"CRISPR-associated protein 1 (Cas1)"

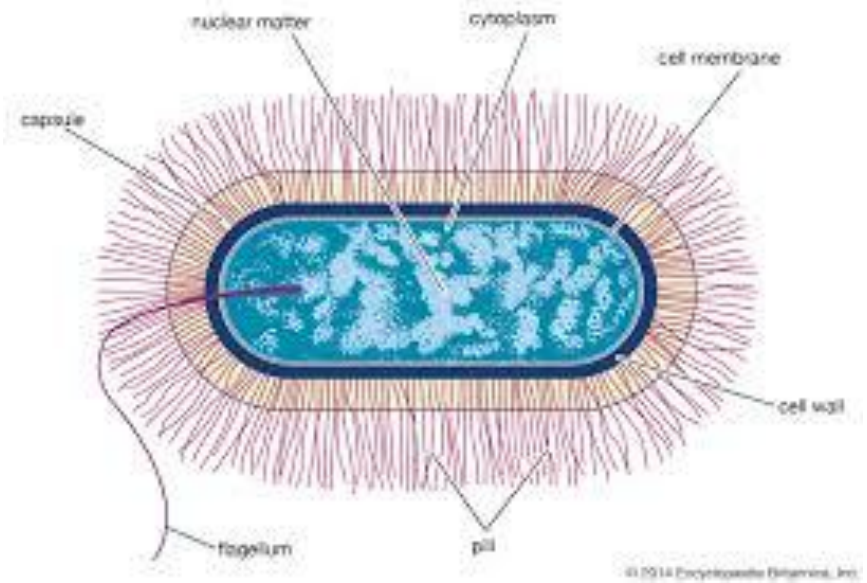
Molecular phylogenetics course project

Sedreh Nassirnia

Fall 2019

What is the story?

- Viruses catch bacterial cell in order to propagate
  - results in bacterial cell lysis and death
- Clustered regularly interspaced short palindromic repeat (CRISPR) and CRISPR associated (Cas) proteins immune defense system is the solution of bacteria to fight with viruses

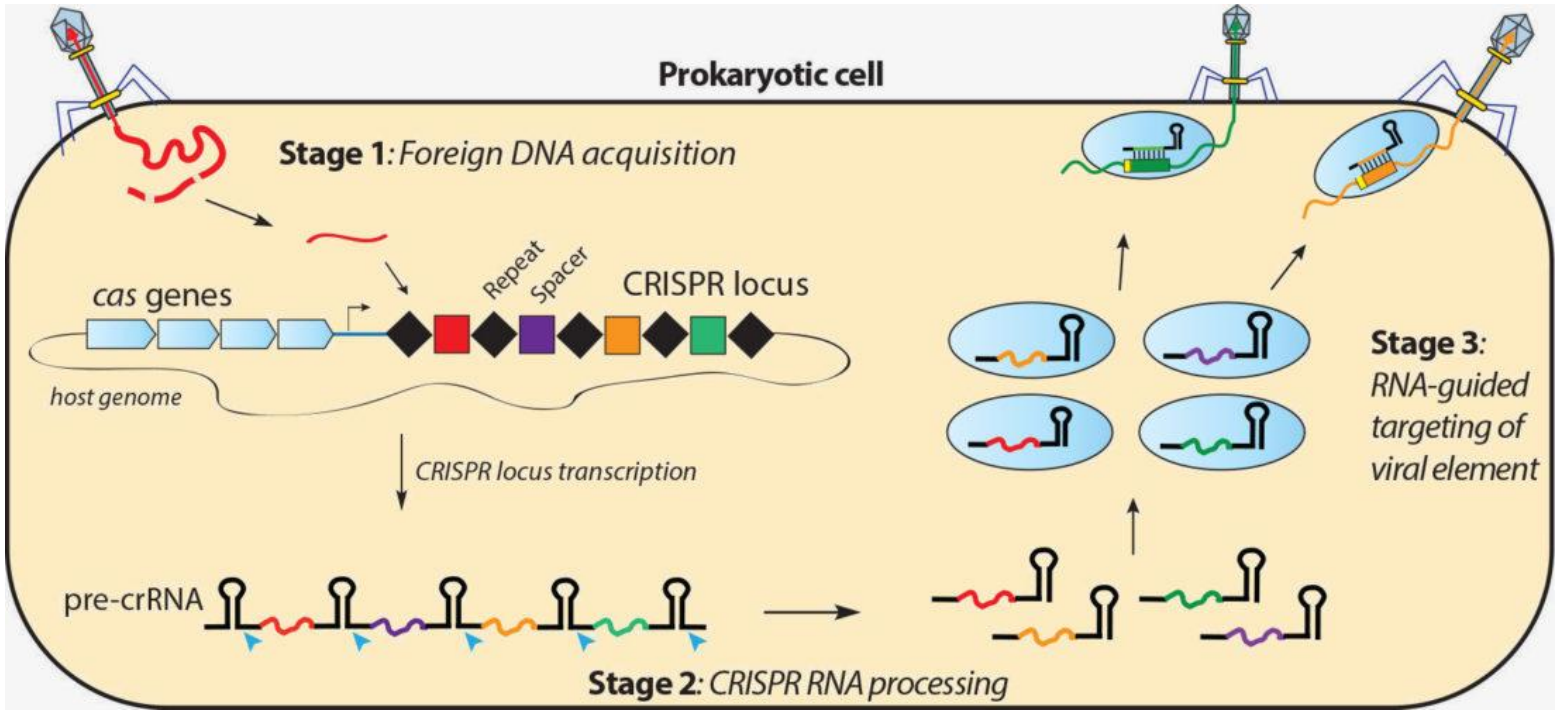


CRISPR-Cas Systems

- CRISPR (clustered regularly interspaced short palindromic repeat)
  - adaptive immune system that provides protection against mobile genetic elements such as genomic islands, plasmids, and transposon-like elements and viruses
- Important for
  - clinical microbiologists, ecologists and evolutionary biologists
  - CRISPR-Cas system potential uses
    - Detection and genotyping of microbial pathogens
    - Host identification in metagenomes
    - Analysis of viral genomes
    - Targeted genome engineering in both prokaryotic and eukaryotic cells

CRISPR associated protein Cas1

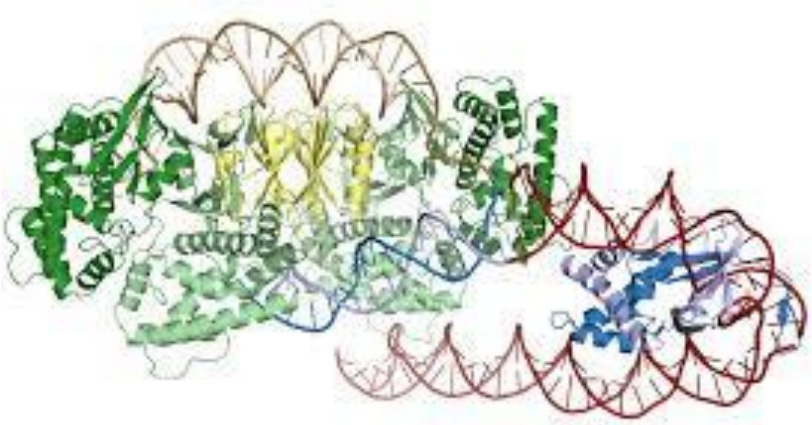
- Cas1 responsible for the ability of the CRISPR immune system in bacteria to adapt to new viral infections
  - Identify the site in the genome where they insert viral DNA



Objective

- **Study evolutionary relationships among bacteria family based on cas1**

- **What percentage of organisms evolved according to viral strains**



Data

KEGG **Azoarcus sp. DD4: CJ010_02280** Help

Entry	CJ010_02280 CDS T06048
Definition	(GenBank) subtype I-C CRISPR-associated endonuclease Cas1
KO	K15342 CRISP-associated protein Cas1
Organism	azr Azoarcus sp. DD4
Brite	KEGG Orthology (KO) [BR:azr00001] 09180 Brite Hierarchies 09182 Protein families: genetic information processing 03400 DNA repair and recombination proteins [BR:azr03400] CJ010_02280 09183 Protein families: signaling and cellular processes 02048 Prokaryotic defense system [BR:azr02048] CJ010_02280 DNA repair and recombination proteins [BR:azr03400] Prokaryotic type DSBR (double strand breaks repair) HR (homologous recombination) Other HR factor CJ010_02280 Prokaryotic defense system [BR:azr02048] CRISPR-Cas system Universal Cas proteins CJ010 02280 BRITE hierarchy
SSDB	Ortholog Paralog Gene cluster GFIT
Motif	Pfam: Cas_Cas1 Motif
Other DBs	NCBI-ProteinID: QDF95462 UniProt: A0A4Y6KN73
LinkDB	All DBs
Position	463572..464612 Genome map
AA seq	346 aa AA seq DB search MRRQLNTLYVTTEGAWLKKDGANIVMEVEGAERARLPVHMLESMVCMGRVMVSPPLLGYC AEQGICVSFLSPNGKFLARMEGPVSGNVLLRREQYRRTDDPARCGLVVRNLLIGKVHNQR AVLGRALRDHGEGMPEADQTALAHARERLRRISARLLLEEKLDVLRGLEGEAAQAYFGVF DHLIRVPEPALRFKGRSRRPPLDAVNALLSFLYTLLTHDCRSALESVGLDPAVGFLHRDR PGRPSLALDLLEEFRPVMADRLALSLINRRQLGERDFVQLDNGAVSLKEESRKTVLTAYQ ERKREEMRHAFLEEKFAVGLFPAVQAQLFARHLRGDLEAYPPFLWK
NT seq	1041 nt NT seq +upstream 0 nt +downstream 0 nt atgcgccgccagctcaacaccctgtatgtgaccaccgaaggcgcctggctgaagaaggat ggcgccaacatcgtcatggaggtggaaggggcggaacgcgcccgtttgccggttcacatg ctggagagcatggtgtgtatggggcgggtgatggtgtcgccgccactgctcggctactgc gccgagcagggcatctgcgtcagcttcctctgcccaacggcaagttcctggcacgcatg gaggggccggtttctggcaacgtgctgctacggcgcgagcaataccgccgtaccgatgat ccggcccgctgtggtctagtcgtccgtaatctgttgatcggcaaggtacataaccaacgc gcggtgcttggccgtgcgctgcgcgatcacggtgagggcatgccggaagccgatcagacc gccctggcgcacgcccgggagcgcctgcggcgcatctccgcgcgcttgttgctggaagag aagctggatgtgttgcgtggcctggaaggagaggcggcgcaggcctatttcggcgtcttc gatcacctcatccgtgtgccggagcctgccttgcgctttaagggccgcagccgccggccg ccactggacgcagtgaatgccttgctctccttcctctacacgttgctgacgcatgactgc

Downloading directly from Pfam doesn't give information about Genus and Species

Link database was used to fetch this information.

Example:

https://www.kegg.jp/entry/azr:CJ010_02280

Fetching IDs

```
import os
import urllib
from bs4 import BeautifulSoup as bs
import sys
import requests
import re

def write_fasta(genus, species, ids, seq):
    with open('/home/sedreh/ITMO/semester3/Molecular_phylogenetic/COURSE_PROJECT/phyloproject_cas1/cas1_pfam.fasta'
        fasta.write(">{} {} {}\n{}\n".format(genus, species, ids, seq))

def fetch_fasta(link):
    page = requests.get(link)
    soup = bs(page.content, 'html.parser')
    children= (list(soup.children)[2]).get_text()
    name = re.findall("(?<=\\xa0\\xa0\xa0)(.*?)(?=\:)", children)[0]
    genus, species=format(re.sub(r'[\d-]', '', name)).split(" ",1)
    species.strip(" ")
    ids = re.findall("(?s)(?<=UniProt:\\xa0)(.*?)(?=\\n)", children)
    seq = re.findall("(?s)(?<=aa \\n)(.*?)(?=\\nNT seq\\n)", children)
    seq = format(re.sub('\n', '', seq[0]))

    return write_fasta(genus, species, ids, seq)

def main():
    list_of_links = list()
    with open('/home/sedreh/ITMO/semester3/Molecular_phylogenetic/COURSE_PROJECT/phyloproject_cas1/links.txt', 'r')
        for link in links:
            list_of_links.append(link.strip('\n'))

    for link in list_of_links:
        fetch_fasta(link)
```

```
aaa:Acav_0268          K15342 CRISP-associated protein Cas1 | (GenBank) CRISPR-associated protein Cas1
aaa:Acav_3874          K15342 CRISP-associated protein Cas1 | (GenBank) CRISPR-associated protein Cas1
aac:Aaci_2651          K15342 CRISP-associated protein Cas1 | (GenBank) CRISPR-associated protein Cas1
aacn:AANUM_1357        K15342 CRISP-associated protein Cas1 | (GenBank) CRISPR-associated Cas1 family protein
aacn:AANUM_1920        K15342 CRISP-associated protein Cas1 | (GenBank) cas1-2; CRISPR-associated protein cas1
aact:ACT75_00725       K15342 CRISP-associated protein Cas1 | (GenBank) type I-C CRISPR-associated endonuclease Cas1
aact:ACT75_02270       K15342 CRISP-associated protein Cas1 | (GenBank) type I-F CRISPR-associated endonuclease Cas1
aad:TC41_2954          K15342 CRISP-associated protein Cas1 | (GenBank) CRISPR-associated protein Cas1
aae:aq_369             K15342 CRISP-associated protein Cas1 | (RefSeq) hypothetical protein
aal:EP13_09450         K07486 transposase | (GenBank) transposase
aalg:AREALGSMS7_00764 K03832 periplasmic protein TonB | (GenBank) transport protein TonB
aan:D7S_00182          K15342 CRISP-associated protein Cas1 | (GenBank) CRISPR-associated protein Cas1
aan:D7S_00548          K15342 CRISP-associated protein Cas1 | (GenBank) CRISPR-associated protein Cas1
aao:ANH9381_1474       K15342 CRISP-associated protein Cas1 | (GenBank) CRISPR-associated protein Cas1
aap:NT05HA_0335        K15342 CRISP-associated protein Cas1 | (GenBank) crispr-associated protein Cas1
aaqu:D3M96_05980       K15342 CRISP-associated protein Cas1 | (GenBank) cas1f; type I-F CRISPR-associated endonuclease Cas1
aar:Acear_0821         K15342 CRISP-associated protein Cas1 | (GenBank) CRISPR-associated protein Cas1
aat:D11S_1150          K15342 CRISP-associated protein Cas1 | (GenBank) CRISPR-associated protein Cas1
```

Cleaning data in R

- Raw data
  - 1413 organisms
  - More than 4000 sequences

- Clean and filtered data
  - Each genera must contain at least 3 entries

Final dataset:
- 47 genera
- 320 entries

```
>Capnocytophaga stomatis
MLYRSIYIGNPAYLKLKDQQMKIVCPETKAEKGSVPVEDLGLLMLDHFQITISHQLIQWLMGNNVVIISCDAHHLPHGQMLPLHGNAIYSQRIKDQIEASEPLKKQLWKQTIEC
>Corynebacterium striatum
MAYSEDAITFSTIPADHQVRLEDRVSFAYVEHAAIRQDRTGVVAYSVVDNSELEQRIQLPVGGLAVLMLGPGTSISAAAATSCTRSGTTIMFTGGGGVPAYTHAASLTSSARWA
>Candidatus Caldiarchaeum
MSELVIDKPGTYLGVRKGLFVVRTKGGGRSEFSPVELSHISIRCRGVGVSVDALRLACRFGIEVSVYSRGRPVGKVVGAFLGGGAVTRRAQLEAWGTERGLAVAREIVSAKLYN
>Corynebacterium singulare
MTTPHEVPLTRQALARVGDRISFLYAERCVINRDGNSLTIVDQRGTAHVPATQIAALLLGPGTKITYAAMALLGDAGVSAVWVGERGVRYYAHGRPPAKSSRMAEIQAEVVTHQ
>Clostridium tetani
MKRSYYIYNNGILKRKDNSMAFIDELGERRYIPIETANEIYVMSEMDFNTSLINYLSQYDVIIHFFNYYSFYTGSFQPRKKLVSGNLLVNQVNHYSDNSKRLEIAKKFVDGASY
>Chlorobaculum tepidum
MKKHLNTLFVTTQGSYLSKEGECVLISIDRVEKTRIPLHMLNGIVCFGQVSCSPFLLGHCAQLGVAVTFLTEHGRFLCQMQGPVKGNILLRRAQYRMADNYDQTATLARLFVIG
>Corynebacterium terpenotabidum
MNKIPFRSSVTTQGRNASGARSLWRATGMTDEDFEKPIIAVANSYTQFVPGHVHLKNVGDIVAEAVKEAGGVAREFNTIAVDDGIAMGHSGMLYSLPSREIISDSVEYMVNAHQ
>Hungateiclostridium thermocellum
MKKSAFIFSDGELKRKDSTVLFESEDSKNYLPIEDISDIYIFGEVTVTKKFLELATQKEILLHFYNYNEYYVGTYYPREHYNSGFMILKQAEHYLDEEKRMAIAKKFIHGSVKN
>Cronobacter turicensis
MSFVPLNPIPLNDRTSMIFLQYGHLDVLDGAFVLVDKTGVRTHVPVGAIACIMLEPGTRVSHAAIRLASQVGTLLVWVGEAGVRLYASGQPGGARADKLLYQAKLALDETLRLK
>Corynebacterium urealyticum
MRTPQQVPIERQSLSQMGDRISFLYVERAVVSRDGNALTVTDQRGVAHVPATQLAALLLGTGTRITNAAIALLGDSGVSTVWVGERGVRYYAHGRPPAKSSRLAELQARVVTNQ
>Corynebacterium ulcerans
MSYSNEALAFSTIPASEQIRLEDRVSFLYLEYCLIRQDRTGVIAVSRGDEKAPAELKDLPIKARIQLPVGGLAVLMLGPGTSISQPAATSCARAGVSVLFTGGGGVQAYSLSTP
>Corynebacterium ureicelerivorans
MGSRISFLYIERATVNRDGNALTITDQRSVAHVAATQLAVLLLGPGTRITYAAMALLGDAGVSIVWVGERGVRYYASGRPPAKSSRMAELQAEIVTNQRKRLACAKRMYSLRFP
```

Method

Step 1: All sequences aligned using Muscle multiple sequence alignment

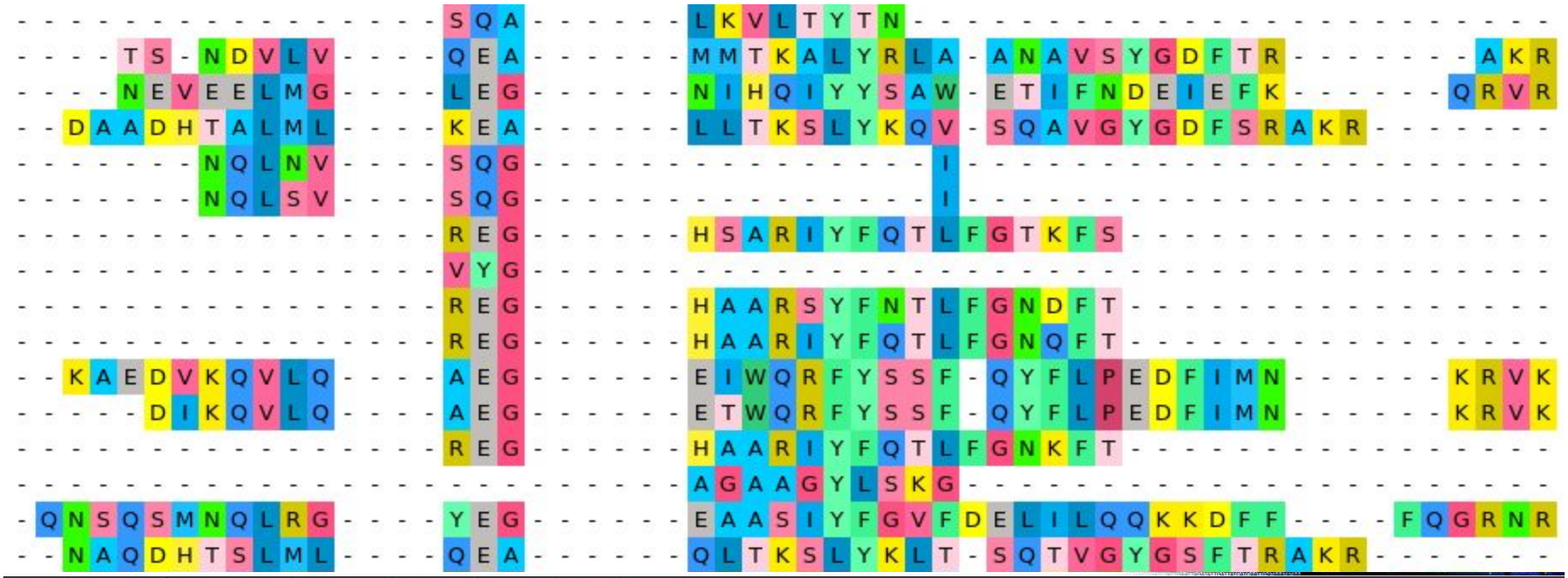
Step 2: Model selection for aligned data using IQ-tree

Step 3: Bootstrapping for 1000 trees in IQ-tree

Step 4: Root the tree and collapse clades with bootstrap support < 95%

Step 5: Analysing the results

Results-Alignment and GBLOCKS



Home | Phylogeny Analysis | Blast: Sequence Explorer | Online Programs | Your Workspace | Documentation | Downloads | Contacts

A new version of phylogeny.fr is available.
Please try it: NGphylogeny.fr.

Gblocks 0.91b (doc)

1. Overview | 2. Data & Settings | 3. Results

Curation: Gblocks

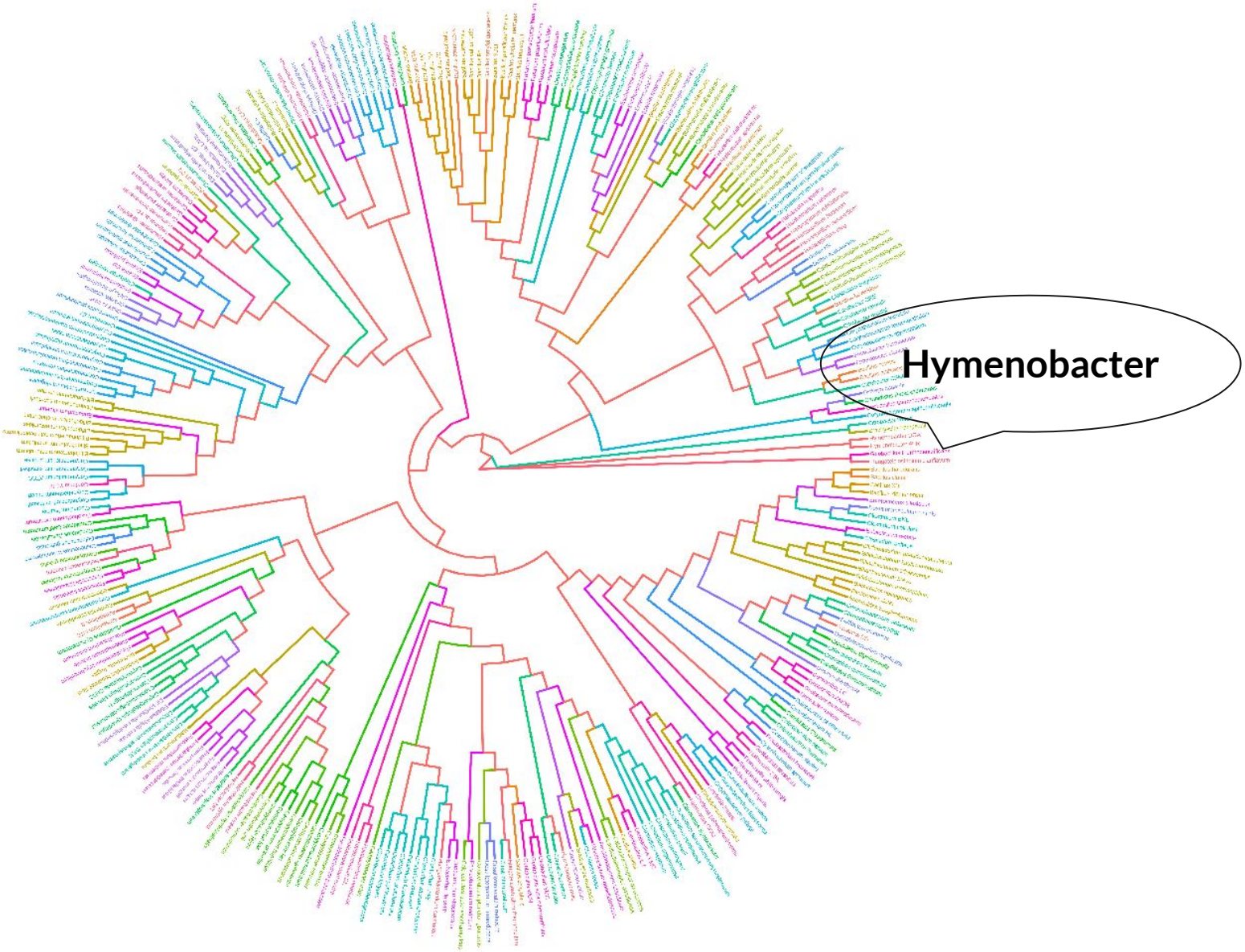
Input:

Input raw alignment

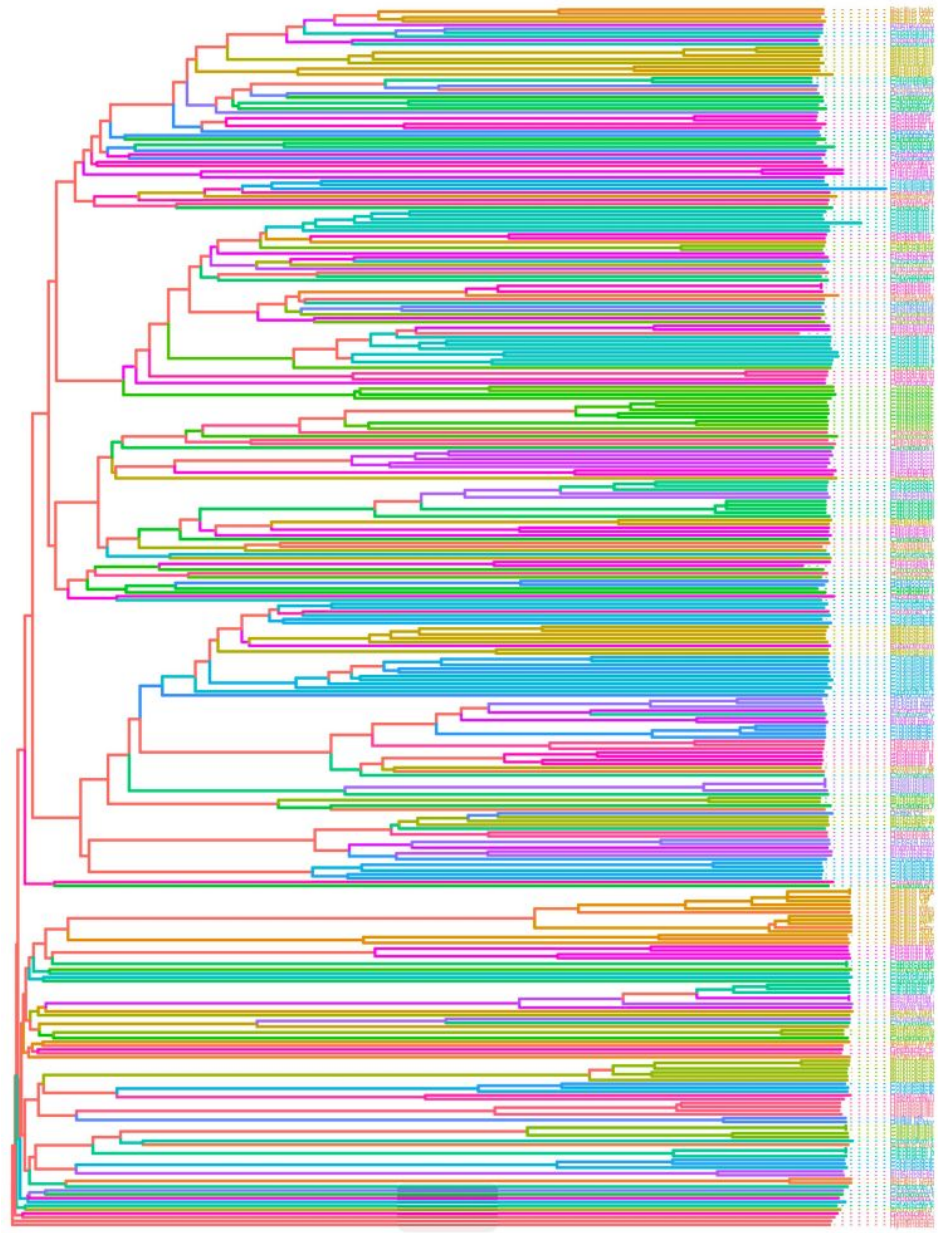
Outputs:

Error: Gblocks returned the following error:
Gblocks did not find any relevant alignment site. This usually means the alignment is not reliable. If you believe it is, you can either try your phylogeny workflow again without Gblocks or with less stringent parameters. If you believe the alignment found is bad, you can try to improve it either yourself using alignment editors or with another alignment program.

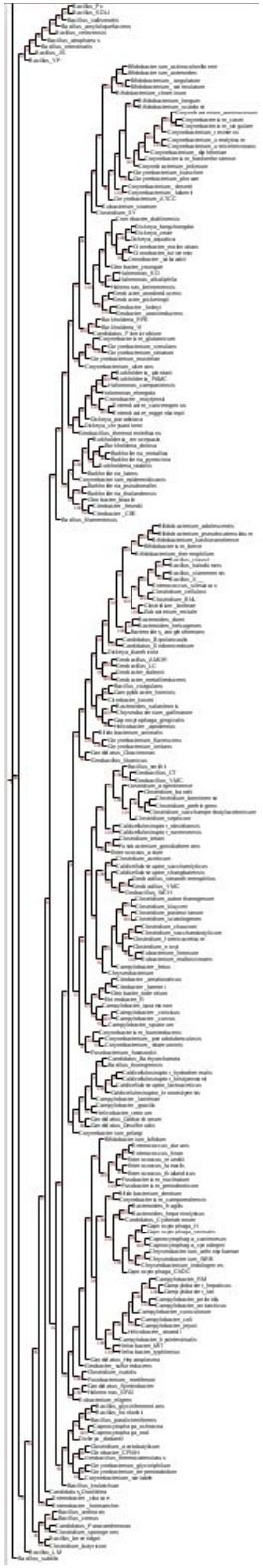
Results-NJ tree construction



Results- Rooted NJ tree



Root tree using Hymenobacter



Results-Unrooted IQtree

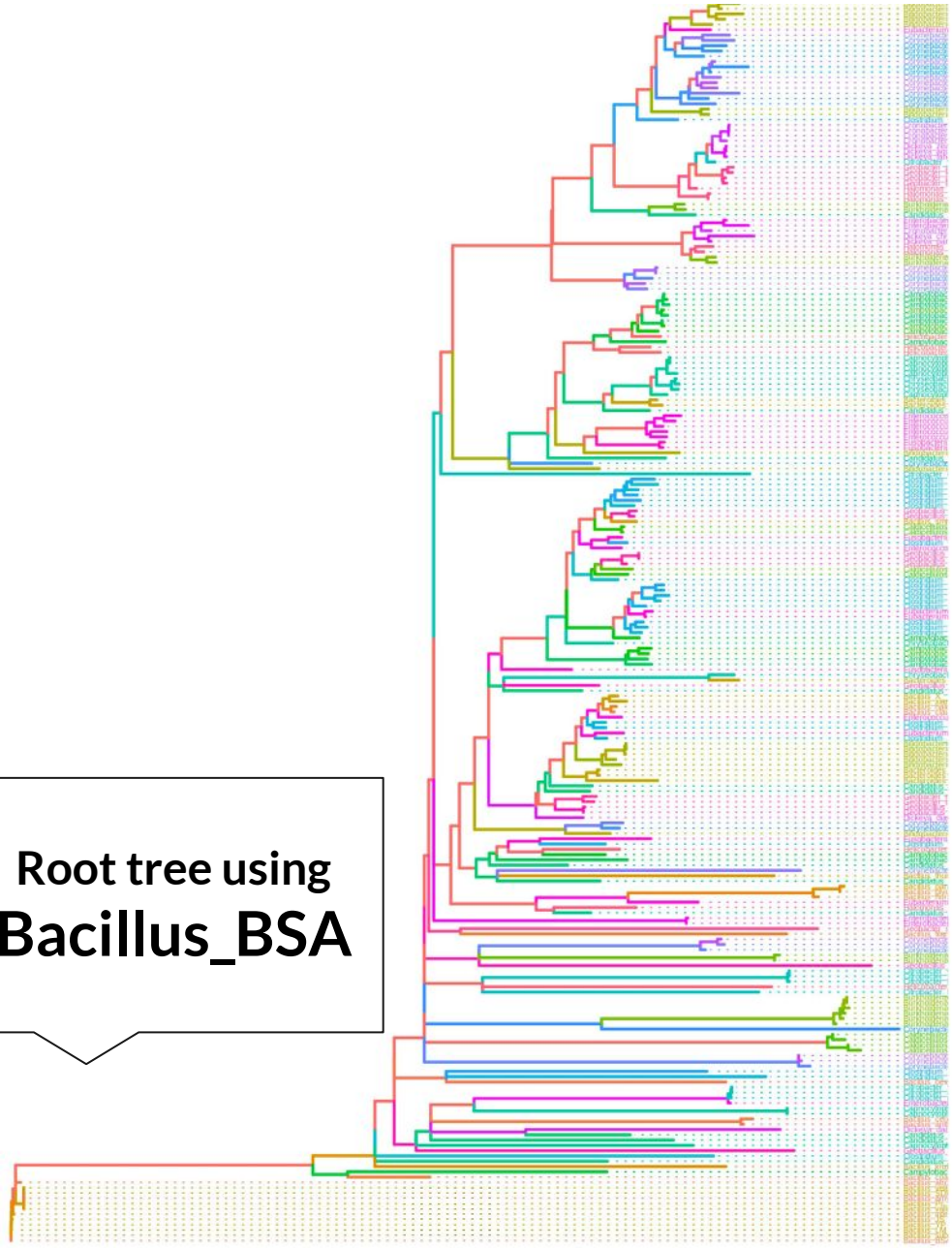
Bacillus_BSA ⟸ **Outgroup**

Best-fit model:
VT+F+G4

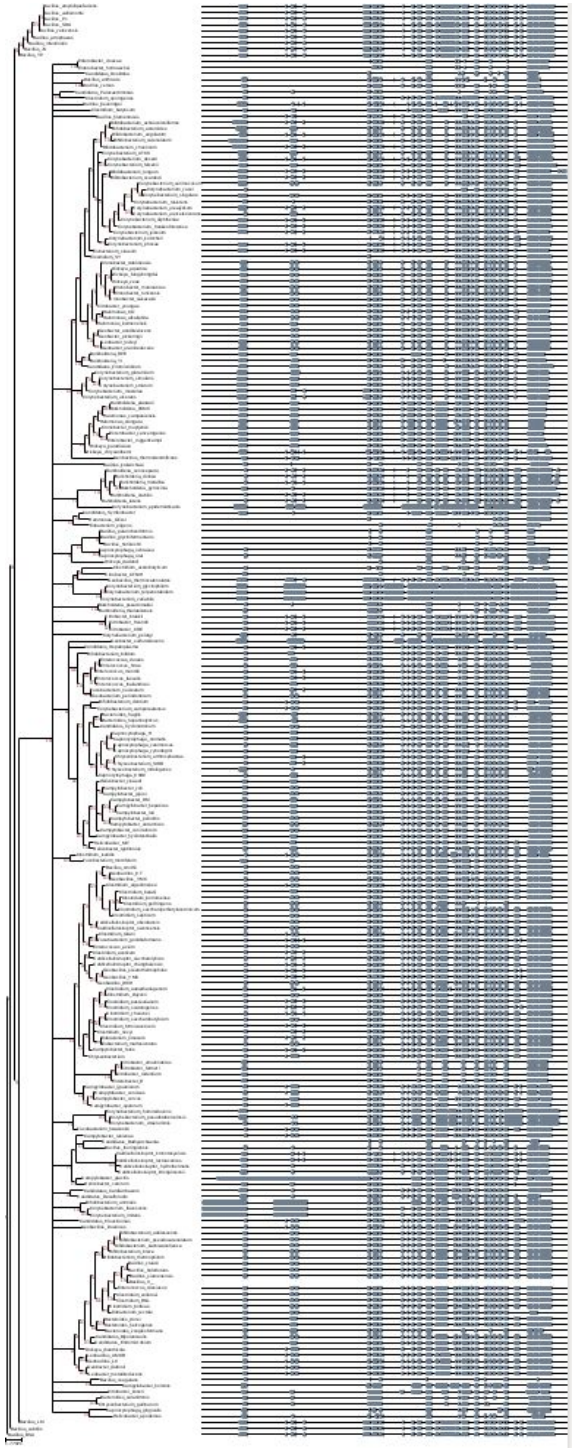
I got clear ml tree from
Phylogenetic tree (newick) viewer

http://etetoolkit.org/treeview/?treeid=ea0100092b9ba8d910322a09a4b64190&algid=f39106a4236c18b06913018514a73427

Results-Rooted IQtree

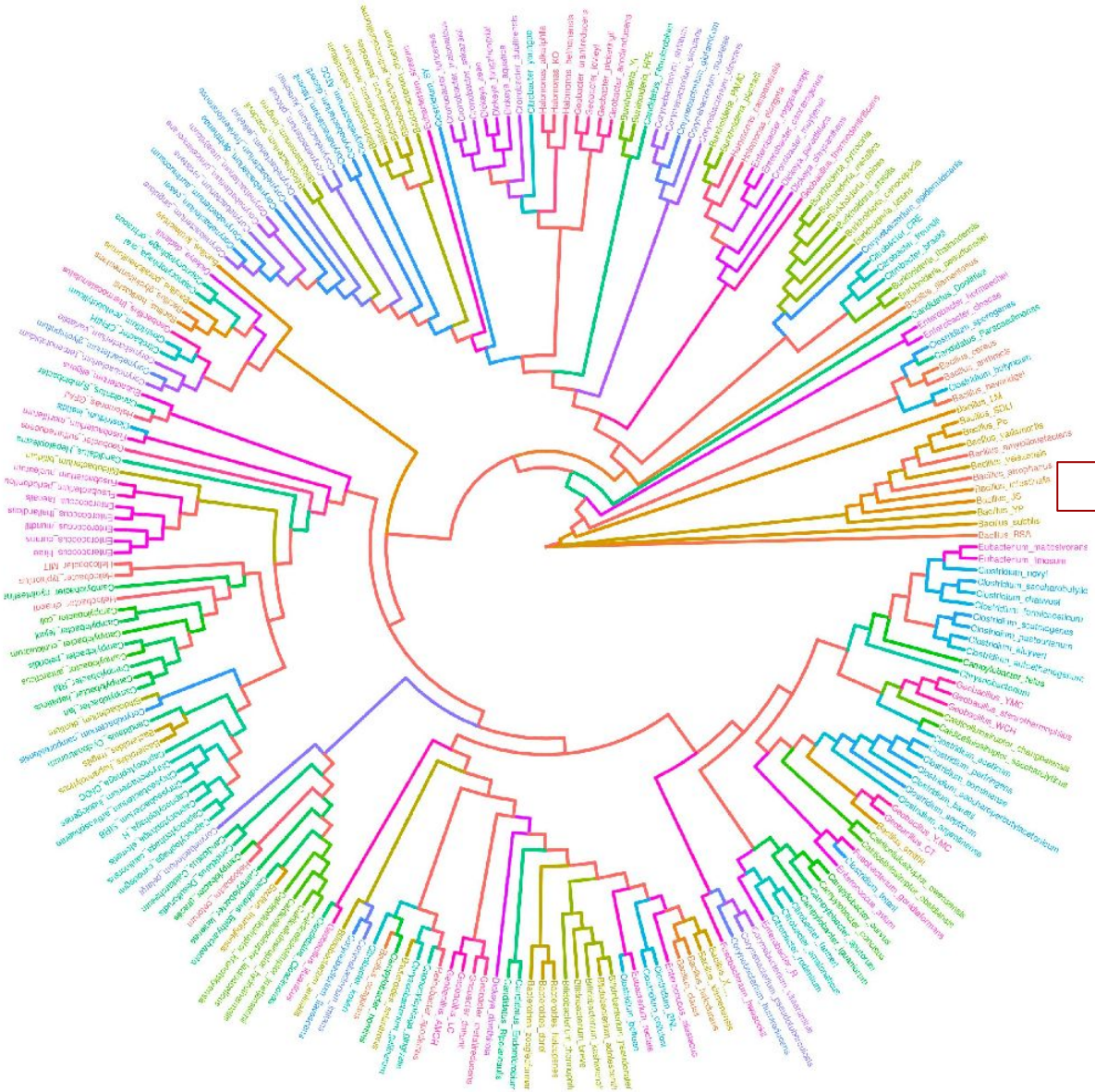


Root tree using
Bacillus_BSA



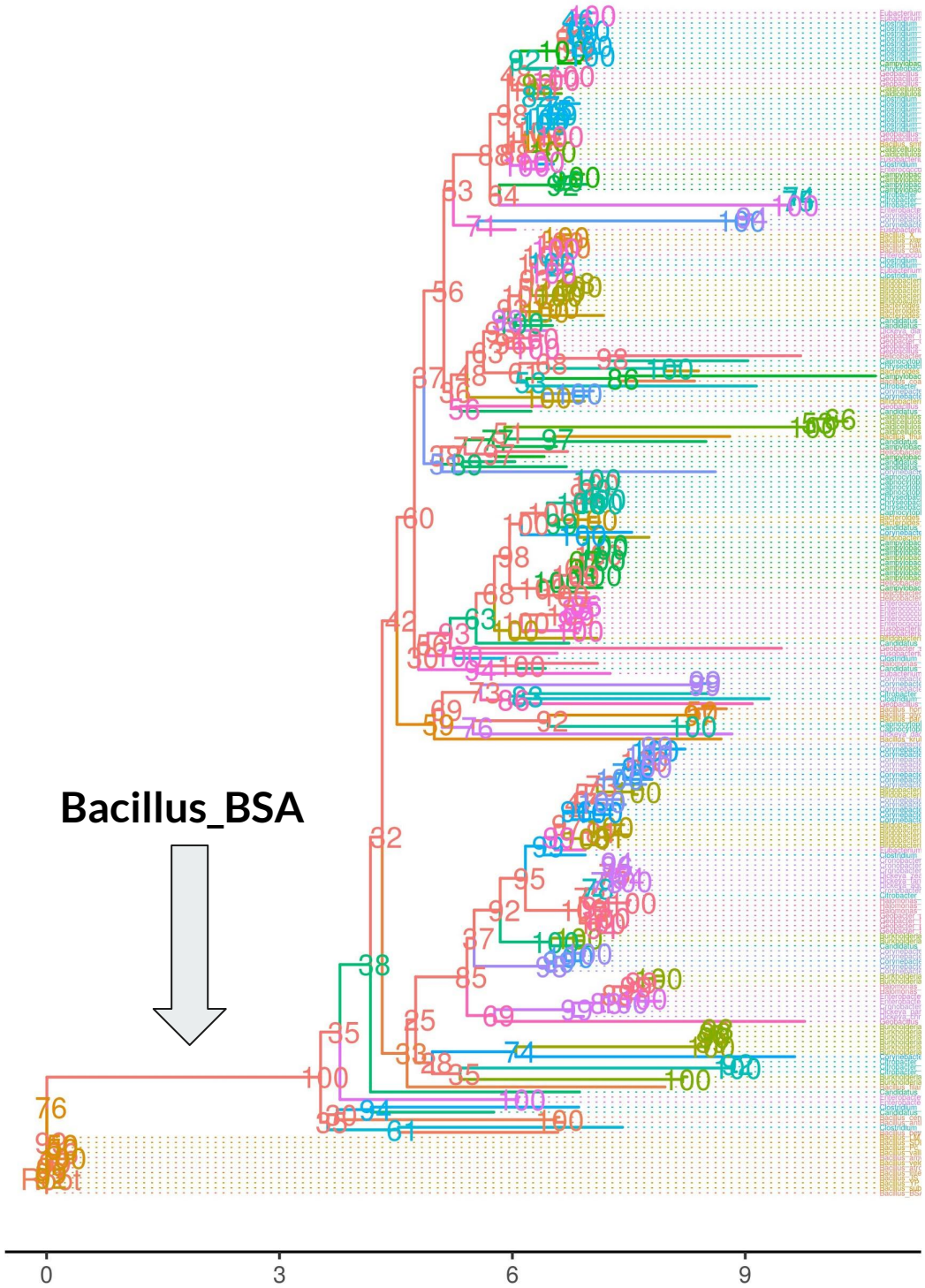
Rooted tree
with alignment

Results- IQtree with bootstrapping



Bacillus_BSA

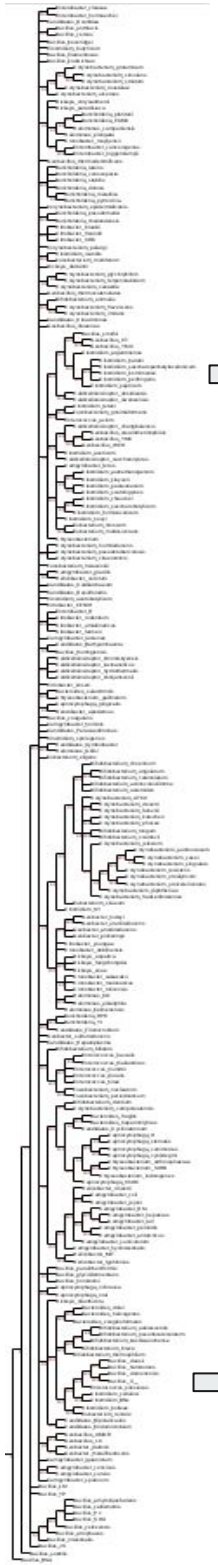
Results- IQtree with bootstrapping

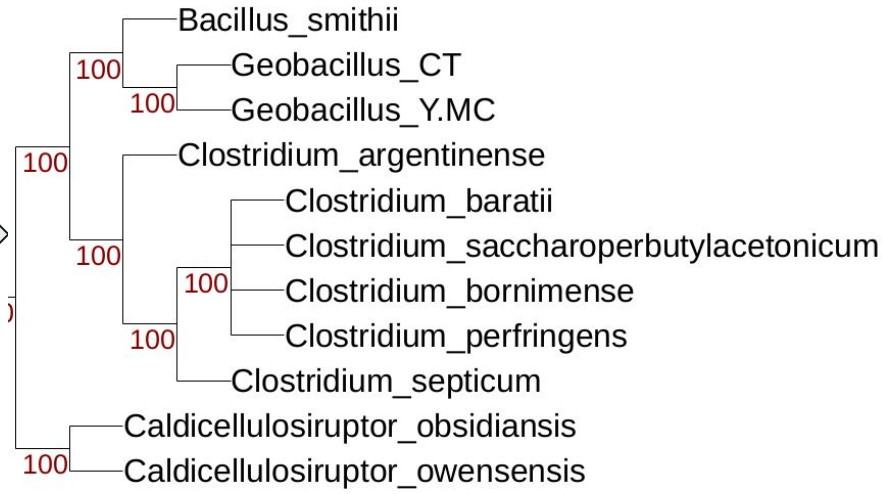


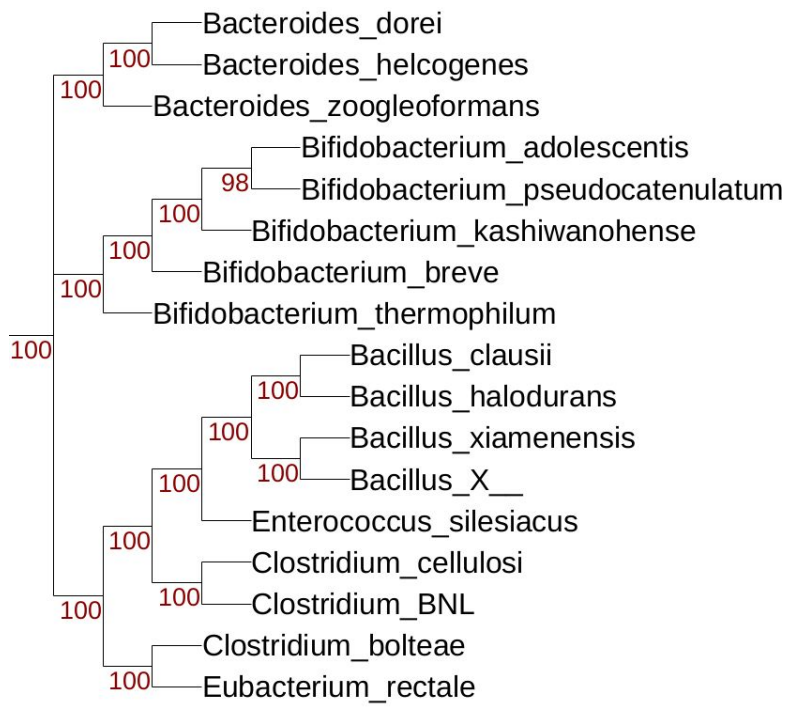
- **Tree is huge and I have to compress it as much as possible!**
- **All clear photos are available on Github**

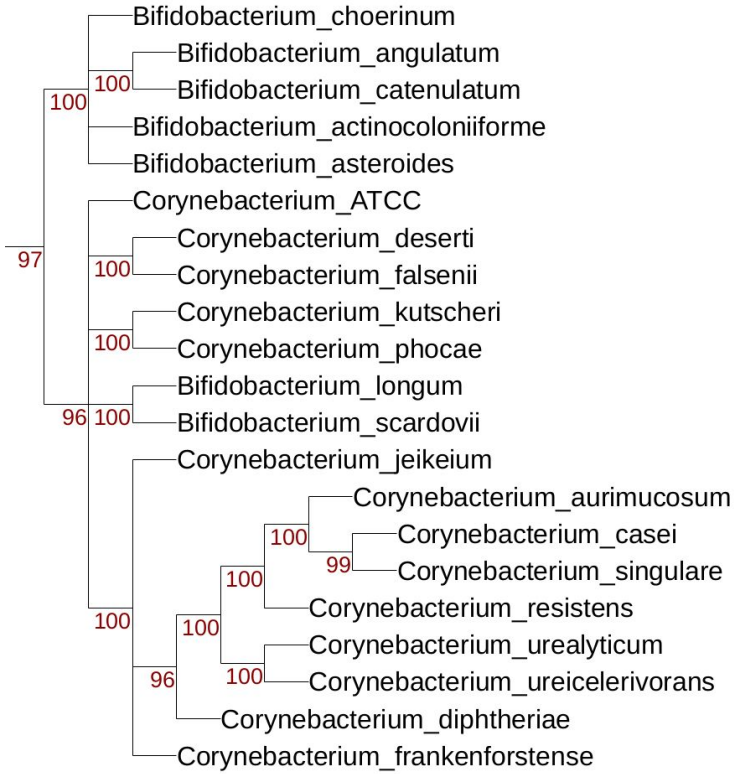
After collapsing (Final tree)

- "collapse" clades with bootstrap support < 95%









Discussion

1.) After bootstrapping “Bacillus_BSA” confirmed as an outgroup!
2.) It can be seen that cas 1 gene have been found scattered in most of the species from Bacilli (Bifidum group) and Clostridia (clostridium group) class!
    a.) both are gram positive species where bacillus are aerobic and clostridium are anaerobi
    b.) Bacillus mycoides is source organism of cas1
3.) most of the organisms evolved according to viral strains

4.) Also significant percentage of species from Actinobacteria class grouped with species from Bacilli and Clostridia class!

Source

All codes will be available on GitHub:

https://github.com/Sedreh/phylogenetics_semester3


THANK
YOU