
DTU

Phylogeny for outbreak investigation

Rolf Sommer Kaas

Overview

1. Outbreak investigations
2. Calling SNPs and Inferring Phylogenies
3. Exercise Introduction
4. Exercise
5. Additional steps and limitations

Outbreak Investigation

Udbrud af livstruende E. coli infektioner i Tyskland

De tyske myndigheder har i de seneste uger registreret et øget antal alvorlige diaré-tilfælde og tilfælde af nyresvigt

Redaktionen | 25/05/2011

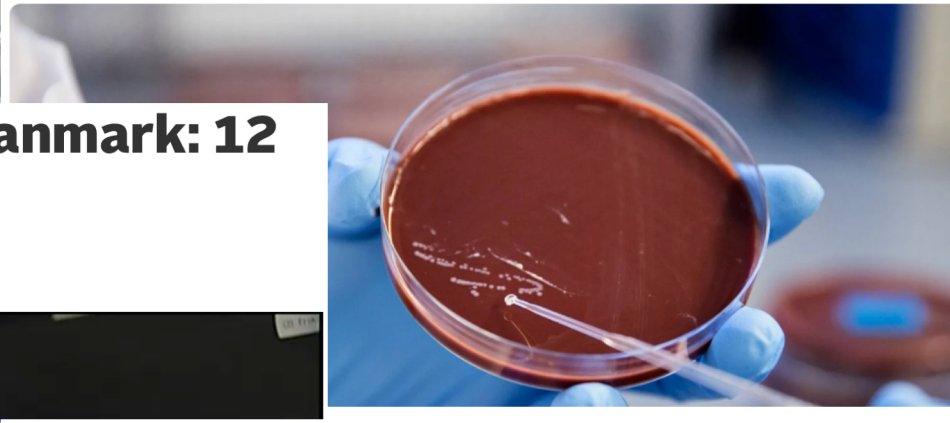
Undersøger udbrud af campylobacter

En lokal madvare mistænkes for at have indeholdt bakterie, der har gjort mange bornholmere syge

23/03/2018 KL. 15:57

Fire danskere knyttes til europæisk udbrud af listeria

Frosne majs mistænkes for at være kilden til udbruddet.



DANMARK

Tre børn har haft akut nyresvigt efter større E. coli-udbrud

Seruminstituttet efterforsker et udbrud med en E. coli-bakterie, som indtil videre omfatter 13 personer.

Listeria-udbrud i Danmark: 12 personer er døde

12. aug. 2014, 12:04
Opd. 12. aug. 2014, 12:51

Højst usædvanligt: 18 danskere smittet med salmonella fra æg

DANMARK

Kålsalater får skylden for større E. coli-udbrud i Danmark

Symptomer på E. coli-infektion er blandt andet utilpashed. Over cirka tre uger har SSI fundet 68 smittede.

DANMARK

Tre døde: Nu kalder producent alle produkter tilbage efter udbrud af salmonella

Tre har mistet livet efter at have fået salmonella af et lægemiddel fra Husk, som producenten tilbagekalder.

Efter alvorligt udbrud: Nu advarer Fødevarestyrelsen

ADVARSLER

Skrevet af Sarah Steenfeldt

50 syge af bakterier fra dansk kyllingekød

Udbrud af livstruende E. coli-infektioner i Tyskland

De tyske myndigheder har i de seneste uger registreret et øget antal alvorlige diaré-tilfælde og tilfælde af nyresvigt

Redaktionen | 25/05/2011

Sygdomstilfældene er især set i Nordtyskland og skyldes infektion med en bestemt E. coli-bakterie (VTEC, verocytotoksin-producerende E. coli), oplyser Statens Serum Institut. Man formoder, at smittekilden er en fødevare, og de tyske myndigheder arbejder intensivt på at finde denne. Det kan ikke udelukkes, at danskere er blevet syge i forbindelse med dette sygdomsudbrud. Indtil videre er der dog ikke kendskab til sygdomstilfælde i Danmark.

Fødevarestyrelsen

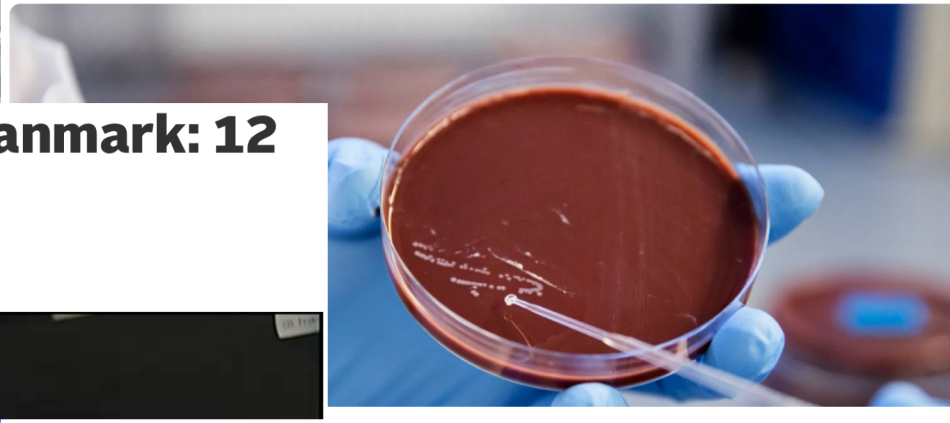
som producenten tilbagekalder.

ange bornholmere syge

23/03/2018 KL. 15:57

Fire danskere knyttes til europæisk udbrud af listeria

Frosne majs mistænkes for at være kilden til udbruddet.



brud i Danmark: 12
r døde

DANMARK

Kålsalater får skylden for større E. coli-udbrud i Danmark

Symptomer på E. coli-infektion er blandt andet utilpashed. Over cirka tre uger har SSI fundet 68 smittede.

ADVARSLER Skrevet af Sarah Steenfeldt

50 syge af bakterier fra dansk kyllingekød

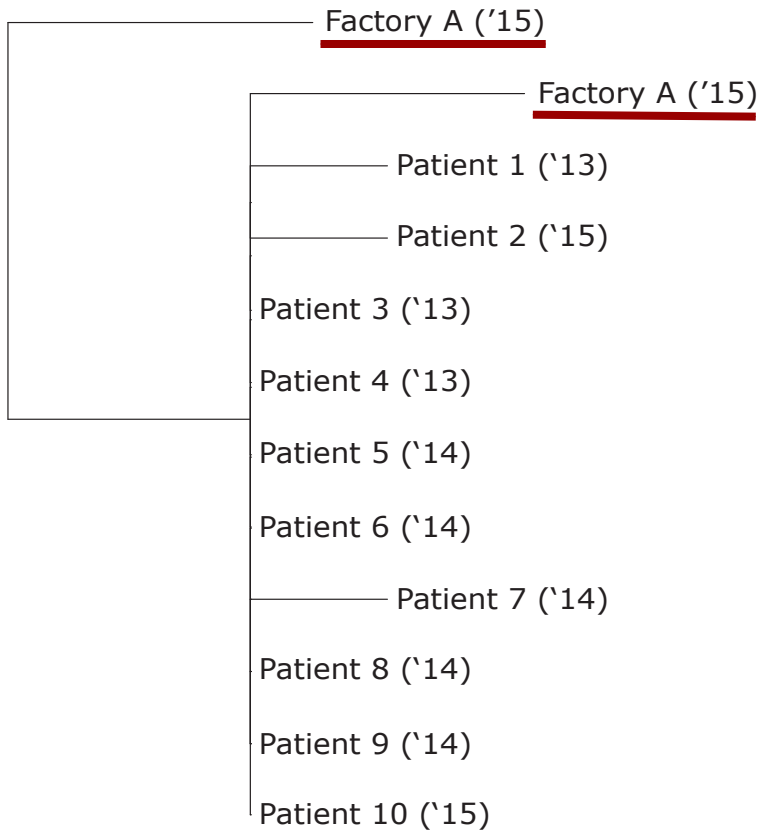
Investigation of Foodborne outbreaks in Denmark

Single Nucleotide Polymorphism Phylogeny

(Real example from Denmark)

Factory A	Factory A									
0	6	3	2	2	2	2	4	2	2	Factory A
6	0	5	4	4	4	4	6	4	4	Factory A
3	5	0	1	1	1	1	3	1	1	
2	4	1	0	0	0	0	2	0	0	
2	4	1	0	0	0	0	2	0	0	
2	4	1	0	0	0	0	2	0	0	
2	4	1	0	0	0	0	2	0	0	
4	6	3	2	2	2	2	0	2	2	
2	4	1	0	0	0	0	2	0	0	
2	4	1	0	0	0	0	2	0	0	

Listeria outbreak in Denmark



Calling SNPs and Inferring Phylogenies

Single Nucleotide Polymorphism Phylogeny

Assumption: Random + Independent

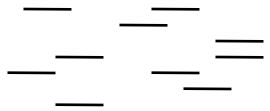
1. Find differences (SNP calling), compared to a reference sequence.
   - Close reference is better (No cross species analysis)

2. Make pseudo-alignment

3. Infer phylogeny

Alignment to Reference

Raw reads



Mappers:

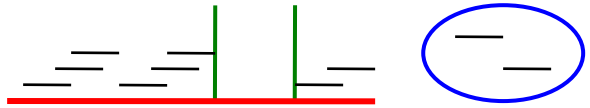
BWA

MiniMap

KMA

Do not match any reads

Do not match reference

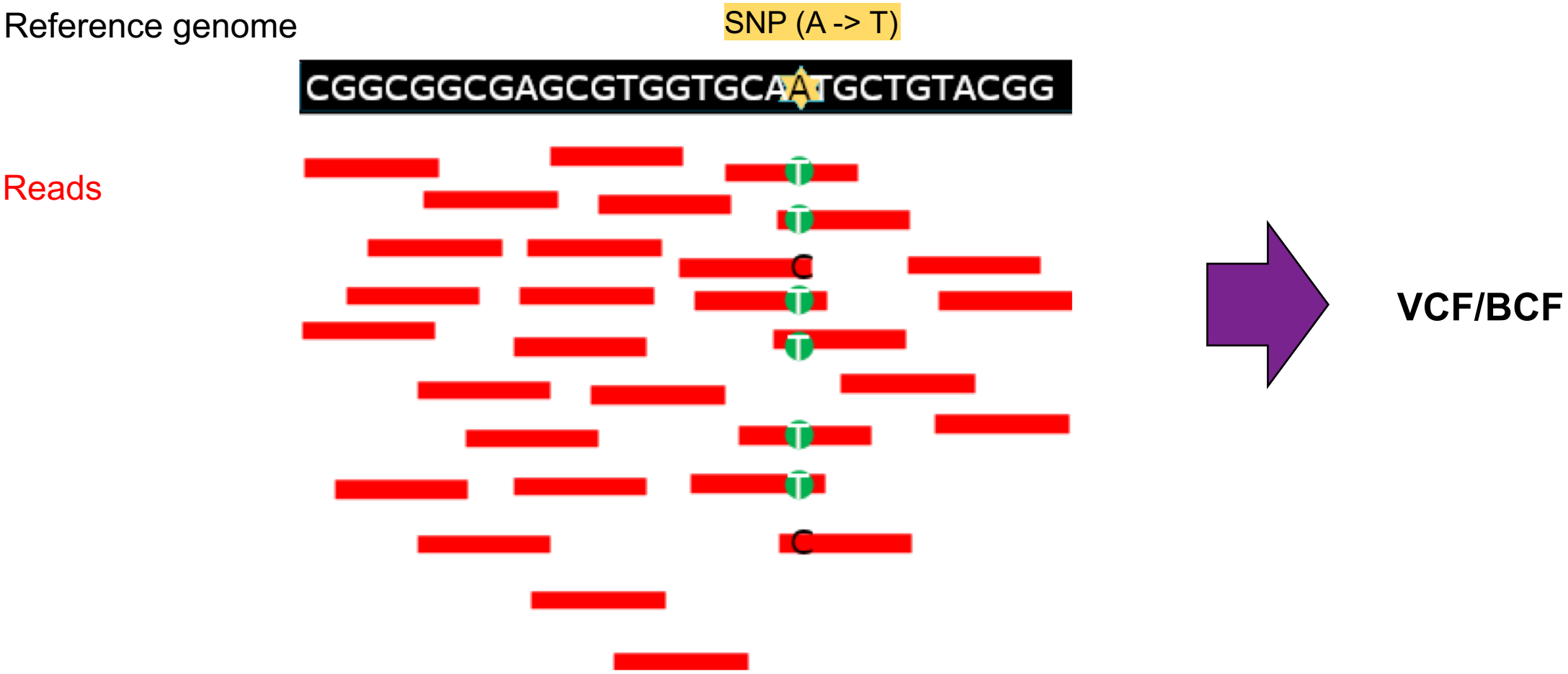


Reference sequence

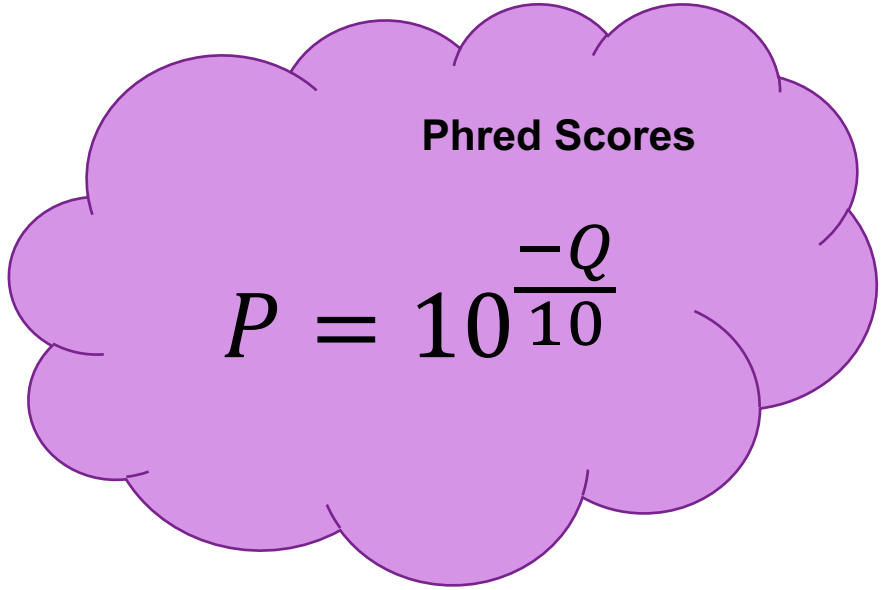
FASTQ

SAM/BAM

Call Single Nucleotide Polymorphisms (SNPs)



VCF Format




```
##fileformat=VCFv4.1
##..
#CHROM   POS      ID       REF      ALT      QUAL     FILTER
NODE_2   3253     .        C        T        73       .
NODE_23  32781    .        G        A        125      .
NODE_32  15351    .        A        C        222      .

INFO     FORMAT   0807T15624_R1.bam
DP=1644;VDB=0.0408;AF1=0.5;AC1=1;DP4=2,2,3,14;MQ=60;FQ=63.7;PV4=0.23,4.1e-09,1,0.23   GT:PL:GQ 0/1:103,0,91:94
DP=82;VDB=0.0418;AF1=0.5163;AC1=1;DP4=1,2,19,28;MQ=60;FQ=-15.1;PV4=1,1e-11,1,1        GT:PL:GQ 0/1:155,0,12:15
DP=86;VDB=0.0421;AF1=1;AC1=2;DP4=0,0,40,39;MQ=60;FQ=-265    GT:PL:GQ 1/1:255,238,0:99
```

VCF format definition: https://samtools.github.io/hts-specs/

Single Nucleotide Polymorphism Phylogeny

Assumption: Random + Independent

1. ~~Find differences (SNP calling), compared to a reference sequence.~~
   - ~~Close reference is better (No cross species analysis)~~

2. Make pseudo-alignment

3. Infer phylogeny

Pseudo-alignment of SNPs

reference nucleotides
CALLED SNPS

```
Strain 1      CTGaGATCCGAGC
Reference     tctatgctatgta
Strain 4      CTGCGATCCGAGC
Strain 6      CTGaGATCCGAGa
```

Single Nucleotide Polymorphism Phylogeny

Assumption: Random + Independent

1. ~~Find differences (SNP calling), compared to a reference sequence.~~
   - ~~Close reference is better (No cross species analysis)~~

2. ~~Make pseudo-alignment~~

3. Infer phylogeny

Infer Phylogeny

Popular algorithms

- Maximum Likelihood (Maximizing Likelihood functions)
  - Most accurate
  - Can’t provide SNP distances

- Maximum Parsimony (Minimum evolution)
  - Provides SNP distances
  - Intuitive method
  - Underestimates actual evolutionary change (LBA)

Infer Phylogeny

Markov models

Used to calculate likelihood of phylogenetic trees

Popular Substitution models

Jukes-Cantor, 69 (JC69)

$$\begin{pmatrix} & A & C & G & T \\ A & * & 0.25 & 0.25 & 0.25 \\ C & 0.25 & * & 0.25 & 0.25 \\ G & 0.25 & 0.25 & * & 0.25 \\ T & 0.25 & 0.25 & 0.25 & * \end{pmatrix}$$

Kimura, 80 (K80)

$$\begin{pmatrix} & A & C & G & T \\ A & * & 0.25\alpha_1 & 0.25 & 0.25\alpha_1 \\ C & 0.25\alpha_1 & * & 0.25\alpha_1 & 0.25 \\ G & 0.25 & 0.25\alpha_1 & * & 0.25\alpha_1 \\ T & 0.25\alpha_1 & 0.25 & 0.25\alpha_1 & * \end{pmatrix}$$

Hasegawa-Kishino-Yano (HKY)

$$\begin{pmatrix} & A & C & G & T \\ A & * & \pi_C\alpha_1 & \pi_G & \pi_T\alpha_1 \\ C & \pi_A\alpha_1 & * & \pi_G\alpha_1 & \pi_T \\ G & \pi_A & \pi_C\alpha_1 & * & \pi_T\alpha_1 \\ T & \pi_A\alpha_1 & \pi_C & \pi_G\alpha_1 & * \end{pmatrix}$$

Generalised Time Reversible (GTR/REV)

$$\begin{pmatrix} & A & C & G & T \\ A & * & \pi_C\alpha_1 & \pi_G & \pi_T\alpha_2 \\ C & \pi_A\alpha_1 & * & \pi_G\alpha_3 & \pi_T\alpha_4 \\ G & \pi_A & \pi_C\alpha_3 & * & \pi_T\alpha_5 \\ T & \pi_A\alpha_2 & \pi_C\alpha_4 & \pi_G\alpha_5 & * \end{pmatrix}$$

Single Nucleotide Polymorphism Phylogeny

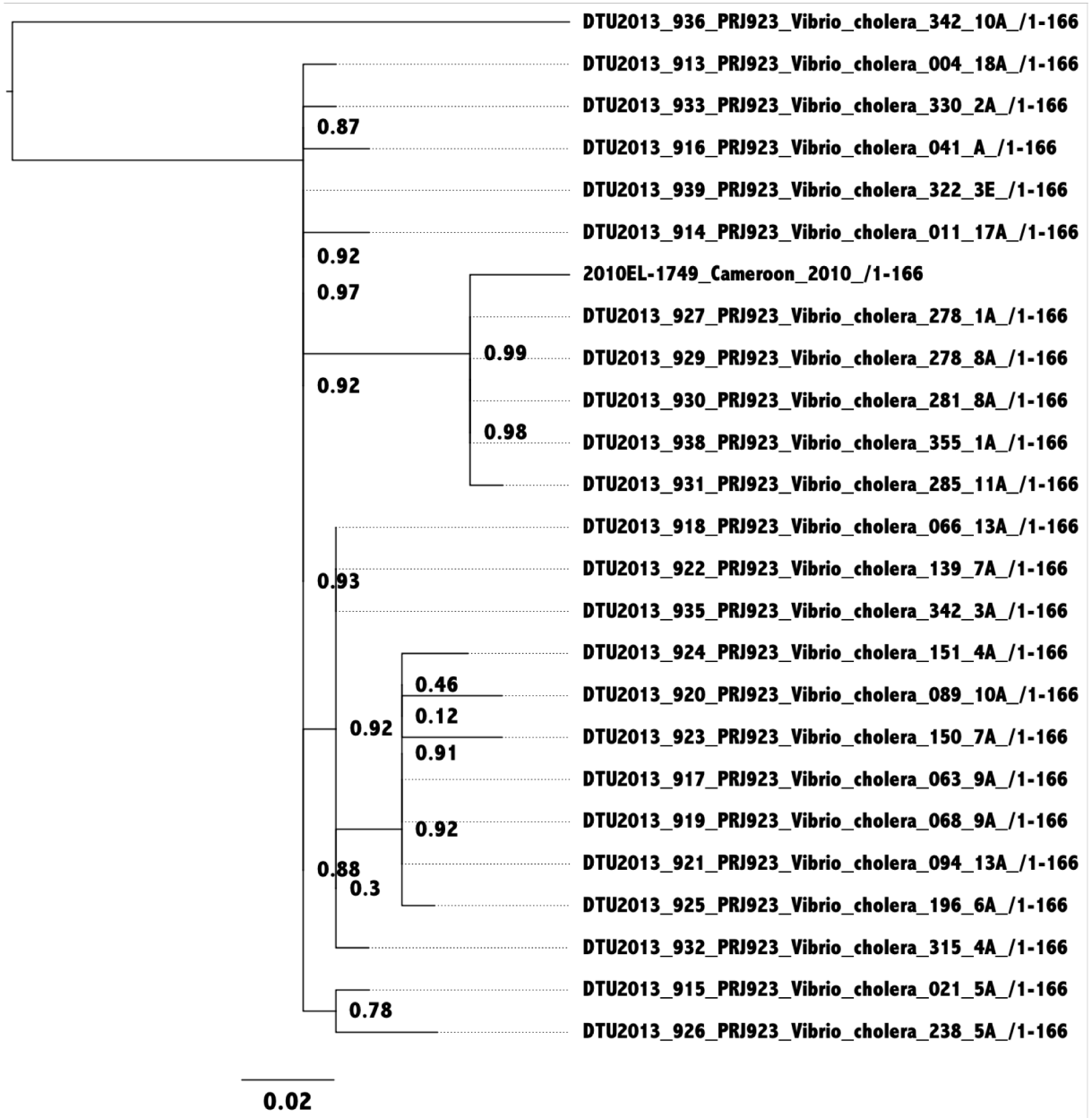
Assumption: Random + Independent

1. ~~Find differences (SNP calling), compared to a reference sequence.~~
   - ~~Close reference is better (No cross species analysis)~~

2. ~~Make pseudo-alignment~~

3. ~~Infer phylogeny~~

Single Nucleotide Polymorphism Phylogeny



Exercise

Additional Steps & Limitations

Additional Steps & Limitations

Additional steps

- Pruning – Remove SNPs caused by mobile elements
- Remove/Handle non-core SNPs
- Further filtering

Limitations

- SNPs are not random & independent
- Diversity => Less confidence