN\_27f\_A02.ab1 | EVN\_27f\_A02\_KM\_unusable.ab1 | unusable | 5 | 0 |
| 16S | EVM2\_27f\_G01.ab1 | EVM2\_27f\_G01\_KM\_unusable.ab1 | unusable | 5 | 0 |
| 16S | E1\_27f\_B02.ab1 | E1\_27f\_B02\_KM\_unusable.ab1 | unusable | 5 | 0 |
| 16S | D1\_27f\_E08.ab1 | D1\_27f\_E08\_KM\_unusable.ab1 | unusable | 5 | 0 |
| 16S | CLS5\_27f\_A05.ab1 | CLS5\_27f\_A05\_KM\_unusable.ab1 | unusable | 5 | 0 |
| 16S | APP1\_27f\_G09.ab1 | APP1\_27f\_G09\_KM\_unusable.ab1 | unusable | 5 | 0 |

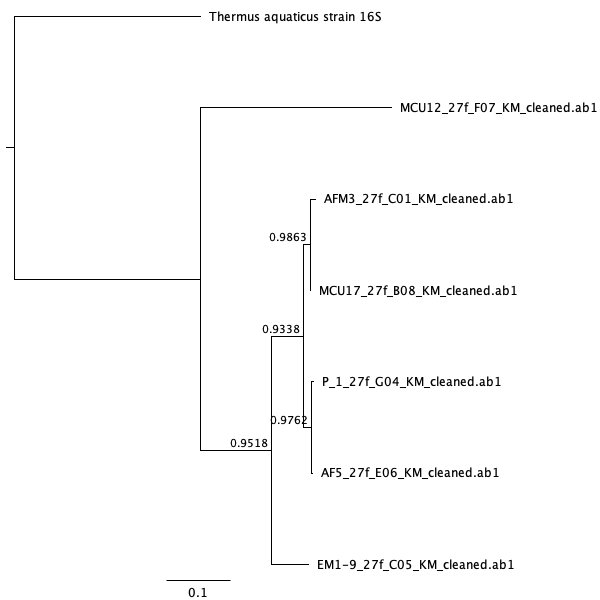
**Table 1.** Short description of each sequence including locus (16S), original filename, filename after edits, usability, sequence length after trimming, and the number of modified bases after cleaning.

## Phylogenies

I found that P\_1\_27f\_G04, AF5\_27f\_E06, AFM3\_27f\_C01, MCU17\_27f\_B08, and EM1-9\_27f\_C05 grouped together (Figure 1; ML bootstrap > 80, Figure 2; Bayesian posterior probability > 95 ). Of this I inferred P\_1\_27f\_G04 and AF5\_27f\_E06 group together (Figure 1; ML bootstrap = 80, Bayesian posterior probaility > 95); and AFM3\_27f\_C01 grouped with MCU17\_27f\_B08 (Figure 1; ML bootstrap > 80, Bayesian posterior probaility > 95). I found a closer relationship between AFM3\_27f\_C01, MCU17\_27f\_G04, P\_1\_27f\_G04 and AF5\_27f\_E06 to EM1-9\_27f\_C05 than MCU12\_27f\_F07 (Figure 1; ML bootstrap > 80, Bayesian posterior probaility > 95). Both maximum likelihood and bayesian phylogenies supported the same groupings for the seven sequences used. Both trees can be seen below in figures 1 and 2.



**Figure 1.** Maximum Likelihood phylogeny of very good sequences using the PhyML algorithm (Guindon *et al.*, 2009). *Thermus aquaticus* was set as an outgroup.



**Figure 2.** Bayesian phylogeny of very good sequences using the MrBayes algorithm (Huelsenbeck and Ronquist, 2001). *Thermus aquaticus* was set as an outgroup.

## BLAST Results

I matched the 22 very good and acceptable sequences to the BLAST database. I found strong matches for 16 of the samples with 5 samples from the genus *Staphylococcus*, three from the genus *Neisseria perflava*, and two *E. coli*. All matches to these samples were bacterial with 7 *Firmicutes* and 6 *Proteobacteria*. Only one sample, 3A-B\_27f\_B04 had a query cover below 90%. Every percent identity was above 80%. This analysis can be seen below in Table 2.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sequence | Description | Percent Identity (%) | Query Cover (%) | Accession | E-value |
| AFM3\_27f\_C01 | *Staphylococcus capitis* partial | 99.51 | 100 | LT978442.1 | 0.0 |
| P\_1\_27f\_G04 | *Staphylococcus hominis strain* WS1-7 | 100.00 | 100 | MN448417.1 | 0.0 |
| MCU17\_27f\_B08 | *Staphylococcus epidermidis* strain 15 | 99.85 | 100 | MK465356.1 | 0.0 |
| AF5\_27f\_E06 | *Staphylococcus hominis* strain WS3-1 | 99.75 | 100 | MN448433.1 | 0.0 |
| EM1-9\_27f\_C05 | *Bacillus licheniformis* strain XIXJ053 | 99.18 | 100 | MH801084.1 | 0.0 |
| MCU12\_27f\_F07 | *Brevundimonas vesicularis* strain A33 | 100.00 | 100 | MN252083.1 | 0.0 |
| PC4A\_27f\_C10 | *Neisseria perflava* strain 17-P17 | 99.82 | 100 | MF374824.1 | 0.0 |
| PC3A\_27f\_A10 | *Neisseria perflava* strain 17-P17 | 99.59 | 100 | MF374824.1 | 1e-119 |
| CVC1B\_27f\_B03 | *Pantoea sp.* strain ST18.17/028 | 100.00 | 100 | MK875130.1 | 0.0 |
| SB3\_27f\_C06 | *Staphylococcus epidermidis* strain IBK-11 | 100.00 | 100 | MN428237.1 | 3e-69 |
| P3\_27f\_C08 | *Bacillus licheniformis* strain Xuyi\_293\_3 | 98.93 | 100 | MN309987.1 | 0.0 |
| CVC7A\_27f\_B03 | *Enterobacteriaceae bacterium* X6/SB50 | 98.87 | 100 | LC007901.1 | 0.0 |
| CSB3B\_27f\_F10 | *Klebsiella granulomatis* strain 2-7-8 | 100.00 | 99 | JF91986.1 | 4e-119 |
| CLS6\_27f\_F04 | Uncultured bacterium partial | 100.00 | 91 | FM162220.1 | 5e-17 |
| FL2A\_27f\_A09 | *Steptococcus salivarius* strain JF | 98.96 | 100 | CP014144.1 | 0.0 |
| Cntrl\_2\_27f\_D05 | *Escherichia coli* stain NCTC9040 | 97.80 | 100 | LR134247.1 | 0.00 |
| I5\_27f\_D10 | *Pseudomonas sp.* strain 273 | 94.12 | 100 | MK681820.1 | 3e-111 |
| 16A\_C\_27f\_F03 | *Escherichia coli* strain SJA92 | 92.12 | 100 | MN147836.1 | 0.0 |
| M3\_2\_27f\_B01 | *Bacillus licheniformis* strain FMB2-2 | 96.86 | 98 | KP992871.1 | 1e-115 |
| 3A-B\_27f\_B04 | *Pseudomonas sp.* m8 | 83.46 | 33 | DQ219815.1 | 6e-58 |
| CLS2\_27f\_B04 | Uncultured *Paenibacillus sp.* clone R7SP1T1H06 | 94.63 | 98 | HQ018234.1 | 2e-103 |
| PC1B\_27f\_A06 | *Neisseria sp.* strain CMUL044 | 85.55 | 96 | MK258031.1 | 4e-168 |
| CLS1\_27f\_E04 | Sequencing Failed |  |  |  |  |
| UE3B\_27f\_G09 | Sequencing Failed |  |  |  |  |
| EN\_27f\_F02 | Sequencing Failed |  |  |  |  |
| I6\_27f\_C10 | Sequencing Failed |  |  |  |  |
| CSB3A\_27f\_E10 | Sequencing Failed |  |  |  |  |
| M2\_2\_27f\_D01 | Sequencing Failed |  |  |  |  |
| EVM3\_27f\_C08 | Sequencing Failed |  |  |  |  |
| D6\_27f\_B09 | Sequencing Failed |  |  |  |  |
| D3\_27f\_G08 | Sequencing Failed |  |  |  |  |
| CVCWA\_27f\_A02 | Sequencing Failed |  |  |  |  |
| b1b\_27f\_B02 | Sequencing Failed |  |  |  |  |
| IF3\_27f\_E07 | Sequencing Failed |  |  |  |  |
| EVN\_27f\_A02 | Sequencing Failed |  |  |  |  |
| EVM2\_27f\_G01 | Sequencing Failed |  |  |  |  |

**Table 2.** BLAST results for each sequence including a description of the match, percent identitiy, query cover, accession number, and e-value.

# Discussion

## Overview of Findings

Of the 40 sequences I was given, 18 were unusable, 16 were acceptable, and 6 were very good high quality reads. The unusable sequences resulted from a variety of reasons including: insufficient template, template contamination, and camera blockage. All 22 of the acceptable and high quality reads had BLAST matches with the GenBank database. All samples were bacterial with 7 *Firmicutes* and 6 *Proteobacteria*. Both the Maximum Likelihood and Bayesian approach produced the same phylogeny, providing strong evidence the correct relationships were identified for the 6 very strong sequences. In addition, the sequences with BLAST matches to *Staphylococcus* species grouped together in a monophyletic clade.

## Unuseable Sequences

There are a few possible reasons for the unuseable sequences in this group of 40 sequences. The 5 sequences in table 1 with only 5 base pairs did not have one element necessary for Sanger sequencing. The most likely cause is insufficient template. The sequences from b1b\_27f\_B02, UE3B\_27f\_G09, D6\_27f\_B09, and D6\_27f\_B09 had multiple peaks throughout suggesting there was template contamination in the sample. CVCWA\_27f\_A02 had messy and low peaks which might be due to poor quality template. D3\_27f\_G08, CLS1\_27f\_E04, UE3B\_27f\_G09, and M2\_2\_27f\_D01 had sharp high-intensity spikes mixed throughout the sequence. One possible cause for this is a blockage of the camera used to sequence the sample, causing light to scatter. EVM3\_27f\_C08 had messy peaks which spanned multiple nucleotides suggesting poor template. I6\_27f\_C10 and EN\_27f\_F02\_KM had a repeated pattern every 20-30 base pairs suggesting the sample was ran through PCR without template.

## BLAST

AFM3\_27f\_C01 is likely to be in the genus *Staphylococcus*, however the specific species could not be identified with high confidence. The top hit on BLAST was *Staphylococcus capitis*, while *Staphylococcus epidermidis* was also highly prevalent. PC4A\_27f\_C10 is likely to be in the genus *Neisseria*, but the species could not be identified with high confidence. The species *perflava*, *flavescens*, and *subflava* were all highly rated matches on BLAST. CVC1B\_27f\_B03 is likely a gram-negative bacterium and could either be in the genus *Pantoea* or the class Gammaproteobacteria. CSB3B\_27f\_F10 is likely to be in the family Enterobacteriaceae, falling under the *Klebsiella* or *Raoultella* genus. CLS6\_27f\_F04 and CLS2\_27f\_B04 both had relatively few matches (only 8) in the GenBank database. 3A-B\_27f\_B04 was highly likely to be a *Pseudomonas sp.*. All other analyzed sequences were matched with strong confidence. Further sequencing and phylogentic inferences should be used to strengthen the identifications of all of the BLAST matches.

## Phylogenetic Context

Based on phylogenetic and BLAST results I inferred the identity of the top six sequences with strong confidence. AFM3\_27f\_C01, MCU17\_27f\_G04, P\_1\_27f\_G04 and AF5\_27f\_E06 were grouped together in a monophyletic clade in both trees. BLAST analysis matched all of these sequences with different species of *Staphylococcus*. In total, this provides strong evidence to suggest these are closely related species of *Staphylococcus*. In addition, P\_1\_27f\_G04 and AF5\_27f\_E06 were grouped together in both phylogenies and BLAST results matched two different strains of *Staphylococcus hominis*. AFM3\_27f\_C01 and MCU17\_27f\_B08 also grouped together in both phylogenies with BLAST results matching *S. capitis* and *S. epidermidis* respectively. These matching results indicate two species *Staphylococcus* which are more related to each other than *Staphylococcus hominis*.

The other two sequences added to the phylogenies are more distantly related with BLAST matches outside of *Staphylococcus*. It is likely the *Brevundimonas* genus is a more recent common ancestor to *Staphylococcus* than the *Bacillus* genus. In fact, previous work shows *Staphylococcus* and *Bacillus* are both *Firmicutes* while *Brevundimonas* is a *Gammaproteobacteria* (Costa *et al.*, 2012). Future work should be done to resolve a larger phylogeny with the other bacterial sequences in the larger dataset these 40 sequences were pulled from, especially with other *Firmicutes*. An additional tree with *proteobacteria* would also be very informative for the other samples in this dataset.

# Sources Cited

Benson,D.A. *et al.* (2000) GenBank. *Nucleic acids research*, **28**, 15–18.

Costa,P.S. *et al.* (2012) Phylogeny of cultivable heterotrophic bacteria derived from mixed colonies. *Aquatic Microbial Ecology*, **66**, 95–106.

Guindon,S. *et al.* (2009) Estimating maximum likelihood phylogenies with phyml. In, *Bioinformatics for dna sequence analysis*. Springer, pp. 113–137.

Huelsenbeck,J.P. and Ronquist,F. (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, **17**, 754–755.

Katoh,K. *et al.* (2002) MAFFT: A novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic acids research*, **30**, 3059–3066.

Kearse,M. *et al.* (2012) Geneious basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, **28**, 1647–1649.

Saul,D. *et al.* (1993) Phylogeny of twenty thermus isolates constructed from 16S rRNA gene sequence data. *International Journal of Systematic and Evolutionary Microbiology*, **43**, 754–760.