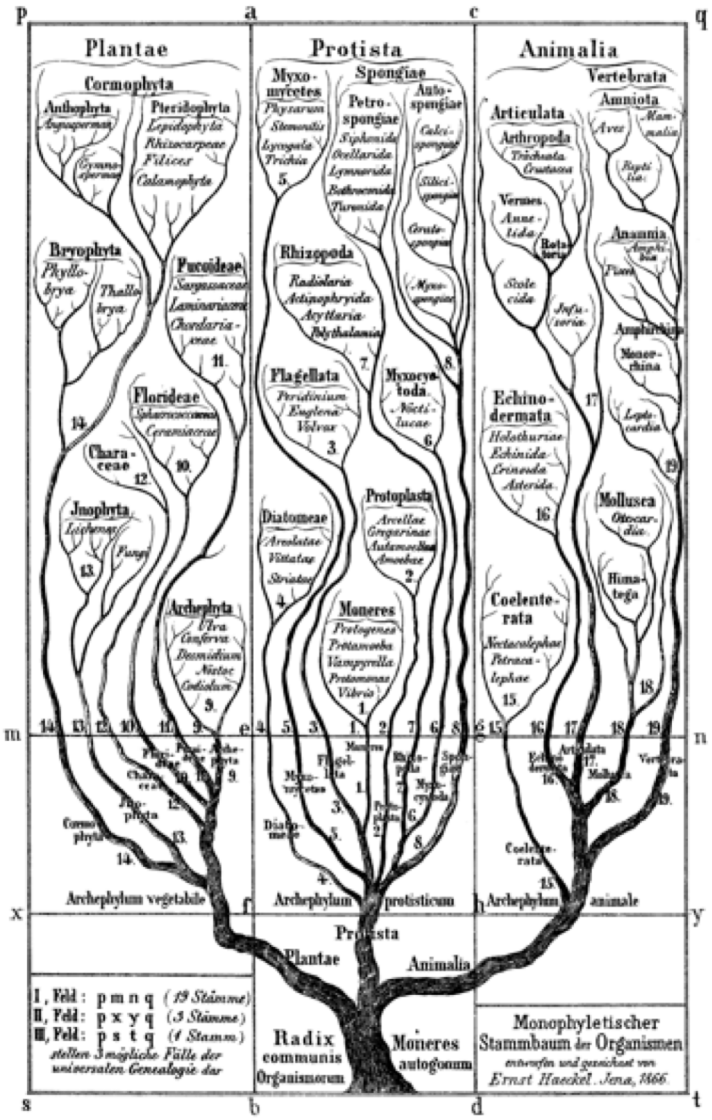
Whole Genome based Phylogeny

Pimlapas Leekitcharoenphon (Shinny)
pile@food.dtu.dk
Researcher
DTU Food
27-05-2022

Overview

- What is Phylogeny
- SNP methods
  - CSI Phylogeny
- Gene by gene approach
  - cgMLST
- Phylogeny with space and time
- How to visualize phylogenetic tree



What is phylogeny?

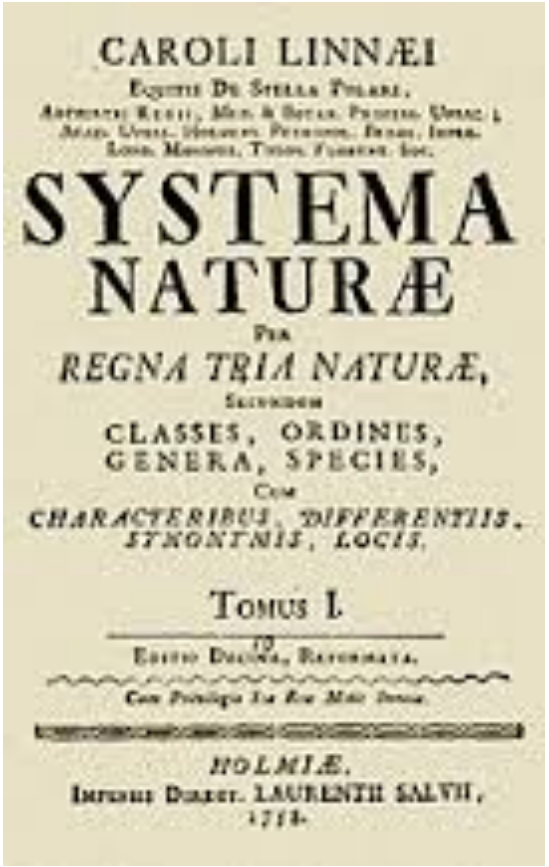
- Early phylogeny
  - Classification
  - Based on phenotypes
- Current phylogeny
  - Based on genotypes
  - DNA mutations as basis for evolution

Classification

Carl Linnaeus 1707-1778

Hierarchical system

- Kingdom
- Phylum
- Class
- Order
- Family
- Genus
- Species



CAROLI LINNÆI

SYSTEMA NATURÆ

PER

REGNA TRIA NATURÆ,

CLASSES, ORDINES, GENERA, SPECIES,

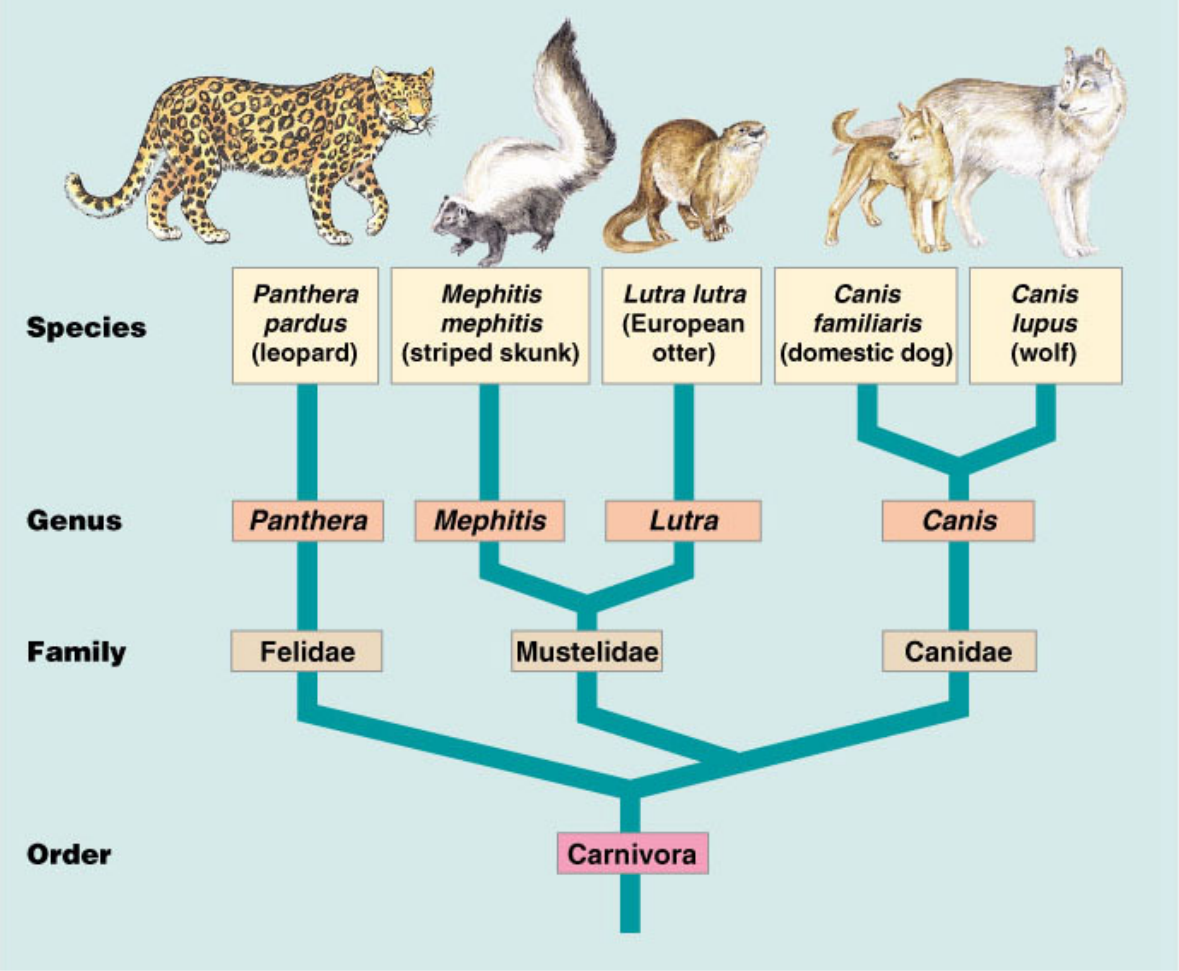
CUM

CHARACTERIBUS, DIFFERENTIIS, SYNONYMIS, LOCIS.

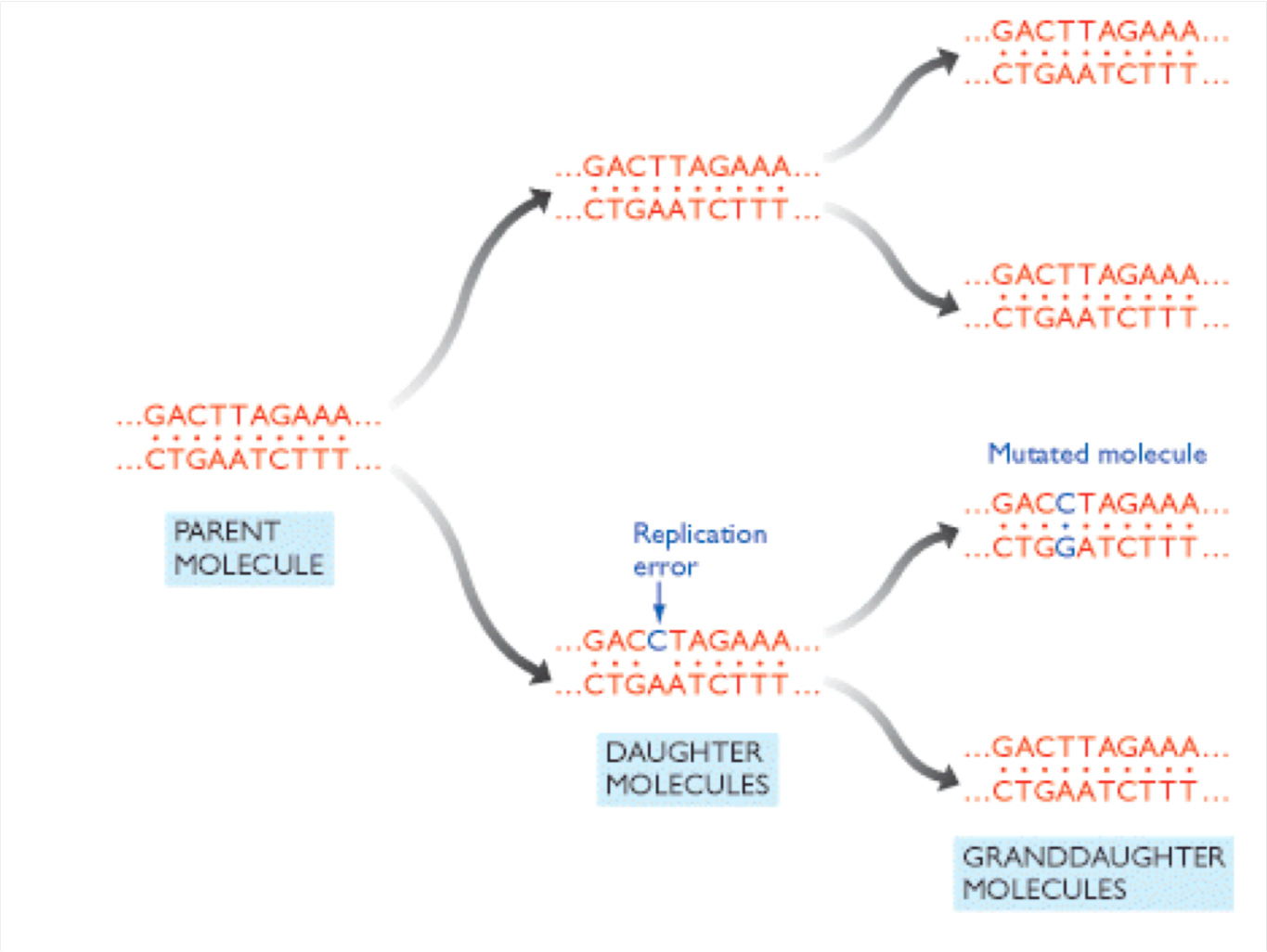
TOMUS I.

HOLMIÆ.

Classification depicted as a tree



DNA mutations as basis for evolution

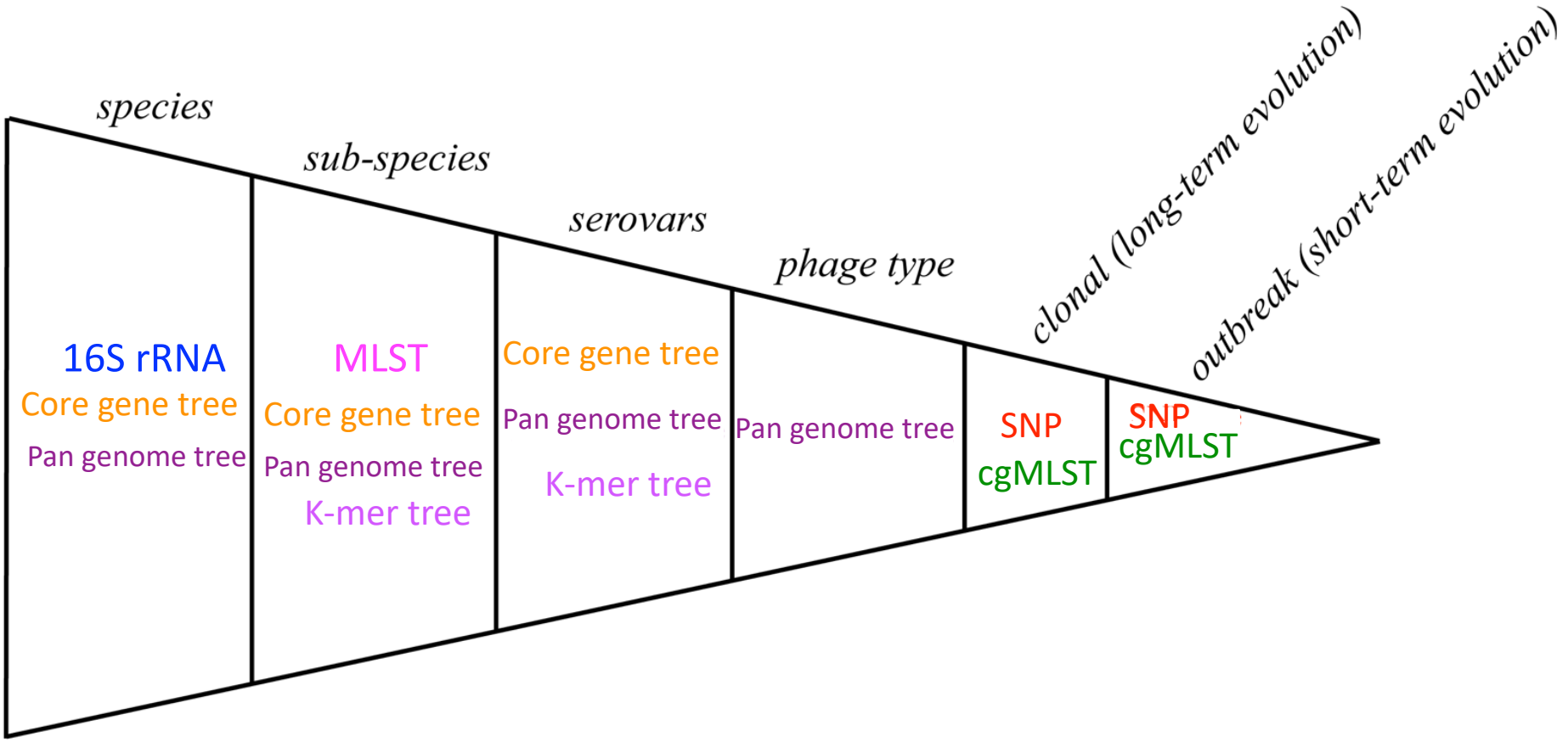


What are phylogenetic trees

- Phylogenetic trees are a visual representation of the genetic relationship between species
- Think of them as family trees
- Phylogeny can also be represented by distance matrices

What are phylogenetic trees

- Trees were traditionally made using aligned sequences of single genes or proteins
- Whole genome data can be used to create trees based on
  - SNP calling
  - K-mer overlap
  - Alignment of genes and genomes


species
sub-species
serovars
phage type
clonal (long-term evolution)
outbreak (short-term evolution)
16S rRNA
Core gene tree
Pan genome tree
MLST
Core gene tree
Pan genome tree
K-mer tree
Core gene tree
Pan genome tree
K-mer tree
Pan genome tree
SNP
cgMLST
SNP
cgMLST

Whole genome based phylogeny

- Single nucleotide polymorphism (SNPs) approach
  - Require reference genome
- Gene by gene approach
  - cgMLST, wgMLST
  - No reference genome required
  - Required species specific cgMLST scheme

What is a SNP

- A Single Nucleotide Polymorphism (SNP) is a DNA sequence variation occurring commonly* within a population (e.g. 1%) in which a Single Nucleotide — A, T, C or G — in the genome (or other shared sequence) differs between members of a biological species or paired chromosomes.

How does it work

Strain A	AT**T**CA**G**T**A**G**T**
Strain B	AT**G**CA**G**T**T**G**A**
Strain C	AT**G**CA**A**T**T**G**T**
Strain D	AT**C**CA**T**T**A**G**C**

Construct distance matrix

Strain A ATTCAGTAGT

Strain B ATGCAGTTGA

Strain C ATGCAATTGT

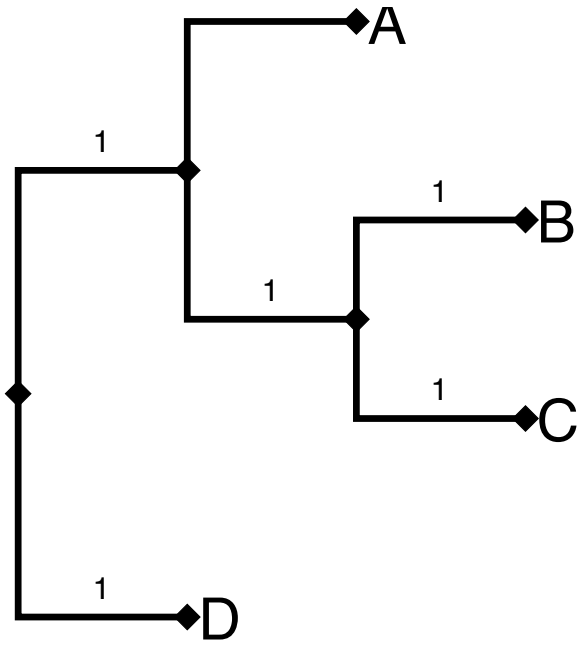
Strain D ATCCATTAGC

	A	B	C	D
A	0	3	3	3
B	3	0	2	4
C	3	2	0	4
D	3	4	4	0

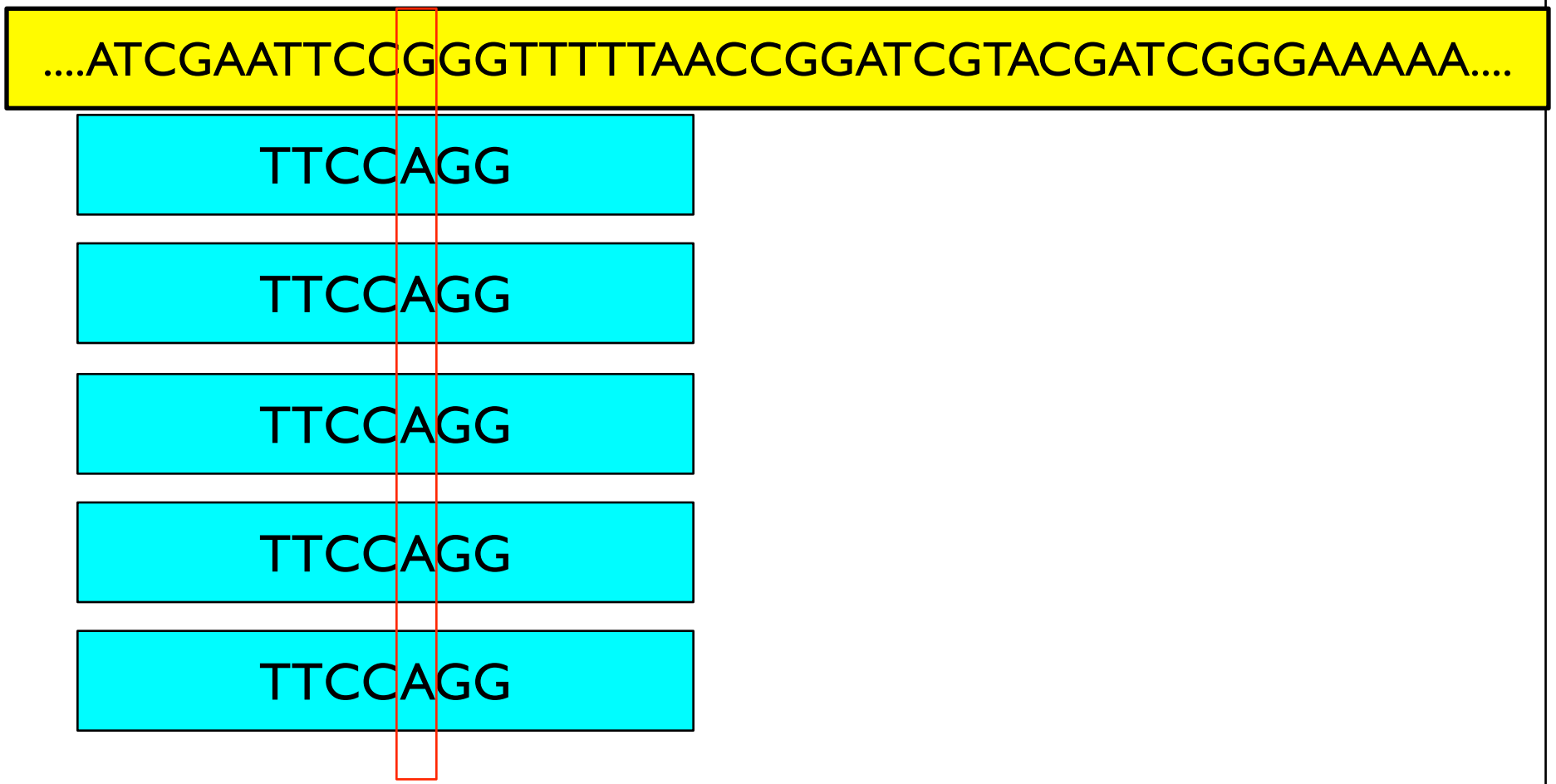
Make Tree

Strain A	AT**T**CA**G**T**A**G**T**
Strain B	AT**G**CA**G**T**T**G**A**
Strain C	AT**G**CA**A**T**T**G**T**
Strain D	AT**C**CA**T**T**A**G**C**

	A	B	C	D
A	0	3	3	3
B	3	0	2	4
C	3	2	0	4
D	3	4	4	0



SNPs identification



SNP identification (SAMtool)

https://cge.food.dtu.dk/services/CSIPhylogeny/

Center for Genomic Epidemiology

Username
Password
New Reset Login

Home | Services | Instructions | Output | Article abstract

CSI Phylogeny 1.1 (Call SNPs & Infer Phylogeny)

CSI Phylogeny calls SNPs, filters the SNPs, does site validation and infers a phylogeny based on the concatenated alignment of the high quality* SNPs.

Note: The old version of this service is still available at: CSI Phylogeny 1.0a. But it is now deprecated and no longer suported.

Service updated (14:30 10-Mar-2016 GMT+1). Service was down for several days due to errors in the queing system. The downtime was exploited to implement a new queing method for this service. It has been tested and should work but please don't hesitate to write Scientific support if your jobs are failing. The update does not affect output results, only where the pipeline is executed on the CGE server.

Input data

Upload reference genome (fasta format)
Note: Reference genome must not be compressed.

Choose File no file selected
☐ Include reference in final phylogeny.

Select min. depth at SNP positions
10x

Select min. relative depth at SNP positions
10 %

Select minimum distance between SNPs (prune)
10 bp

Select min. SNP quality
30

Select min. read mapping quality
25

Select min. Z-score
1.96

CSI Phylogeny

https://cge.food.dtu.dk/services/CSIPhylogeny/

- Strict sorting of SNPs
  - Depth
  - Relative depth
  - Distance between SNPs
  - SNP quality
  - Read mapping quality

Rolf S. Kaas, Pimlapas Leekitcharoenphon, Frank M. Aarestrup, Ole Lund. Solving the Problem of Comparing Whole Bacterial Genomes across Different Sequencing Platforms. PLoS ONE 2014; 9(8): e104984.

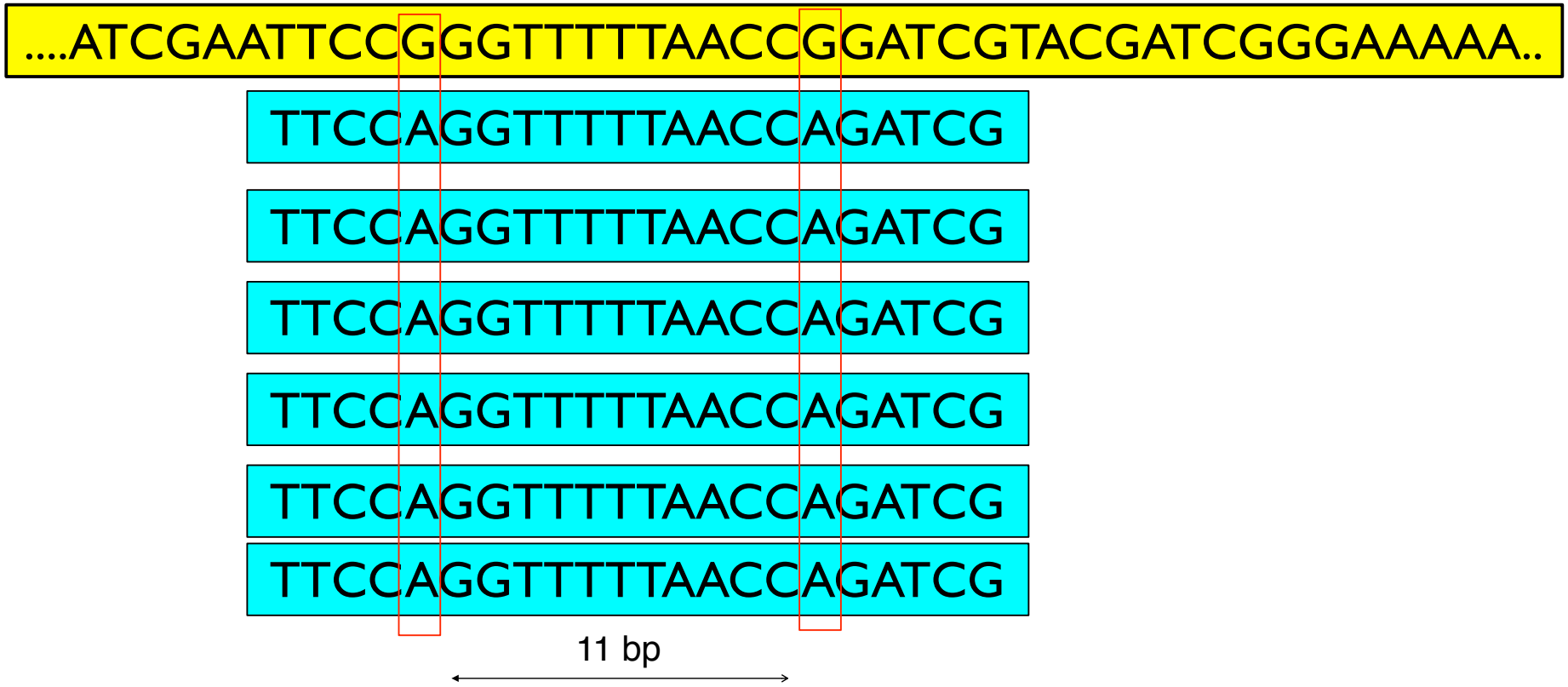
CSI Phylogeny

- Requires all SNPs to be significant
  - Z-score higher than 1.96 for all SNPs

$$Z = \frac{X - Y}{\sqrt{X+Y}}$$

- X is the number of reads, with the most common nucleotide at that position, and Y the number of reads with any other nucleotide.

Pruning


....ATCGAATTCCGGGTTTTTAACCGGATCGTACGATCGGGAAAAA..
TTCCAGGTTTTTAACCAGATCG
TTCCAGGTTTTTAACCAGATCG
TTCCAGGTTTTTAACCAGATCG
TTCCAGGTTTTTAACCAGATCG
TTCCAGGTTTTTAACCAGATCG
TTCCAGGTTTTTAACCAGATCG
11 bp

Variant calling format (VCF)

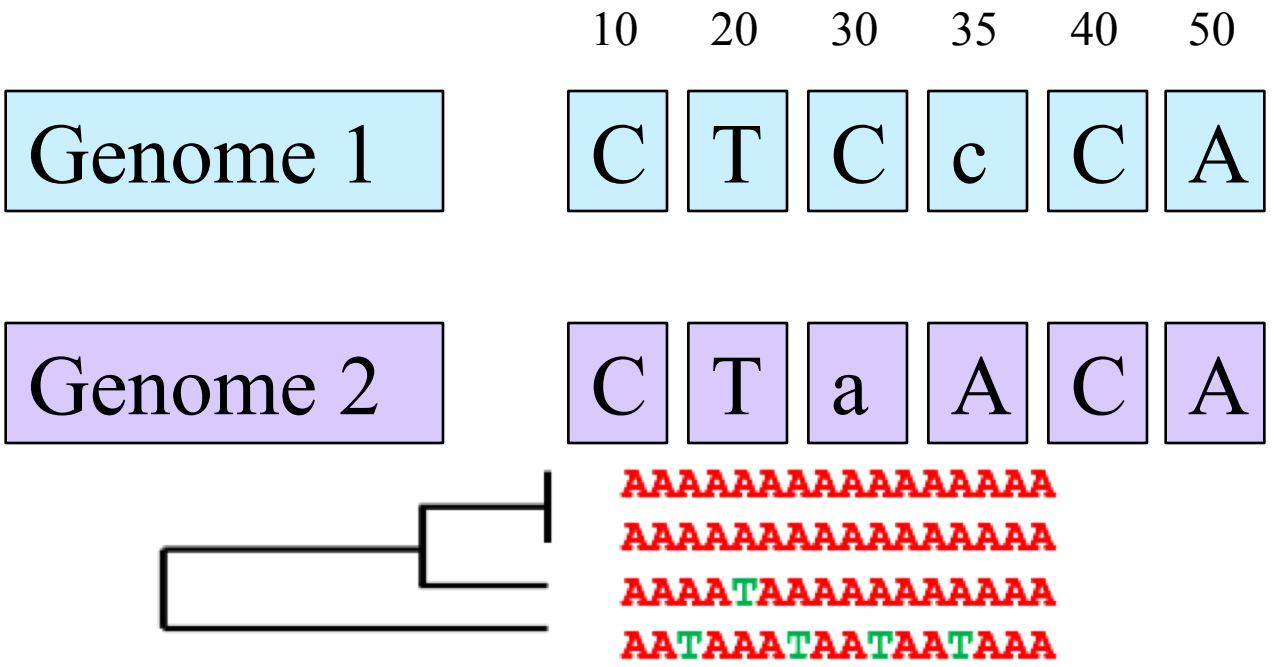
Genome 1	position	ref	change
Ref_genome	10	T	C
Ref_genome	20	C	T
Ref_genome	30	A	C
Ref_genome	40	A	C
Ref_genome	50	G	A

Genome 2	position	ref	change
Ref_genome	10	T	C
Ref_genome	20	C	T
Ref_genome	35	C	A
Ref_genome	40	A	C
Ref_genome	50	G	A

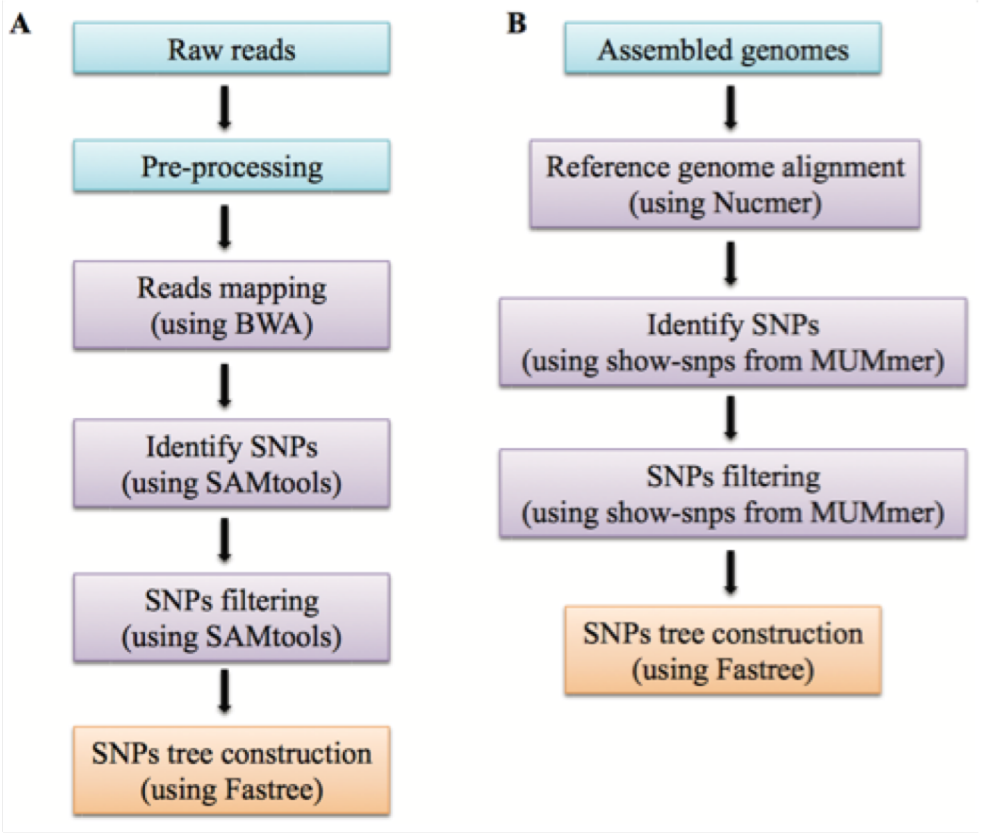
Variant calling format (VCF)

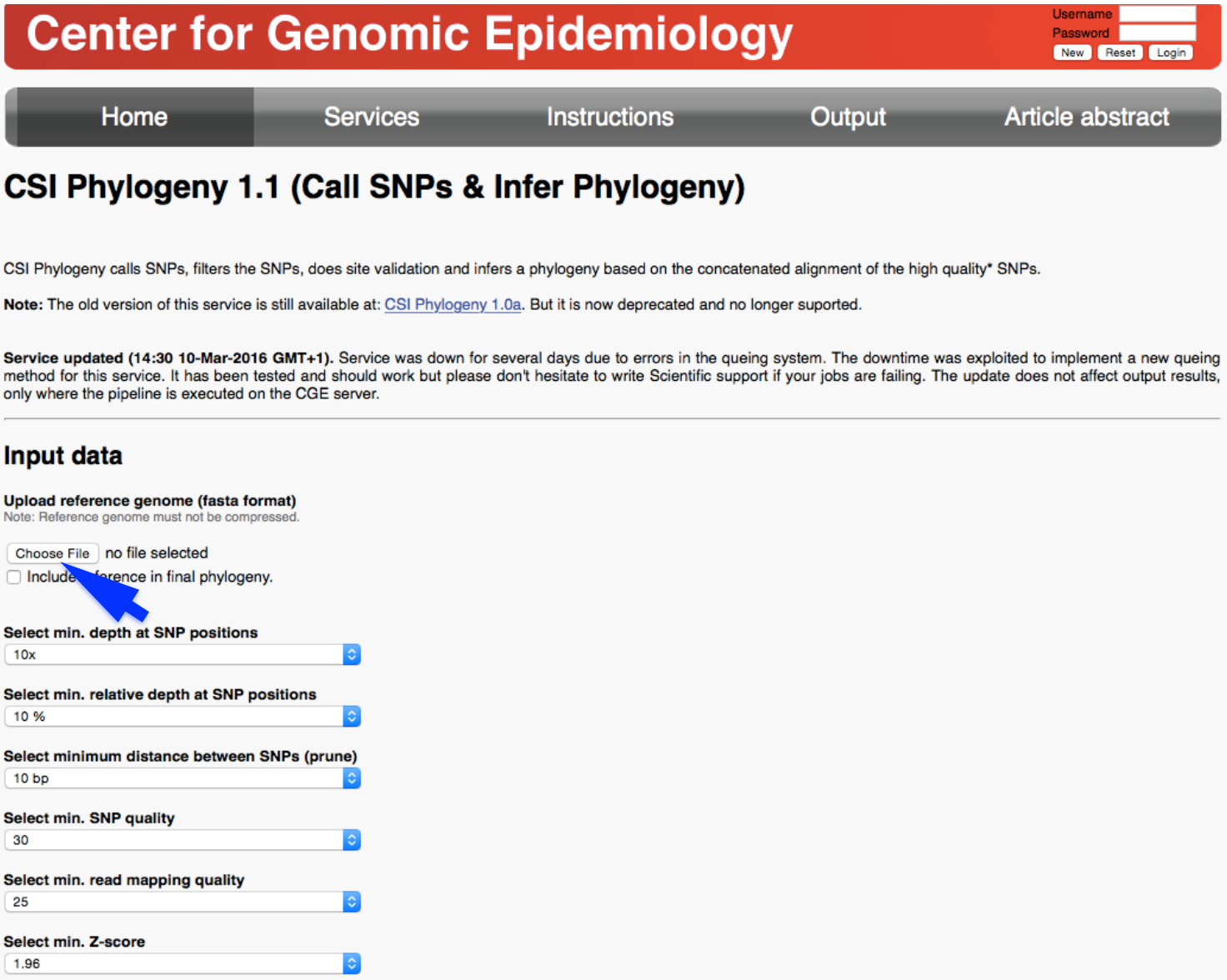
Genome 1	position	ref	change
Ref_genome	10	T	C
Ref_genome	20	C	T
Ref_genome	30	A	C
Ref_genome	40	A	C
Ref_genome	50	G	A

Genome 2	position	ref	change
Ref_genome	10	T	C
Ref_genome	20	C	T
Ref_genome	35	C	A
Ref_genome	40	A	C
Ref_genome	50	G	A



SNP tree flow





Center for Genomic Epidemiology

Username
Password
New Reset Login

Home | Services | Instructions | Output | Article abstract

CSI Phylogeny 1.1 (Call SNPs & Infer Phylogeny)

CSI Phylogeny calls SNPs, filters the SNPs, does site validation and infers a phylogeny based on the concatenated alignment of the high quality* SNPs.

Note: The old version of this service is still available at: CSI Phylogeny 1.0a. But it is now deprecated and no longer suported.

Service updated (14:30 10-Mar-2016 GMT+1). Service was down for several days due to errors in the queing system. The downtime was exploited to implement a new queing method for this service. It has been tested and should work but please don't hesitate to write Scientific support if your jobs are failing. The update does not affect output results, only where the pipeline is executed on the CGE server.

Input data

Upload reference genome (fasta format)
Note: Reference genome must not be compressed.

Choose File no file selected
Include reference in final phylogeny.

Select min. depth at SNP positions
10x

Select min. relative depth at SNP positions
10 %

Select minimum distance between SNPs (prune)
10 bp

Select min. SNP quality
30

Select min. read mapping quality
25

Select min. Z-score
1.96

Use altered FastTree (more accurate)
Note: Read more here

Upload read files and/or assembled genomes (fasta or fastq format)
Note: Read files must be compressed with gzip (compressed files often ends with .gz).
If you get an "Access forbidden. Error 403": Make sure the start of the web adress is https and not just http. Fix it by clicking here.

Isolate File

Name	Size	Progress	Status

Upload Remove

***High quality SNPs**
A high quality SNP are defined as a SNP that obeys the following rules:

Confidentiality:
The sequences are kept confidential and will be deleted after 48 hours.

Use altered FastTree (more accurate)
Note: Read more here

Upload read files and/or assembled genomes (fasta or fastq format)
Note: Read files must be compressed with gzip (compressed files often ends with .gz).
If you get an "Access forbidden. Error 403": Make sure the start of the web adress is https and not just http. Fix it by clicking here.

Isolate File

Name	Size	Progress	Status
Salmonella-spp-02-03-002.fna	4.80 MB		
Salmonella-spp-02-03-008.fna	4.81 MB		
Salmonella-spp-05-102.fna	4.81 MB		
Salmonella-spp-07-022.fna	4.80 MB		

Upload Remove

***High quality SNPs**
A high quality SNP are defined as a SNP that obeys the following rules:

Confidentiality:
The sequences are kept confidential and will be deleted after 48 hours.

CSI Phylogeny

Output

Tree build by FastTree algorithm, in Newick format

- Branch lengths is substitutions per site **at** the variable sites

Matrix of SNP pair counts in text (.txt) format

- Diagonal SNP matrix

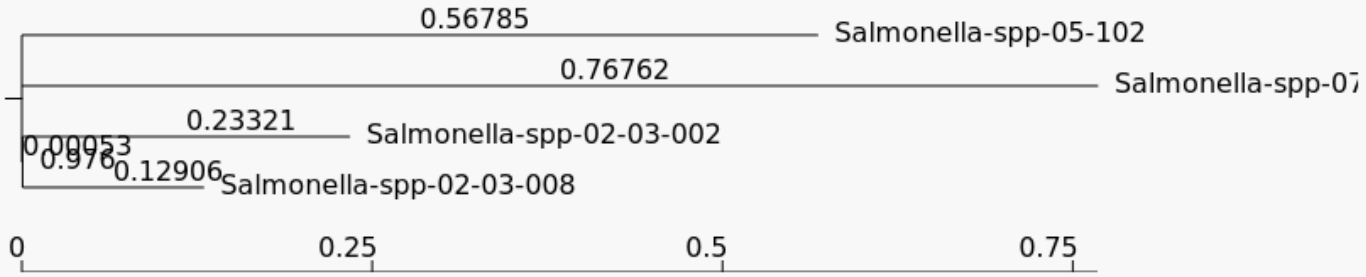
Center for Genomic Epidemiology

Home | Services | Instructions | Output | Article abstract

Mapper: BWA # Submitting 7 jobs. Waiting for vcfwiz.sh to finish... 0

CSIPhylogeny Results

The tree presented in the picture below is only meant as a preview. If the tree is meant to be shared or published, we strongly recommend that the 'Newick' file is downloaded and processed using software created for this purpose. We suggest (FigTree).



Dowload phylogeny as: Newick PDF SVG

Download the filtered SNP calls in Variant Calling Format (VCF):
Note: VCF files are compressed with gzip.
VCF files

Download matrix of SNP pair counts:
Dowload matrix as: TXT EPS

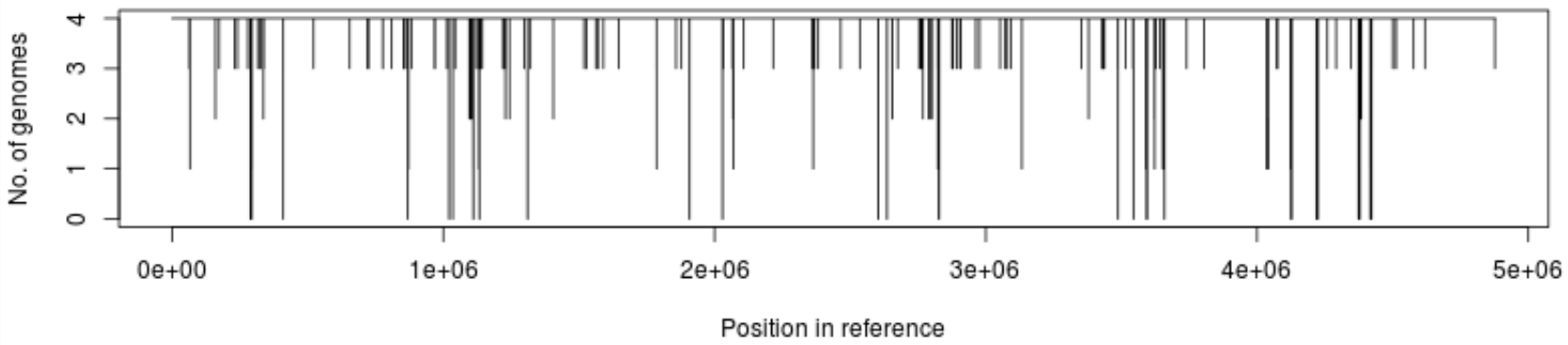
Dowload SNP alignment: FASTA

Percentage of reference genome covered by all isolates: 98.3155029799234
4797128 positions was found in all analyzed genomes.
Size of reference genome: 4879320

Below is listed the number of positions that are shared and trusted between each isolate and the reference genome.

File	Valid positions	Pct. of reference
Salmonella-spp-02-03-008.ignored_snps	4848802	99.3745439938352
Salmonella-spp-02-03-002.ignored_snps	4847669	99.3513235450841
Salmonella-spp-05-102.ignored_snps	4861431	99.6333710435061
Salmonella-spp-07-022.ignored_snps	4821309	98.8110843314232

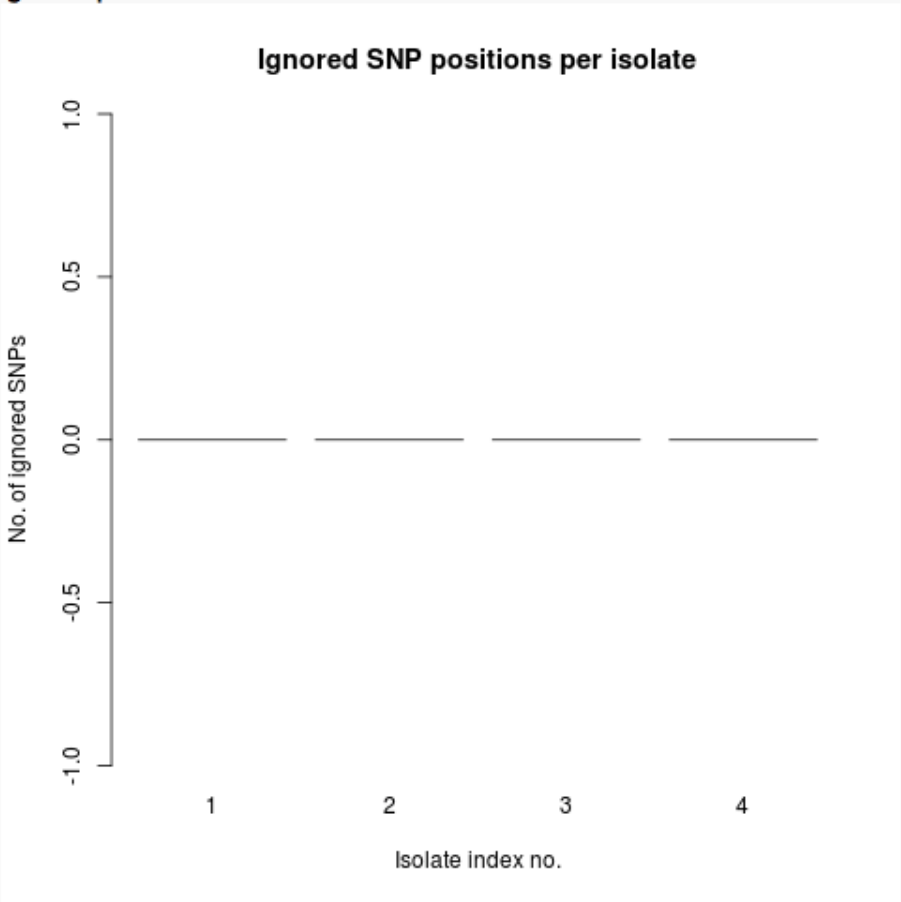
Genomes covering each Position



Download plot:

PDF

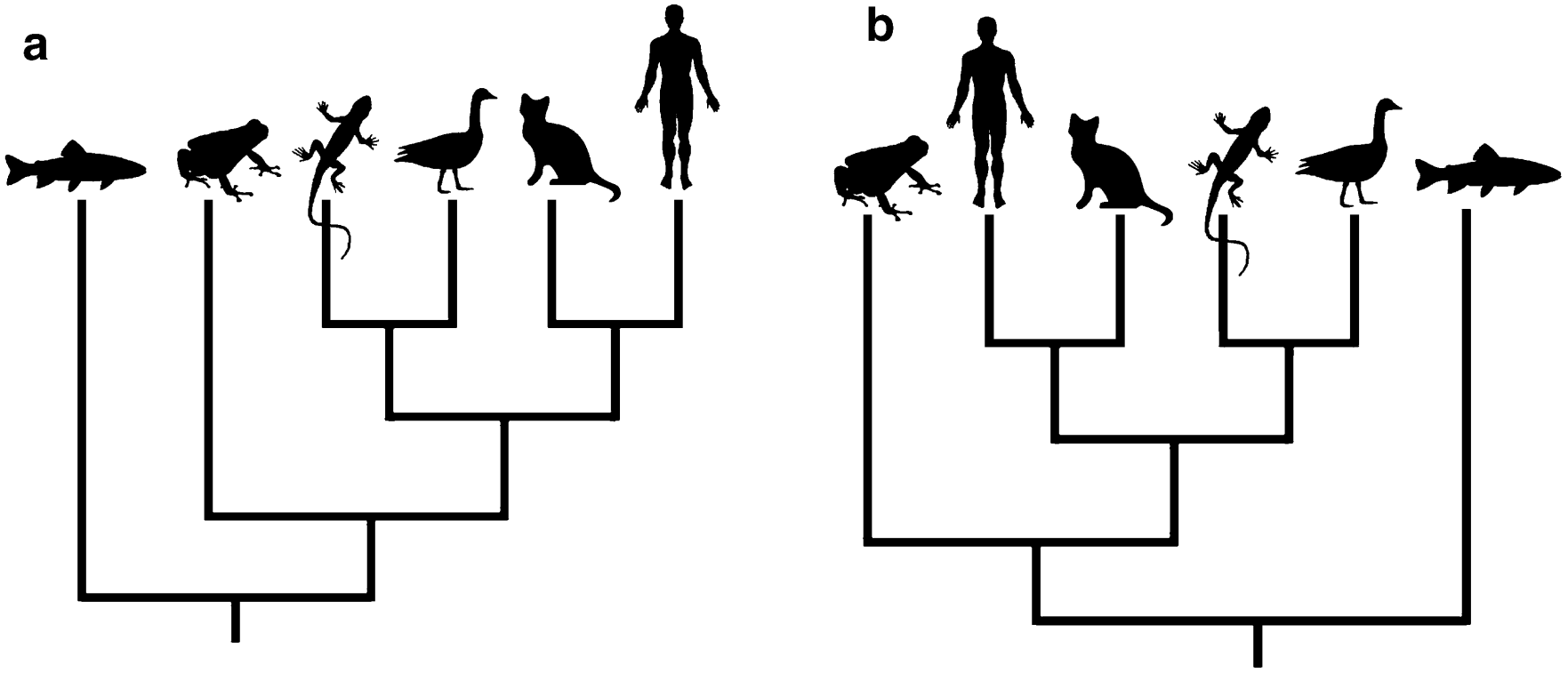
Ignored positions in each isolate



Below is listed the number of positions, covered by at least one read, in each isolate that did not meet the minimum thresholds.

Index	Isolate	Ignored pos.
1	Salmonella-spp-02-03-008	0
2	Salmonella-spp-05-102	0
3	Salmonella-spp-02-03-002	0
4	Salmonella-spp-07-022	0

How to read phylogenetic trees



T. Ryan Gregory. Understanding Evolutionary Trees. Evo Edu Outreach (2008) 1:121–137
DOI 10.1007/s12052-008-0035-x

What is phylogeny used for

- Classify taxonomy – The classic use

- Outbreak detection – Increasing with WGS data

What is phylogeny used for

- Cholera outbreak in Haiti 2010
- Listeria outbreak 2014

  Whole-genome Sequencing Used to Investigate a Nationwide Outbreak of Listeriosis Caused by Ready-to-eat Delicatessen Meat, Denmark, 2014.

  Kvistholm Jensen et al. Clin Infect Dis. (2016) 63 (1): 64-70. doi: 10.1093/cid/ciw192

Case story

- *Vibrio Cholerae* outbreak in Haiti followed the 2010 earthquake
- Rumors said that the outbreak may have come from Nepal, travelling along with UN soldiers from Nepal
- No proof had been given of this until the Hendriksen *et al*. paper in 2011

Population Genetics of Vibrio cholerae from Nepal in 2010: Evidence on the Origin of the Haitian Outbreak. Hendriksen et al. 23 August 2011 mBio vol. 2 no. 4 e00157-11. doi: 10.1128/mBio.00157-11

Case story

- Data
  - 24 recent *V. cholerae* strains from Nepal
  - 10 previously sequenced *V. cholerae* isolates, including 3 from the Haitian outbreak
- Analysis
  - Antimicrobial susceptibility testing
  - PFGE (pulsed-field gel electrophoresis) to analyze for genetic relatedness
  - Whole genome sequencing, SNP identification and phylogenetic analysis

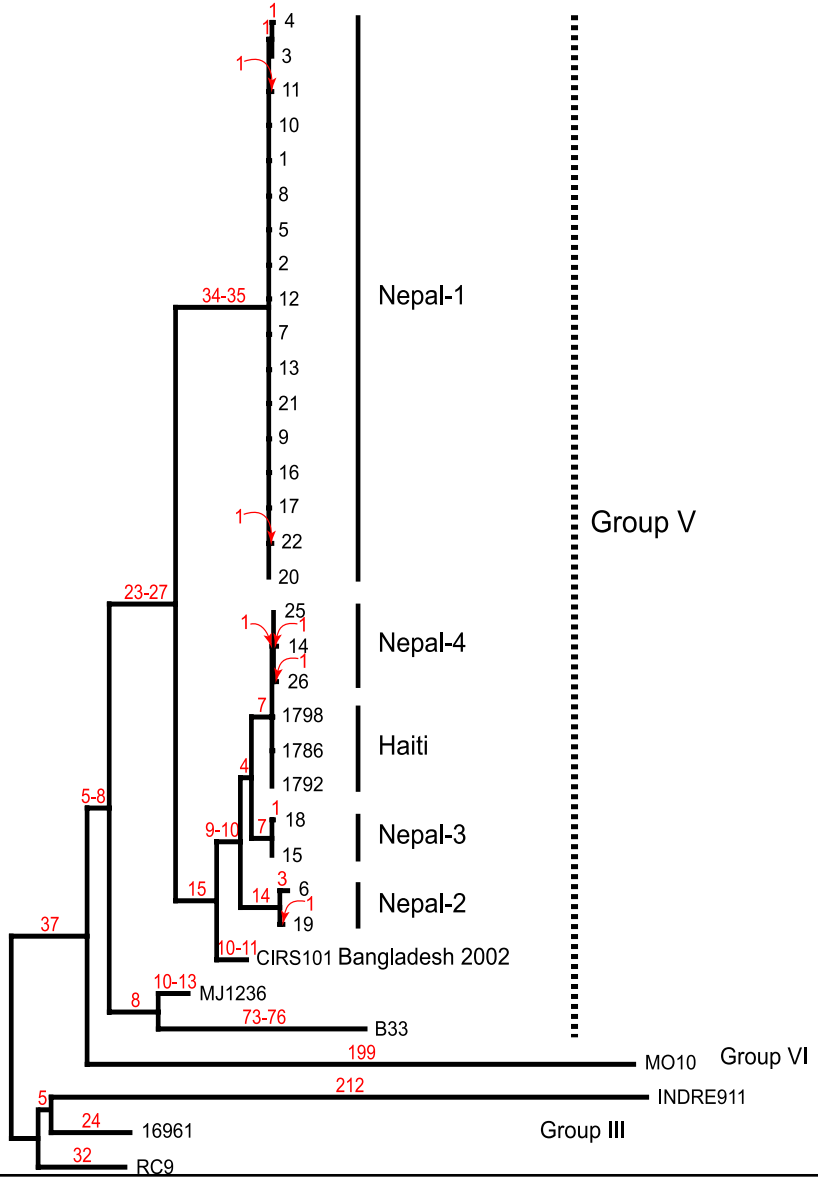
Case story - Results

Resistance profile	Susceptible	Decreased susceptibility	Resistant
Nepalese strains Hendriksen *et al. 2011*	Tetracycline	Ciprofloxacin	Trimethoprim, Sulfamethoxazole Nalidixic
Haitian outbreak strains Centers for Disease Control and Prevention, 2010	Tetracycline	Ciprofloxacin	Trimethoprim, Sulfamethoxazole Nalidixic

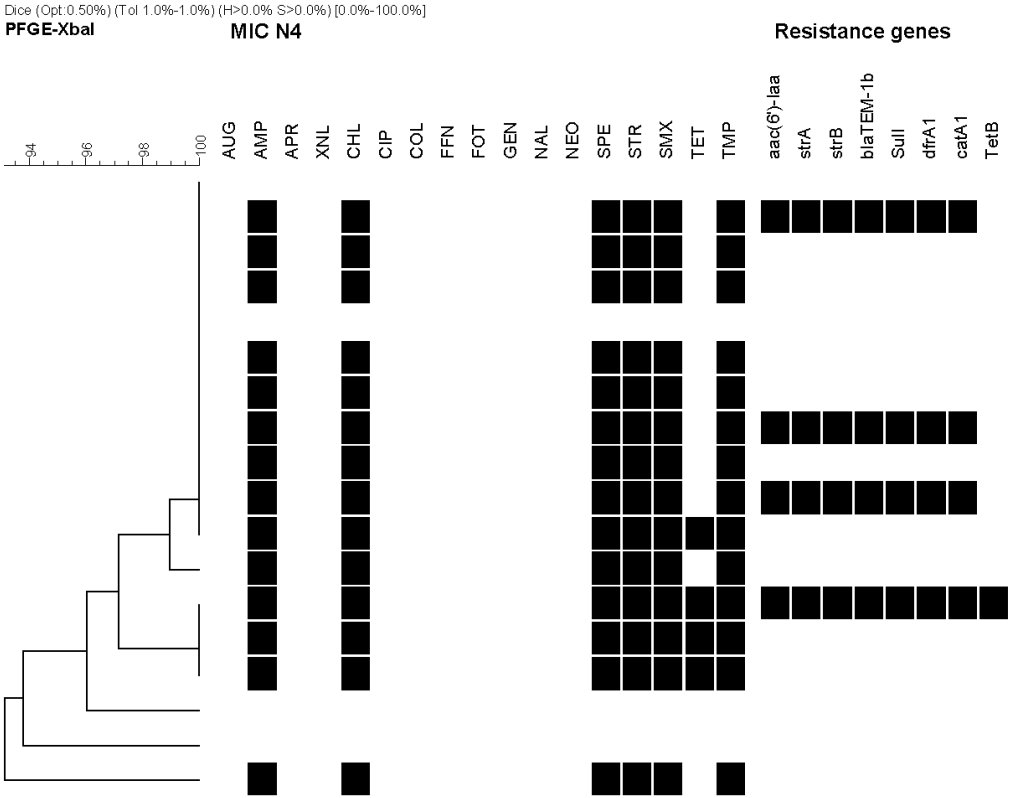
Case story - Results

- Pulsed-field gel electrophoresis (PFGE)
  - Nepalese isolates divided in 4 groups
  - Most common Haitian type in same group as four Nepalese strains

Case story - Results

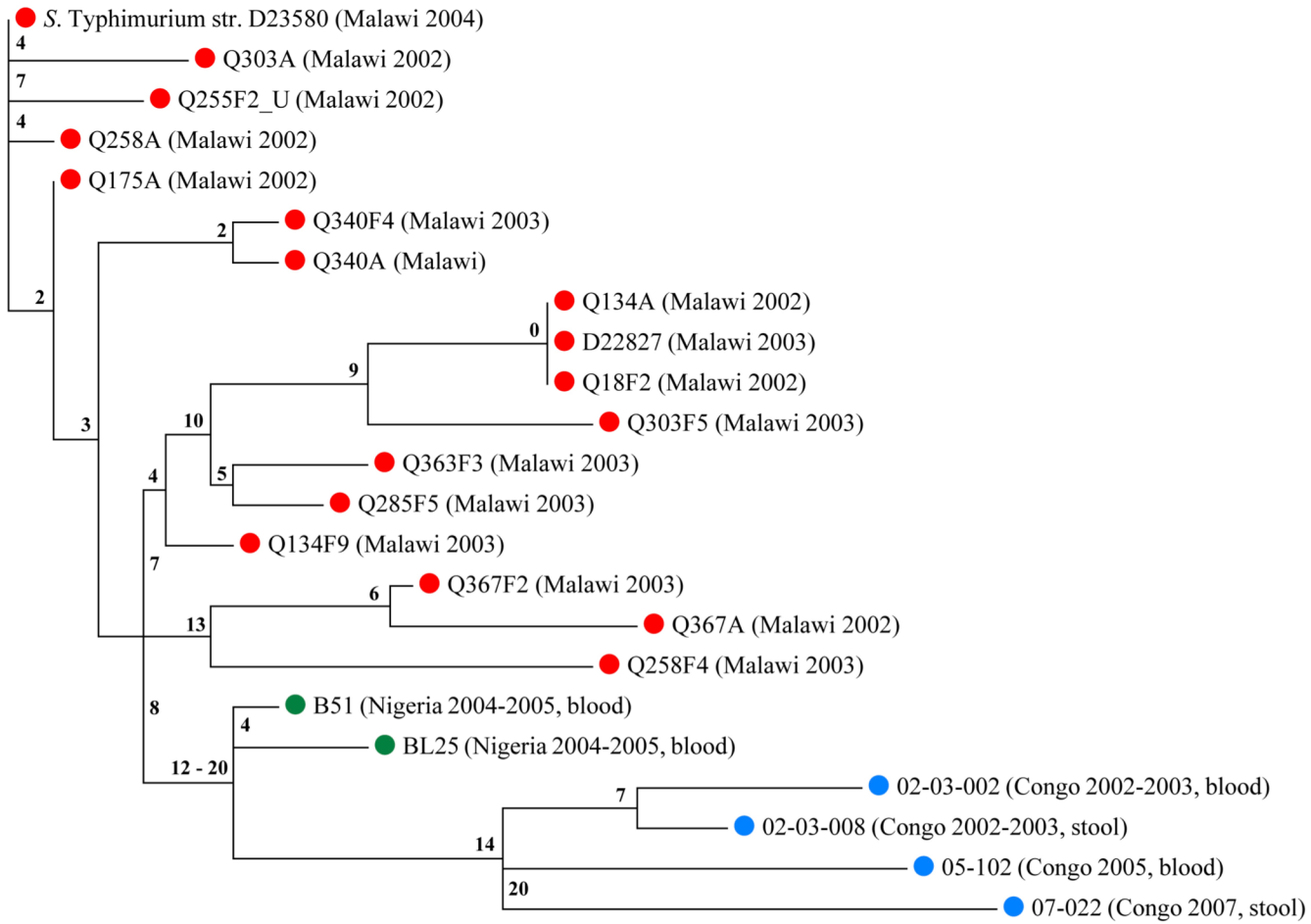


PFGE vs SNP tree



Key	Country of isolation	Year of isolation	SourceType	Seq.
02-03/002				
02-03/008	Congo	2002-2003	Stool	+
05/157	Congo	2005	Stool	
05/159	Congo	2005	Stool	
06/001				
06/013	Congo	2006	Stool	
07/015	Congo	2007	Stool	
07/022	Congo	2007	Stool	+
BL	Nigeria	2004-2005	Blood	
BL 51	Nigeria	2004-2005	Blood	+
F 86	Nigeria	2004-2005	Blood	
02-03/028	Congo	2002-2003	Stool	
BL 25	Nigeria	2004-2005	Blood	+
BL 39	Nigeria	2004-2005	Blood	
BL 95	Nigeria	2004-2005	Blood	
05/102				
07/006				
BL 67	Nigeria	2004-2005	Blood	

PFGE vs SNP tree

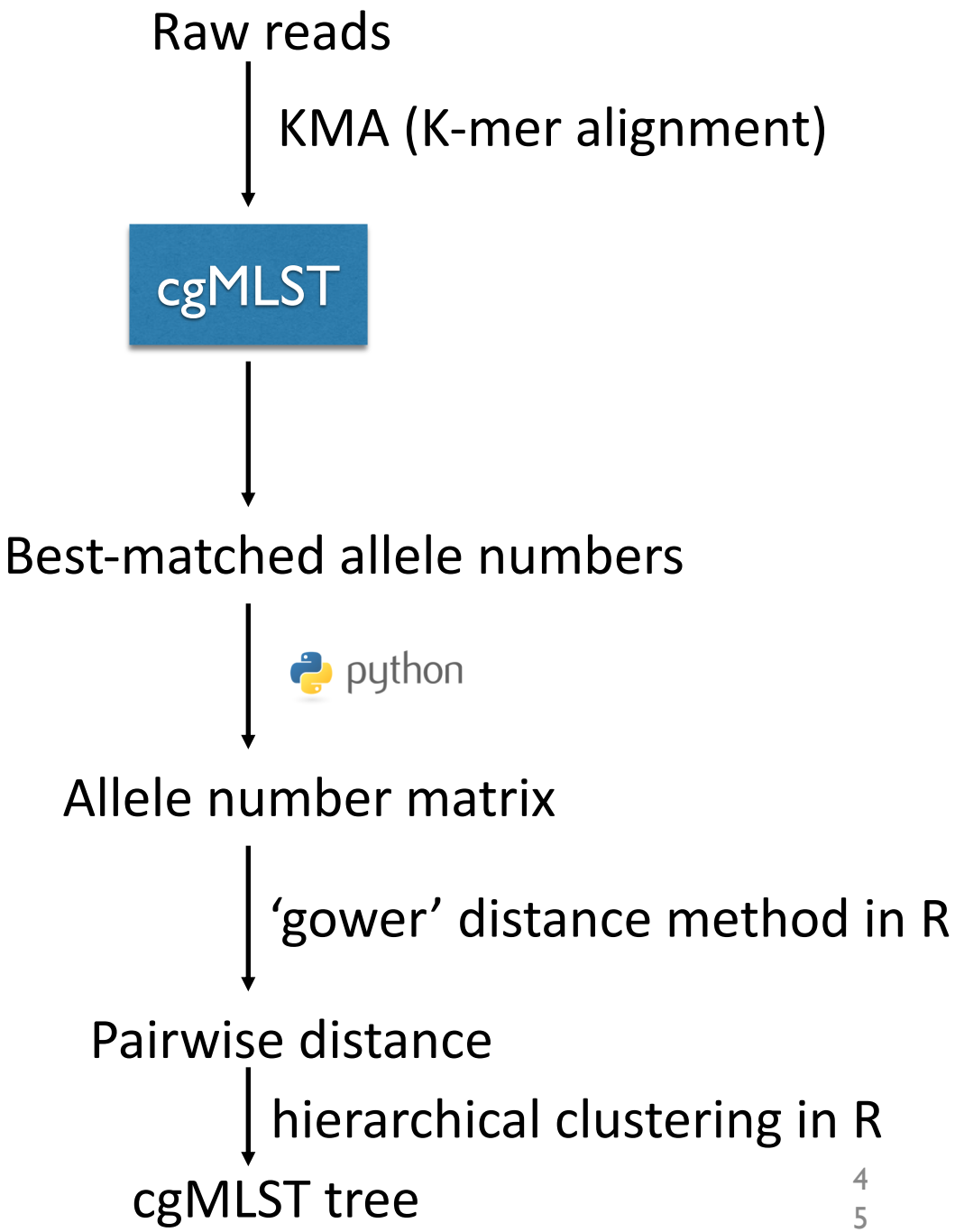


Choosing a reference genome

For comparison of very closely related isolates, a better level of detail is given by using a closely related reference genome.

Would you prefer completed reference genome or draft reference genome (contigs) ?

cgMLST tree


Raw reads
KMA (K-mer alignment)
cgMLST
Best-matched allele numbers
python
Allele number matrix
'gower' distance method in R
Pairwise distance
hierarchical clustering in R
cgMLST tree

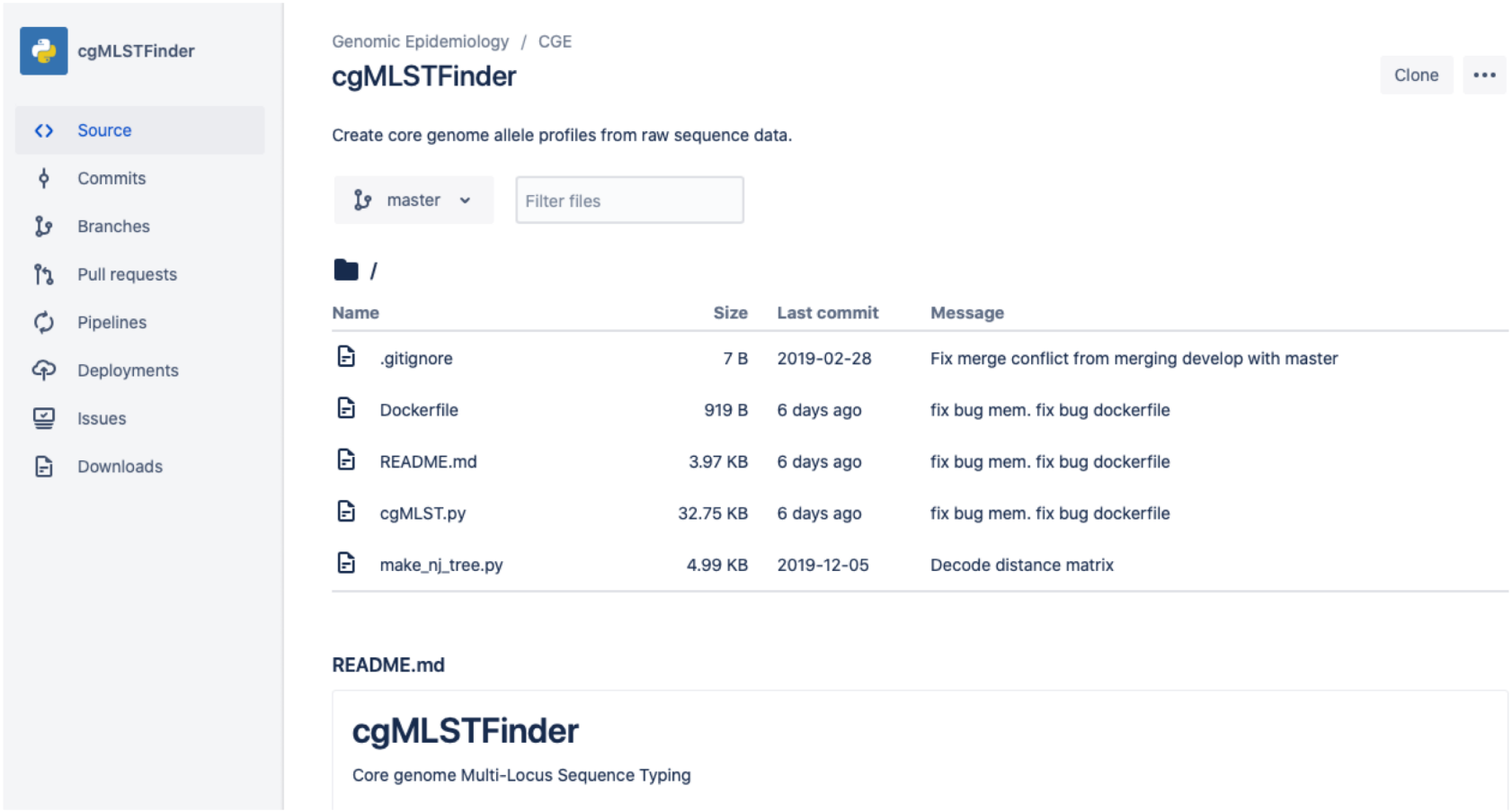
Allele number matrix

Genome	CAMP0001	CAMP0002	CAMP0003	CAMP0006	CAMP0007	CAMP0009	CAMP0010	CAMP0012	CAMP0013	CAMP0015	CAMP0016	CAMP0017	CAMP0018	CAMP0021	CAMP0022
14035391_S3_L001	724	139	202	80	118	81	18	1	363	1	1	1	1	2	96
14035392_S4_L001	1	30	29	1	1	1	1	1	1	1	1	14	1	22	22
14036372-CAM_S93_L001	17	148	21	20	1040	18	18	18	211	17	4	20	1	15	13
14036375-CAM_S19_L001	30	190	277	23	71	1	1	1	1	31	1	32	1	17	32
14038468_S7_L001	573	480	43	159	679	357	35	2	244	181	4	41	1	2	73
14040169_S24_L001	140	124	170	102	168	67	13	77	64	53	33	37	1	41	55
14040195_S28_L001	55	57	60	2	65	22	22	22	225	221	14	49	1	127	21
14040713-CAM_S94_L001	111	111	72	51	81	58	28	31	36	251	25	37	1	33	40
14044103_S26_L001	107	107	29	26	59	48	35	2	1	1	2	46	1	2	2
14044105_S28_L001	95	97	645	31	138	92	20	71	97	22	4	41	1	2	73
14052919_S11_L001	43	31	49	14	51	45	1	1	333	22	24	41	1	2	40
14057476_S17_L001	30	31	33	23	26	23	20	20	208	21	15	24	1	17	1
14058444_S18_L001	135	125	235	23	215	80	32	81	87	28	4	140	1	32	24
14058458_S19_L001	129	122	153	299	621	336	32	40	125	330	4	75	1	28	13
14061139_S21_L001	496	73	82	281	573	102	18	14	106	53	33	37	1	41	55
14061942_S22_L001	1	31	33	79	119	1	1	1	84	1	1	28	1	89	67
C140112_S24_L001	43	31	49	14	51	45	1	1	333	22	24	41	1	2	40
C140316_S27_L001	17	127	21	20	165	18	53	18	18	17	4	20	1	72	13
C140319_S30_L001	17	19	214	20	21	18	18	18	18	17	4	20	1	15	13
C140369_S17_L001	1	1	1	1	1	12	13	2	1	81	2	83	1	2	2
C140476_S18_L001	578	109	132	31	153	95	20	29	99	28	4	88	18	28	13
C140575-CAM_S23_L001	17	148	21	20	165	18	18	18	18	17	4	20	1	15	13
C140637_S22_L001	287	67	73	1	347	30	1	27	28	28	4	29	1	23	24
C140660_S23_L001	13	13	14	14	14	12	13	14	104	1	1	177	1	2	13
C140662-CAM_S59_L001	669	24	157	51	828	42	32	40	97	2	2	2	1	2	107
C140665-CAM_S95_L001	62	66	72	51	81	58	28	31	36	22	25	37	1	27	49
C140695-CAM_S84_L001	31	58	61	35	67	37	29	42	37	33	4	25	1	28	19
DTU2016_1950_PRJ1085_Campylobact	565	39	406	1	1	1	1	1	1	1	2	2	1	2	2
DTU2016_1951_PRJ1085_Campylobact	68	23	25	2	84	22	22	22	22	20	14	187	1	17	1
DTU2016_1959_PRJ1085_Campylobact	27	27	29	26	59	48	35	2	1	1	2	46	1	2	2
DTU2016_1960_PRJ1085_Campylobact	762	195	796	51	365	157	54	72	107	1	1	1	1	28	139
DTU2016_1961_PRJ1085_Campylobact	20	23	25	2	25	22	22	22	22	20	14	23	1	532	21
DTU2016_1962_PRJ1085_Campylobact	535	436	1488	670	1485	94	20	29	228	724	130	239	1	2	13
DTU2016_1963_PRJ1085_Campylobact	43	31	49	14	51	45	1	1	45	22	24	41	1	2	40
DTU2016_1964_PRJ1085_Campylobact	32	2	25	602	265	81	18	109	172	772	80	80	1	119	426
DTU2016_1965_PRJ1085_Campylobact	43	31	49	14	51	45	1	1	45	22	24	41	1	2	40

cgMLST_table.txt

Genome	AEJV01_03887	C_RS24035	C_RS24040	EAKF1_RS07940	ECs0267	ECs4266	ECs4518
110_R1	2	143	1	11	43	76	71
125_R1	2	143	1	11	43	76	71
038_R1	2	143	1	11	43	76	71
049_R1	2	143	1	10	43	76	71
005_R1	2	143	1	11	43	76	71
008_R1	2	143	1	44	43	76	71
089_R1	2	143	1	11	43	76	71
098_R1	2	143	1	44	43	76	71
011_R1	2	143	1	11	43	76	71
036_R1	2	143	1	11	43	76	71

https://bitbucket.org/genomicepidemiology/cgmlstfinder/src/master/



```
---
title: 'cgMLST tree'
author: "Shinny Pimlapas Leekitcharoenphon"
date: "September 21, 2018"
output:
  html_document:
    theme: sandstone
    code_folding: hide
---

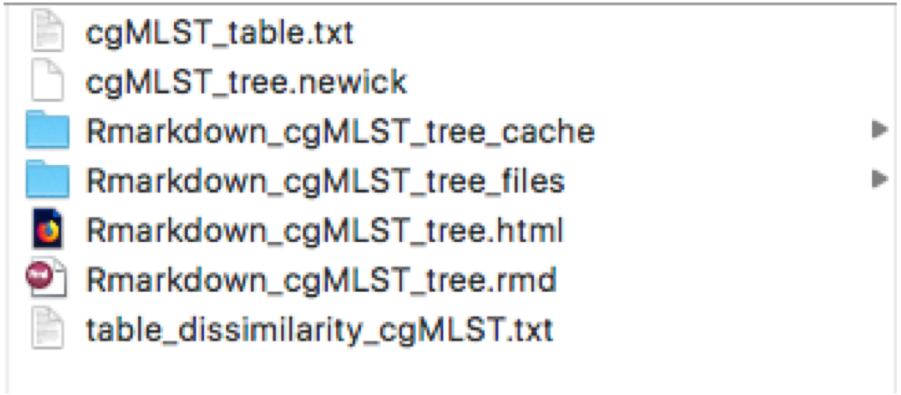

```{r init, message=FALSE}
knitr::opts_chunk$set(cache = TRUE, autodep = TRUE, warning=FALSE, message=FALSE)
#install.packages("cluster")
library(cluster)
#install.packages("ape")
library(ape)
#install.packages("reshape")
library(reshape)
```
cgMLST tree (distance not included) {#E2A}
```{r cgMLST tree, eval=TRUE}
data <- read.table("cgMLST_table.txt", sep = "\t", row.names=1, colClasses = "factor", header = T)
cgMLST_tree <- as.phylo(hclust(daisy(data, metric="gower")))
write.tree(phy=cgMLST_tree, file="cgMLST_tree.newick")
plot(hclust(daisy(data, metric="gower")))

```

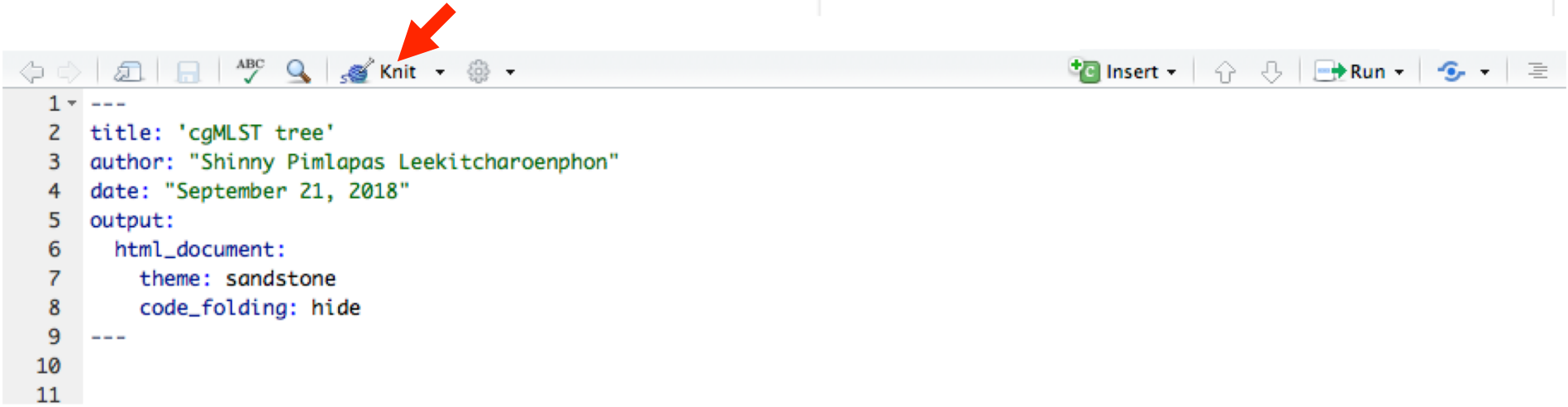

cgMLST dissimilarity table {#E2A}
```{r dissimilarity tree, eval=TRUE}

m <- as.matrix(daisy(data, metric="gower"))
m2 <- melt(m)[melt(upper.tri(m))$value,]
names(m2) <- c("c1", "c2", "distance")
m
m2
write.table(m, 'table_dissimilarity_cgMLST.txt', sep='\t')
```
```

Install following packages in R before running the script

```
install.packages("cluster")
install.packages("ape")
install.packages("reshape"
```





~/Meetings/175_EURL_September2018/exercise3/cgMLST/c

Rmarkdown_cgMLST_tree.html | Open in Browser | Find

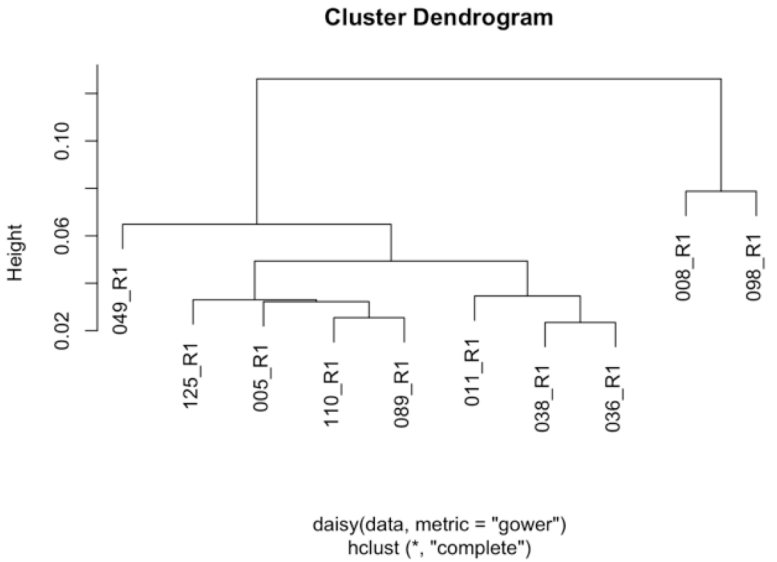
cgMLST tree

Shinny Pimlapas Leekitcharoenphon

September 21, 2018

cgMLST tree (distance not included)

- cgMLST_table.txt
- cgMLST_tree.newick
- Rmarkdown_cgMLST_tree_cache
- Rmarkdown_cgMLST_tree_files
- Rmarkdown_cgMLST_tree.html
- Rmarkdown_cgMLST_tree.rmd
- table_dissimilarity_cgMLST.txt



cgMLST dissimilarity table

CODE

```
##            110_R1     125_R1     038_R1     049_R1     005_R1     008_R1
## 110_R1 0.00000000 0.02666136 0.03461998 0.04974135 0.02785515 0.11221647
## 125_R1 0.02666136 0.00000000 0.04058894 0.05531238 0.03263032 0.11778750
## 038_R1 0.03461998 0.04058894 0.00000000 0.05531238 0.03302825 0.12017509
## 049_R1 0.04974135 0.05531238 0.05531238 0.00000000 0.04974135 0.12614405
```

table_dissimilarity_cgMLST.txt

110_R1	125_R1	038_R1	049_R1	005_R1	008_R1	089_R1	098_R1	011_R1	036_R1	
110_R1	0	0.026661361	0.034619976	0.049741345	0.027855153	0.112216474	0.025467569	0.097095105	0.042180661	0.030242738
125_R1	0.026661361	0	0.040588938	0.055312376	0.032630322	0.117787505	0.033028253	0.102666136	0.047751691	0.035415838
038_R1	0.034619976	0.040588938	0	0.055312376	0.033028253	0.12017509	0.037405491	0.105053721	0.034222045	0.023477915
049_R1	0.049741345	0.055312376	0.055312376	0	0.049741345	0.126144051	0.05252686	0.111022682	0.064862714	0.050935137
005_R1	0.027855153	0.032630322	0.033028253	0.049741345	0	0.11539992	0.032232392	0.101074413	0.041384799	0.029844807
008_R1	0.112216474	0.117787505	0.12017509	0.126144051	0.11539992	0	0.116195782	0.07879029	0.124552328	0.114206128
089_R1	0.025467569	0.033028253	0.037405491	0.05252686	0.032232392	0.116195782	0	0.101472344	0.049343414	0.034222045
098_R1	0.097095105	0.102666136	0.105053721	0.111022682	0.101074413	0.07879029	0.101472344	0	0.111022682	0.101074413
011_R1	0.042180661	0.047751691	0.034222045	0.064862714	0.041384799	0.124552328	0.049343414	0.111022682	0	0.034619976
036_R1	0.030242738	0.035415838	0.023477915	0.050935137	0.029844807	0.114206128	0.034222045	0.101074413	0.034619976	0

Multiply dissimilarity value with total loci,
you will get actual number of allele difference

SNP vs gene-by-gene approach

No nomenclature (SNP addresses has been indroduced by PHE)	cgST type
Comparability	Comparability
- Only by using the same reference genome	- higher as there is only one or a few cgMLST scheme
- Only by using the same SNP pipeline and parameters	
No limitation in term of reference genome	Limition by avaiable scheme
	- Public scheme have been developed and are maintained for many
	but not all pathogens
Restriced to regions of the genome present in all analyzed genomes	
useful information in the accessory genome is discarded	
Not suitable in scenarios where plasmid-mediated rather than clonal outbreak	

A.C. Church et al. Clinical Microbiology and Infection 24 (2018) 350-354

What defines an outbreak

- We can’t tell for certain
- It depends on the species and clone
- But a rule of thump is:
  - Within 10 SNPs it is definitely an outbreak
  - Within 30 SNPs it might be an outbreak
  - Above 100 SNPs it is most likely not an outbreak

Isolate relatedness

Table 1
Examples of relatedness criteria for wg/cgMLST and SNP typing schemes of representative clinically relevant bacteria

Organism	Relatedness threshold[a]		References
	wg/cgMLST (allele)	SNPs	
Acinetobacter baumannii	≤8	≤3	[25,26]
Brucella spp.	Epidemiologic validation in progress[b]		http://www.applied-maths.com/applications/wgmlst
Campylobacter coli, C. jejuni	≤14	≤15	[27,28]
Cronobacter spp.	Epidemiologic validation in progress[b]		http://www.applied-maths.com/applications/wgmlst
Clostridium difficile	Epidemiologic validation in progress[b]	≤4	[29], http://www.cgmlst.org/ncs, http://www.applied-maths.com/applications/wgmlst
Enterococcus faecium	≤20	≤16	[30]
Enterococcus raffinosus	Epidemiologic validation in progress[b]		http://www.applied-maths.com/applications/wgmlst
Escherichia coli	≤10	≤10	[31,32], https://enterobase.warwick.ac.uk/
Francisella tularensis	≤1	≤2	[33,34]
Klebsiella oxytoca	Epidemiologic validation in progress[b]		http://www.applied-maths.com/applications/wgmlst
Klebsiella pneumonia	≤10	≤18	[35,36]
Legionella pneumophila	≤4	≤15	[37]
Listeria monocytogenes	≤10	≤3	[38,39]
Mycobacterium abscessus		≤30	[40]
Mycobacterium tuberculosis	≤12	≤12	[41]
Neisseria gonorrhoeae	Epidemiologic validation in progress[b]	≤14	[42], http://www.applied-maths.com/applications/wgmlst
Neisseria meningitidis	Epidemiologic validation in progress[b]		http://www.cgmlst.org/ncs
Pseudomonas aeruginosa	≤14	≤37	[31,43]
Salmonella dublin	Epidemiologic validation in progress[b]	≤13	[44], https://enterobase.warwick.ac.uk/
Salmonella enterica	Epidemiologic validation in progress[b]	≤4	[45], http://www.cgmlst.org/ncs, http://www.applied-maths.com/applications/wgmlst, https://enterobase.warwick.ac.uk/
Salmonella typhimurium	Epidemiologic validation in progress[b]	≤2	[46], https://enterobase.warwick.ac.uk/
Staphylococcus aureus	≤24	≤15	[47,48]
Streptococcus suis		≤21	[49]
Vibrio parahaemolyticus	≤10		[50]
Yersinia spp.	0		[51]

A.C. Church et al. Clinical Microbiology and Infection 24 (2018) 350-354

Isolate relatedness

TABLE 1 | Maximum pairwise SNPs measured during investigations into foodborne illness outbreaks and contamination events.

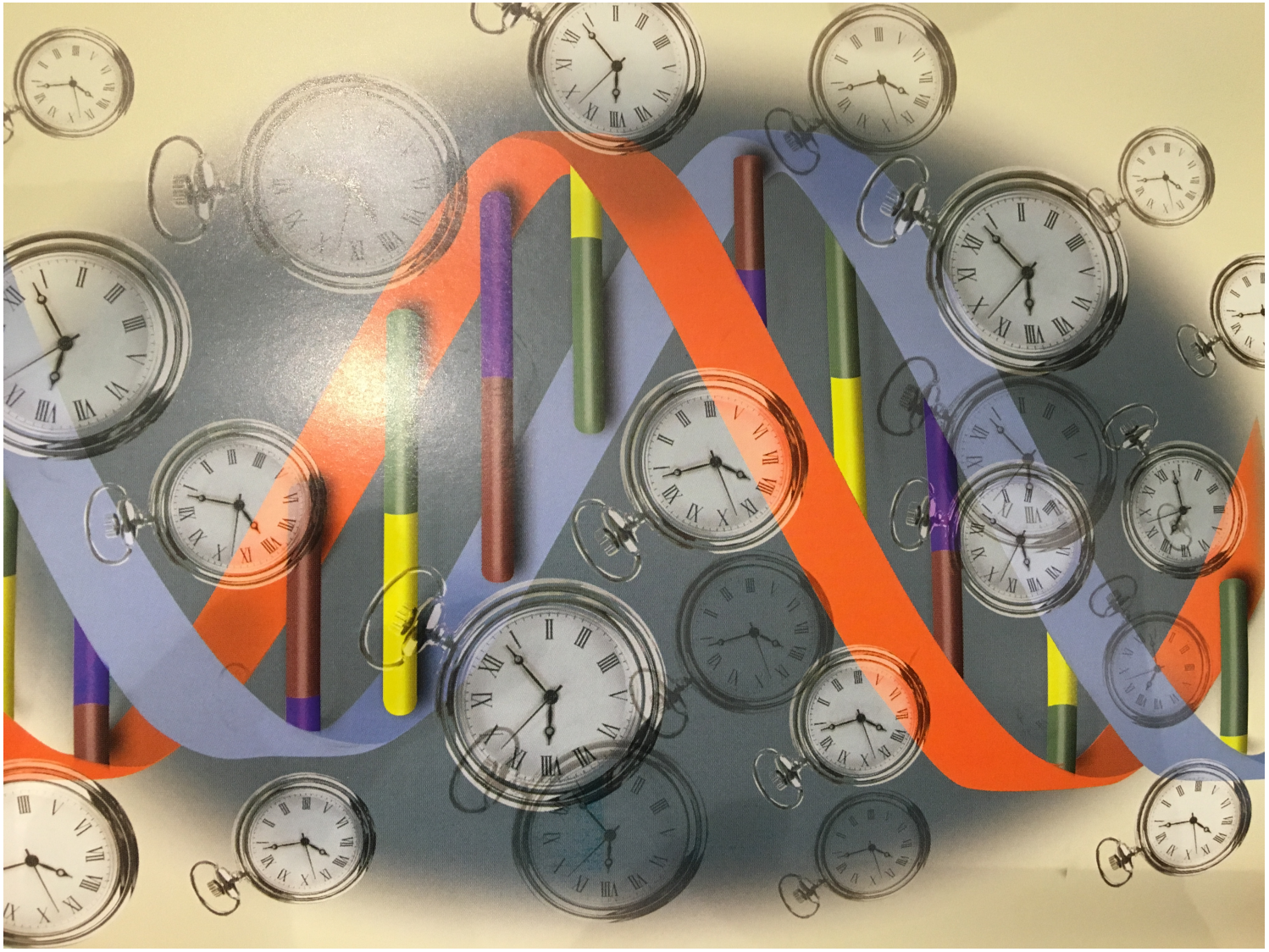
Organism	Maximum SNP count (number)	Maximum SNP count (range)			Reference
		<21	21–100	>100	
E. coli	4	X			Underwood et al., 2013
E. coli	15	X			Eppinger et al., 2011
L. monocytogenes	9	X			Chen et al., 2017c
L. monocytogenes	12	X			Chen et al., 2017a
L. monocytogenes	18	X			Li et al., 2017
L. monocytogenes	20	X			Wang et al., 2015
L. monocytogenes	21		X		Nielsen et al., 2017
L. monocytogenes	28		X		Gilmour et al., 2010
L. monocytogenes	29		X		Chen et al., 2017b
L. monocytogenes	42		X		Chen et al., 2016
L. monocytogenes	67		X		Jackson et al., 2016
S. enterica	2	X			Wuyts et al., 2015
S. enterica	3	X			Allard et al., 2016
S. enterica	3	X			Taylor et al., 2015
S. enterica	6	X			Hoffmann et al., 2016
S. enterica	12	X			Octavia et al., 2015
S. enterica	30		X		Leekitcharoenphon et al., 2014

The maximum SNP counts for isolates that were traced back to the same source in the original study are presented. Whether the maximum SNP counts are less than 21 SNPs, 21 to 100 SNP, or greater than 100 SNPs is also indicated.

Pightling AW et al. Front. Microbiol. 9:1482 (2018)

Phylogenetic tree with space and time

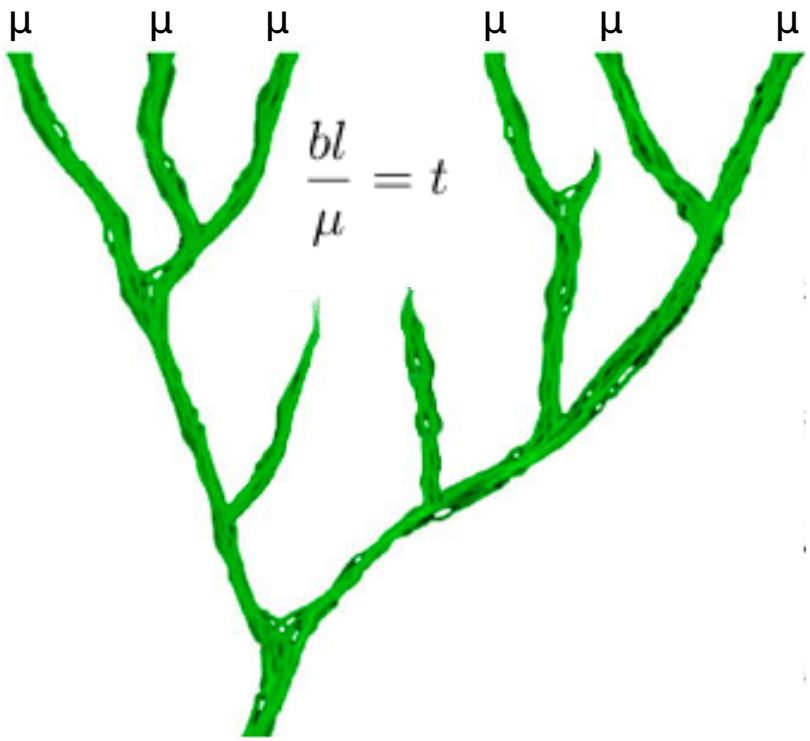
BEAST



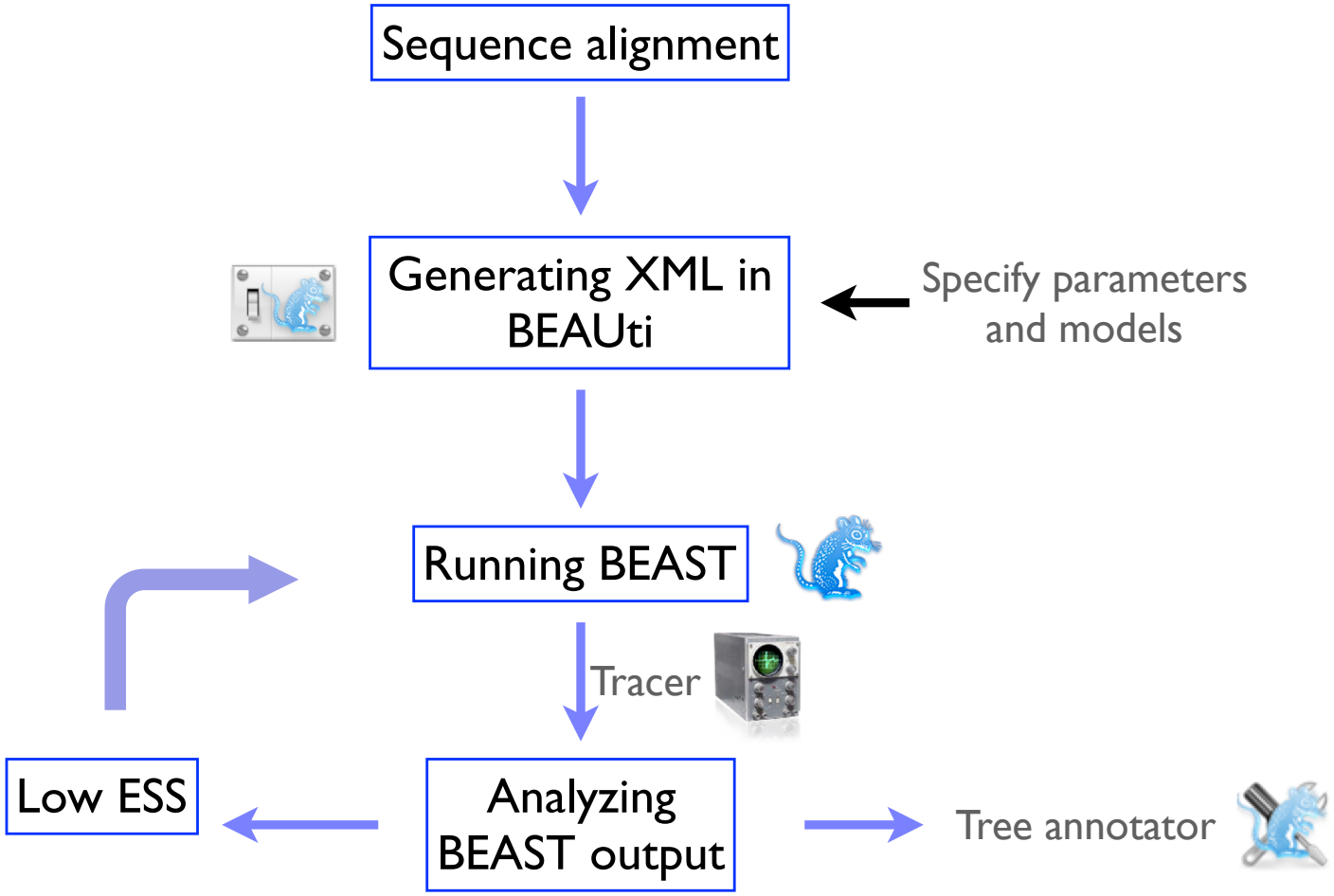
BEAST

- BEAST - Bayesian Evolutionary Analysis Sampling Trees

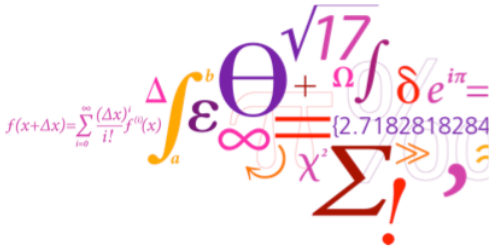
- BEAST is a program to reconstruct and date phylogenetic tree



BEAST workflow



Genomic epidemiology of the global occurrence *S.* Typhimurium DT104



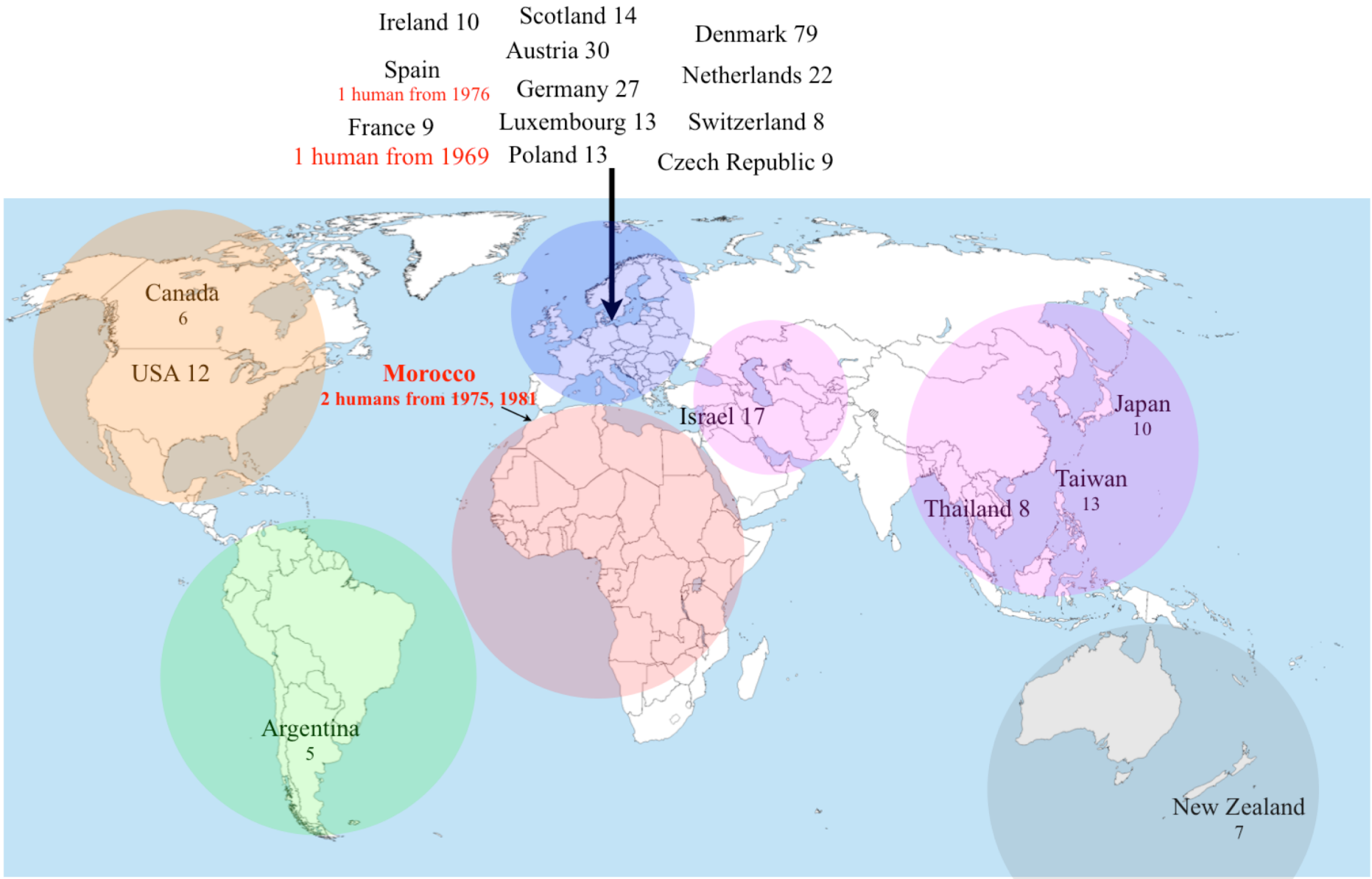
S. Typhimurium DT104

- During the last three decades, *S.* Typhimurium phage type DT104 emerged as the most important phage type and one of the best-studied because of its rapid global dissemination [Lan R, *et al.* Infect Genet Evol. 2009] [Helms M, *et al.* Emerg Infect Dis. 2005]

- DT104 has a multiple antimicrobial resistance pattern to ampicillin, chloramphenicol, streptomycin, sulphonamide and tetracycline (ACSSuT) [Mulvey MR, *et al.*Microbes Infect. 2006]

- Previous epidemics with MDR phage types of *S.* Typhimurium, such as DTs 29, 204, 193 and 204c, were mostly restricted to cattle [Threlfall EJ. J Antimicrob Chemother. 2000]

- DT104 spread among all domestic animals including cattle, poultry, pigs and sheep [Threlfall EJ. J Antimicrob Chemother. 2000]



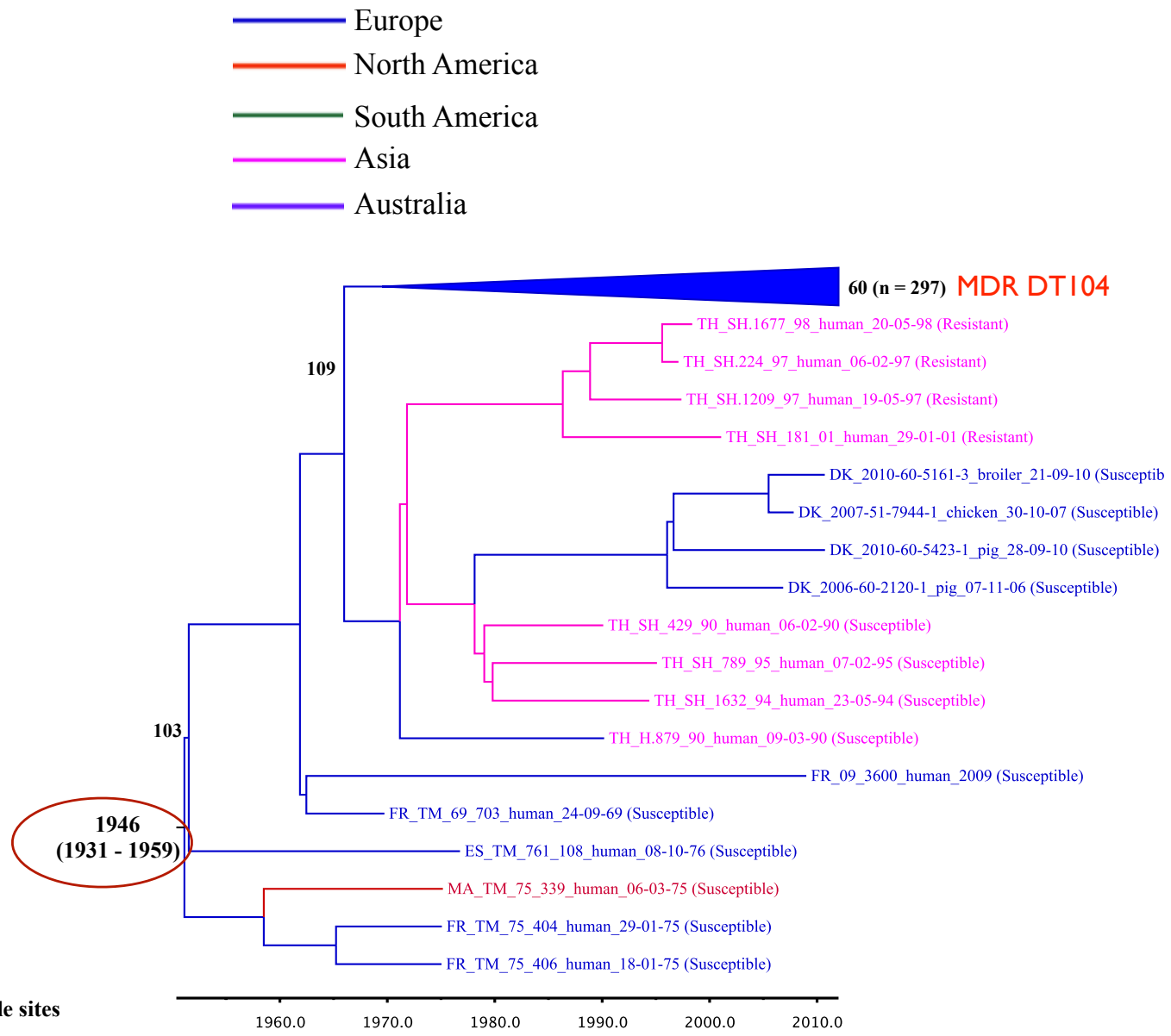
315 genomes

197 animal isolates

118 human isolates

x

Global phylogeny of DT104

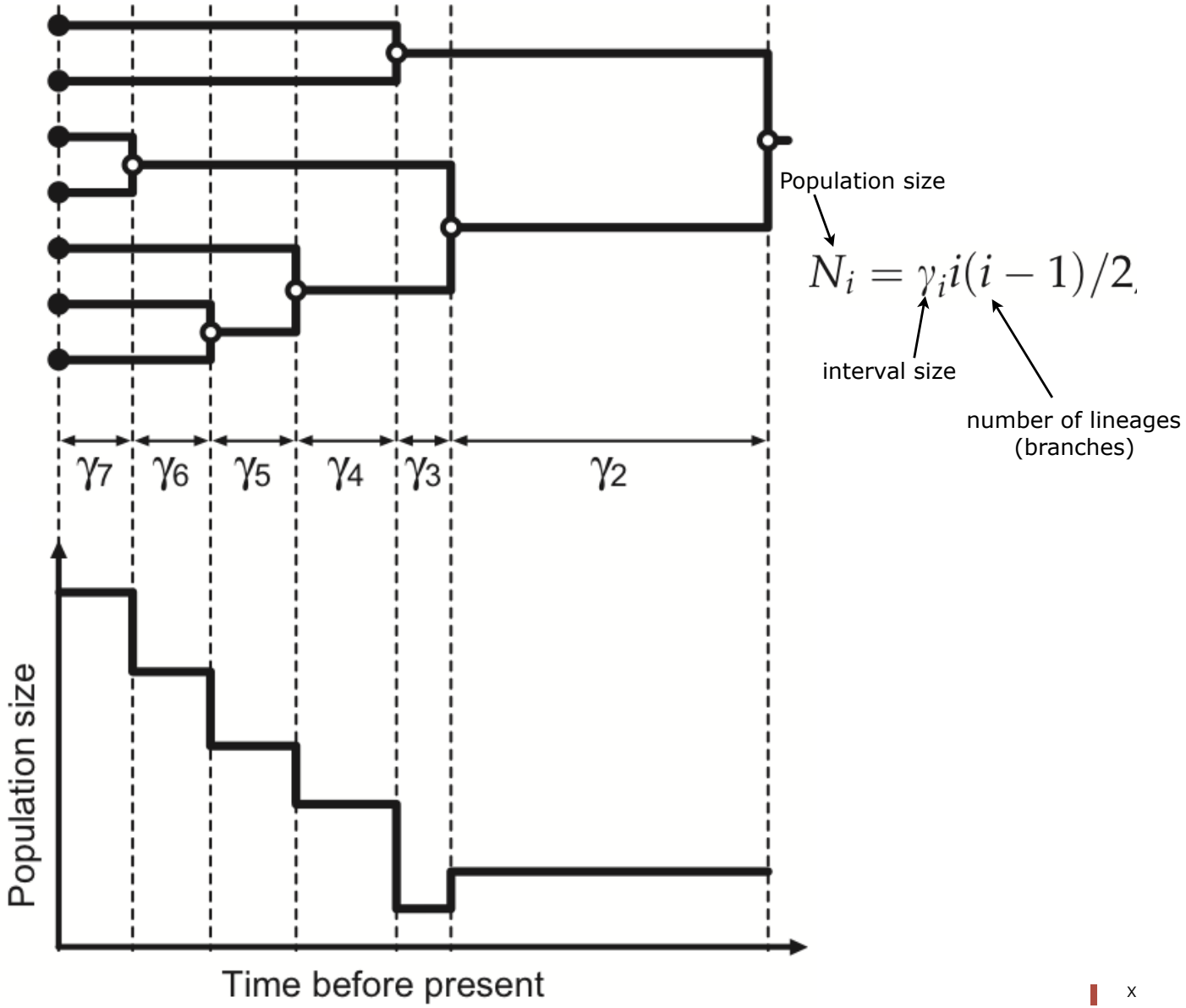


4619 variable sites

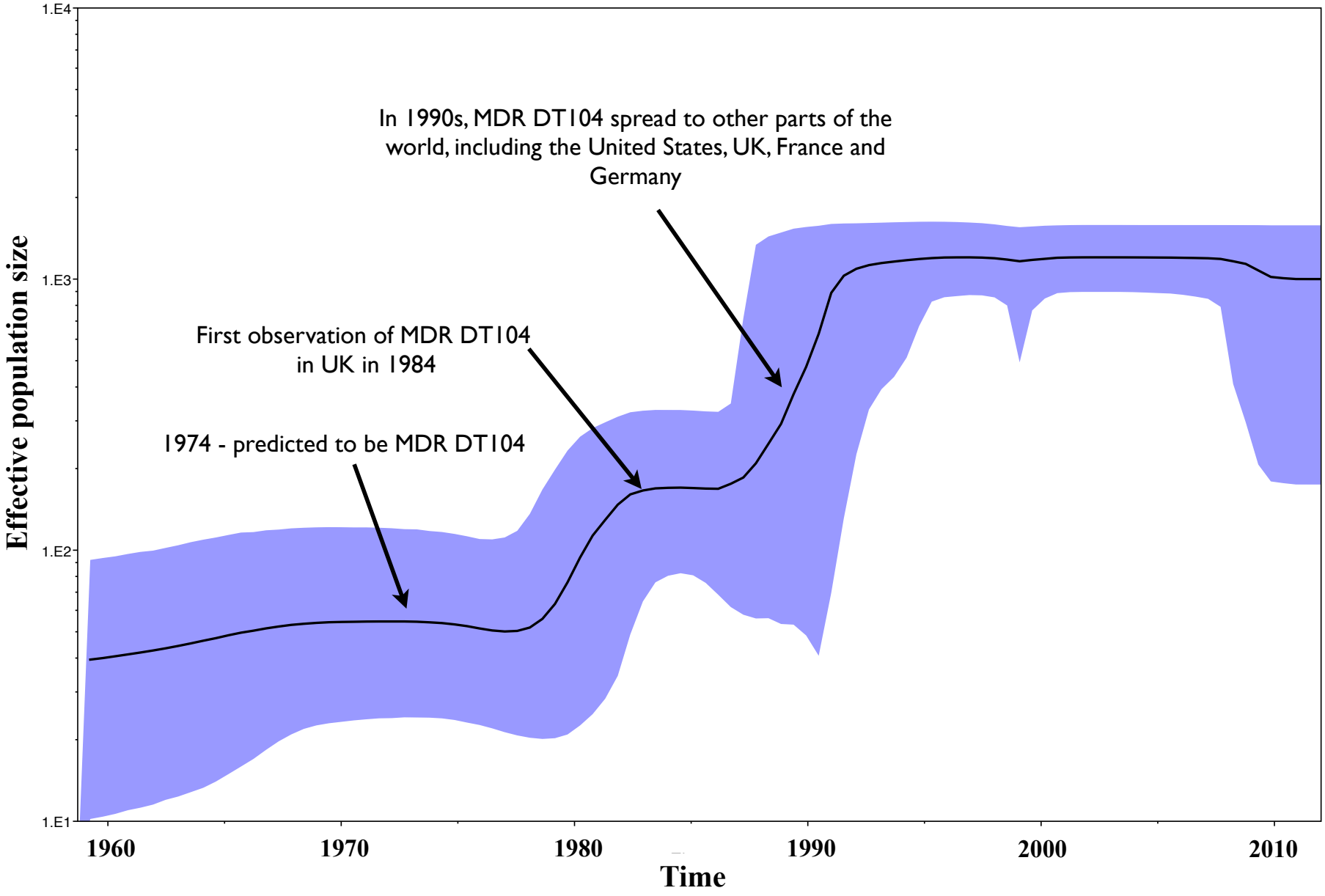
2.97×10^{-7} SNP/site/year ~ 1.47 SNP/year

Leekitcharoenphon P. *et al*, Appl Environ Microbiol. 2016

Demographic history using Bayesian skyline plot

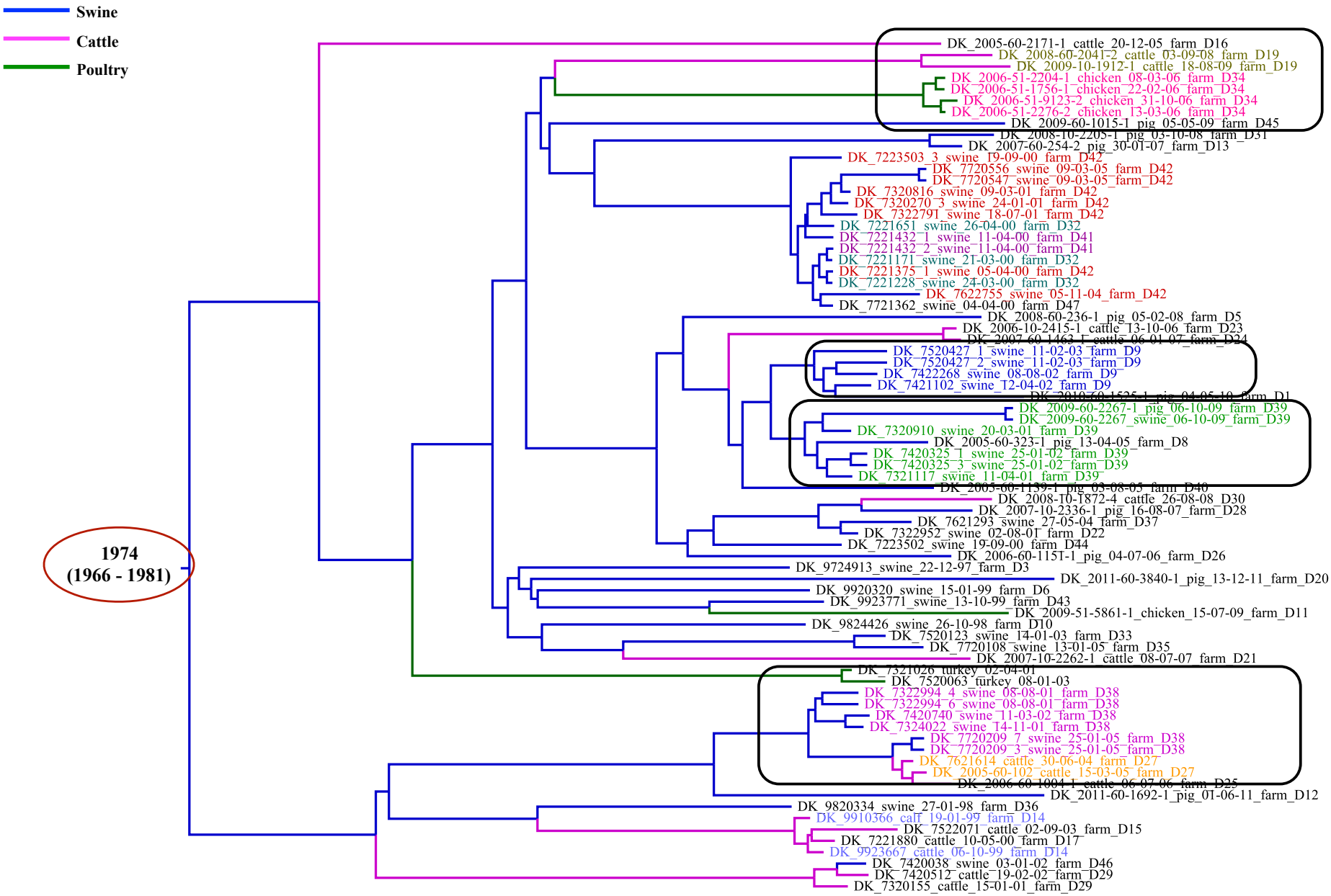


Demographic history of global DT104



Leekitcharoenphon P. *et al*, Appl Environ Microbiol. 2016

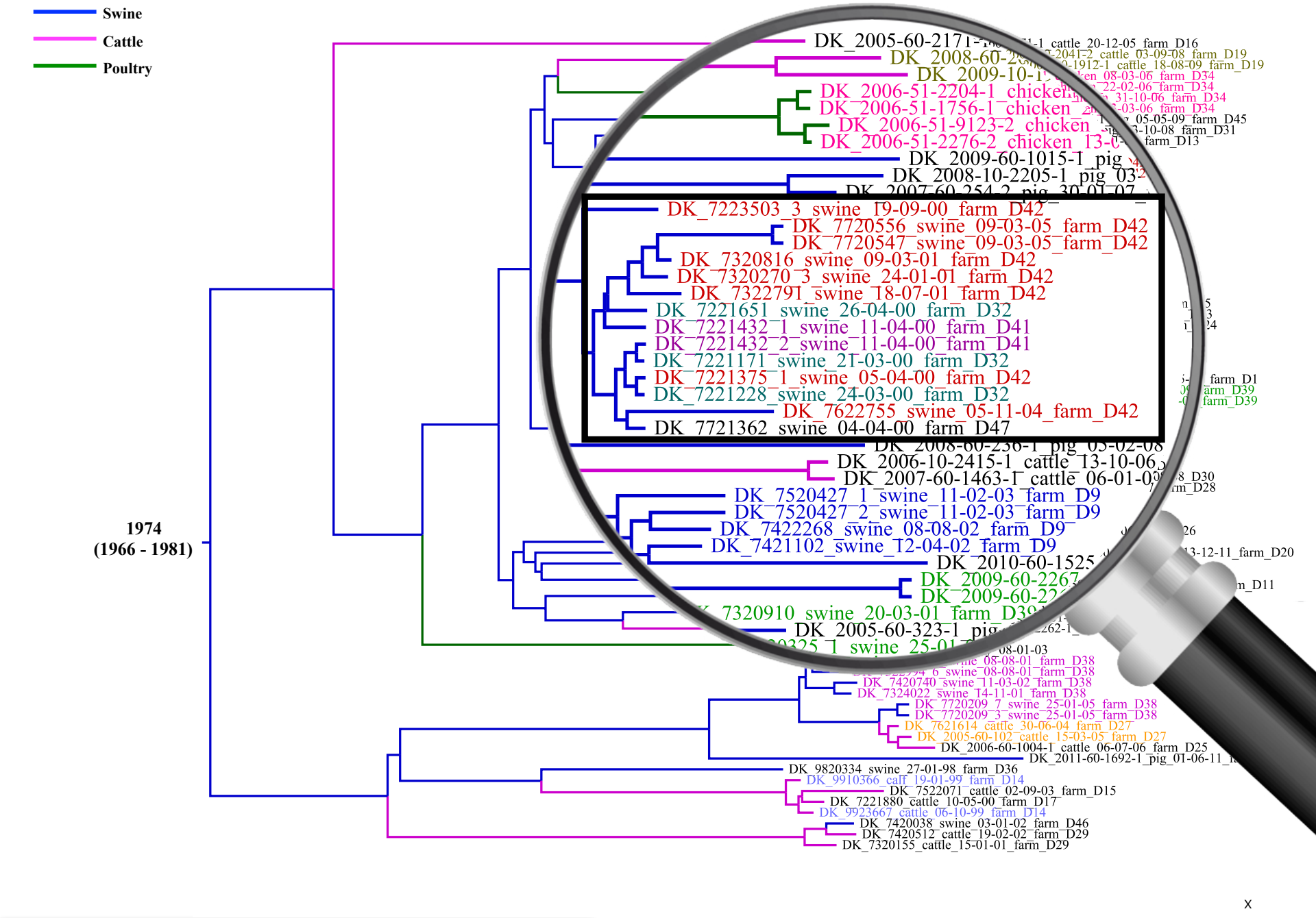
Local phylogeny of DT104 (Denmark)



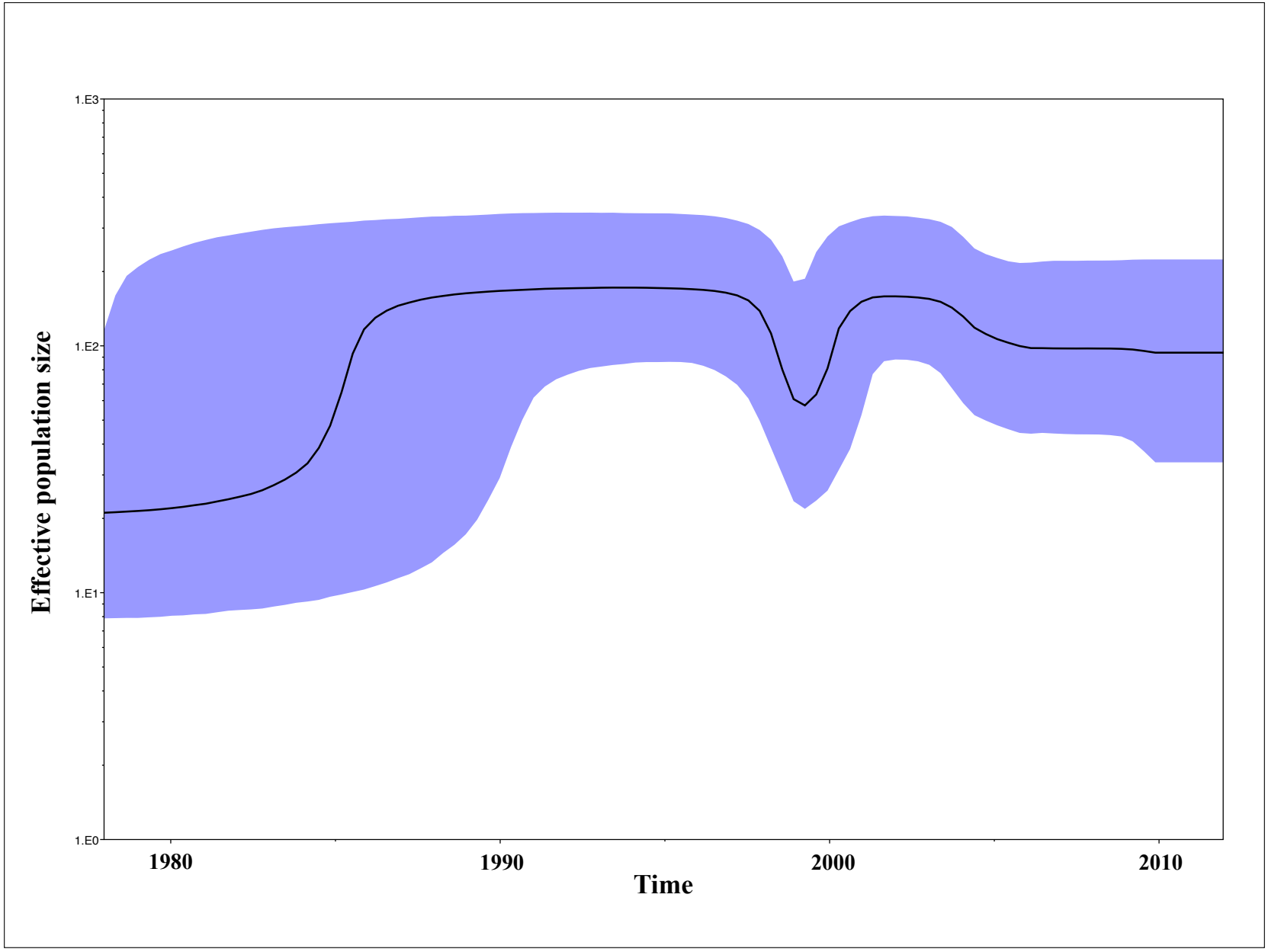
Leekitcharoenphon P. *et al*, Appl Environ Microbiol. 2016

x

Local phylogeny of DT104 (Denmark)

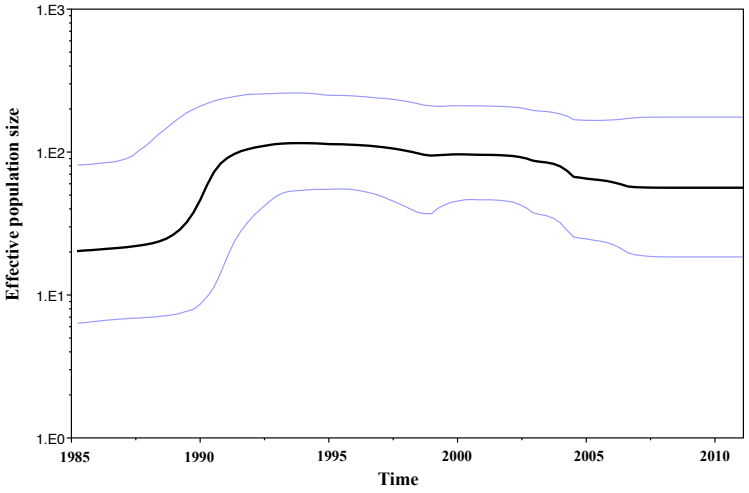


Demographic history of Danish MDR DT104

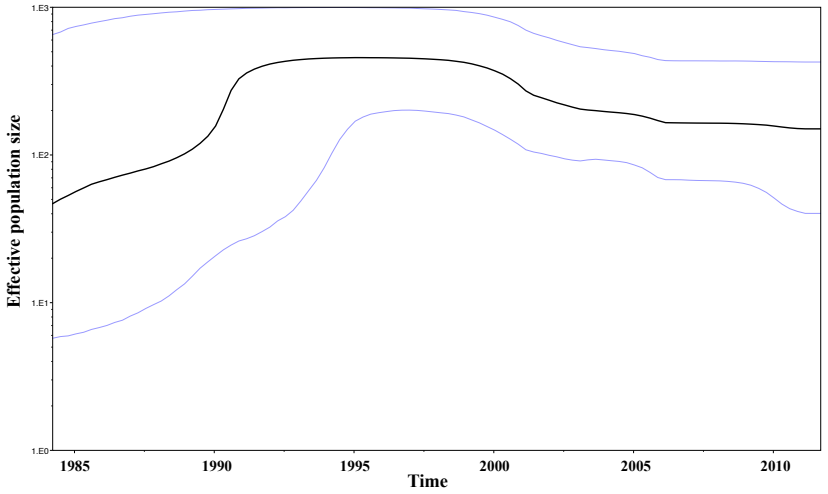


Leekitcharoenphon P. *et al*, Appl Environ Microbiol. 2016

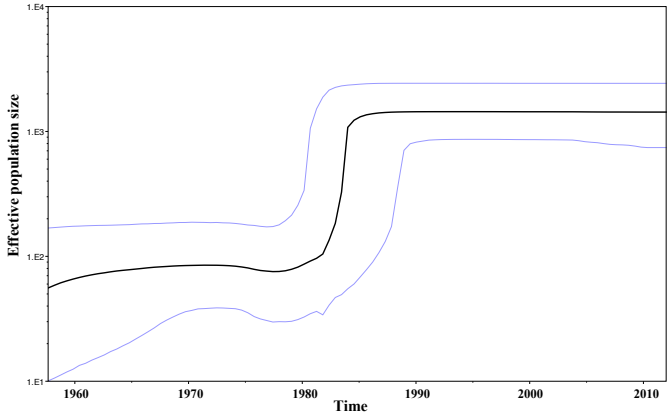
Cattle



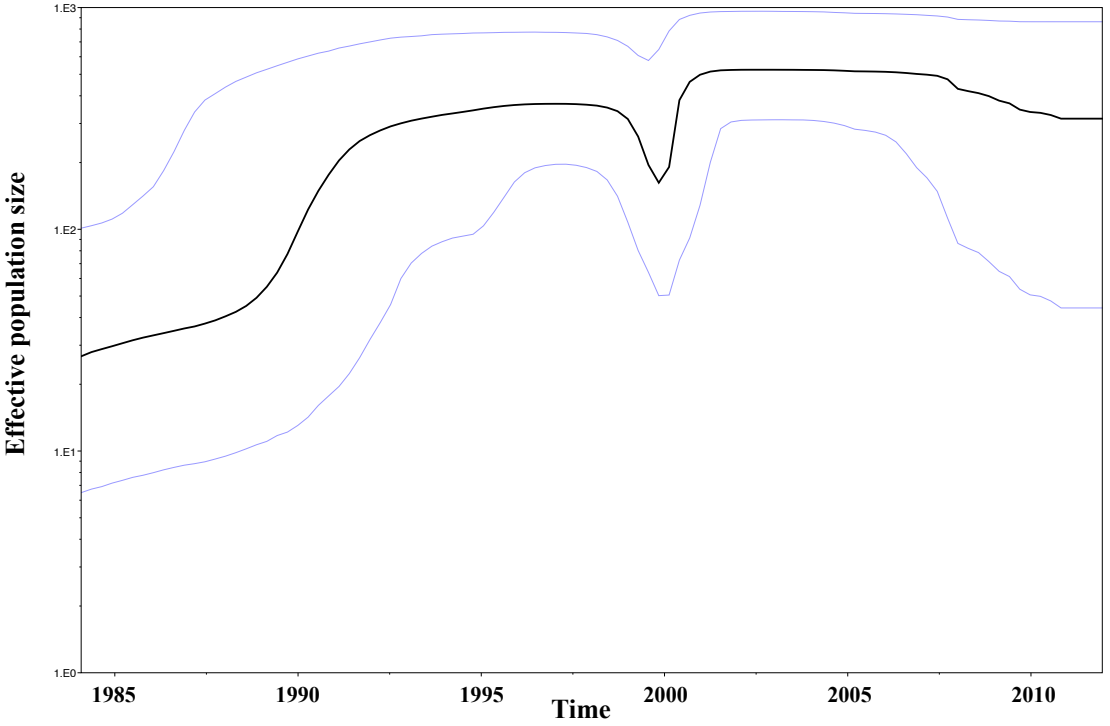
Poultry



Human



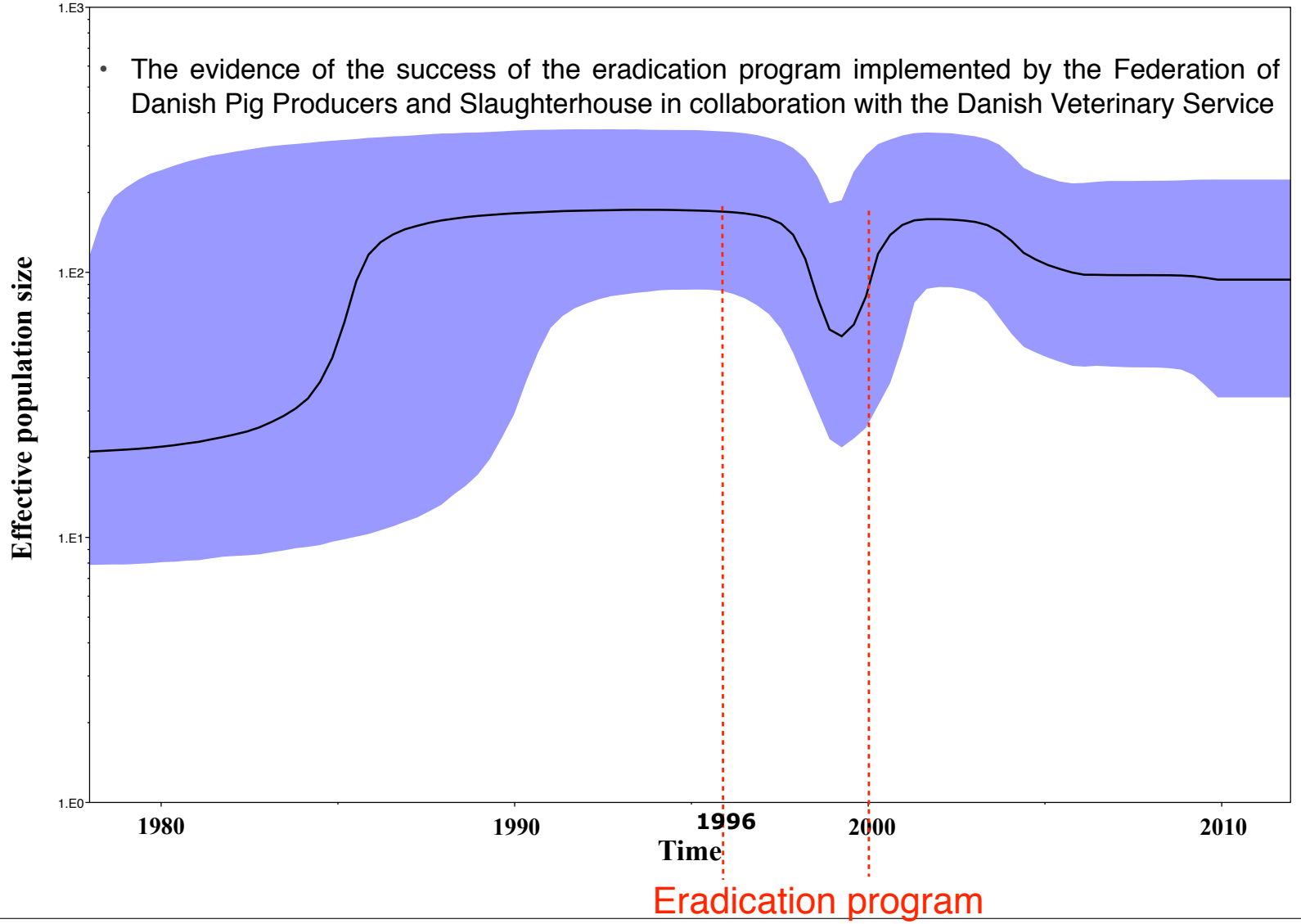
Swine

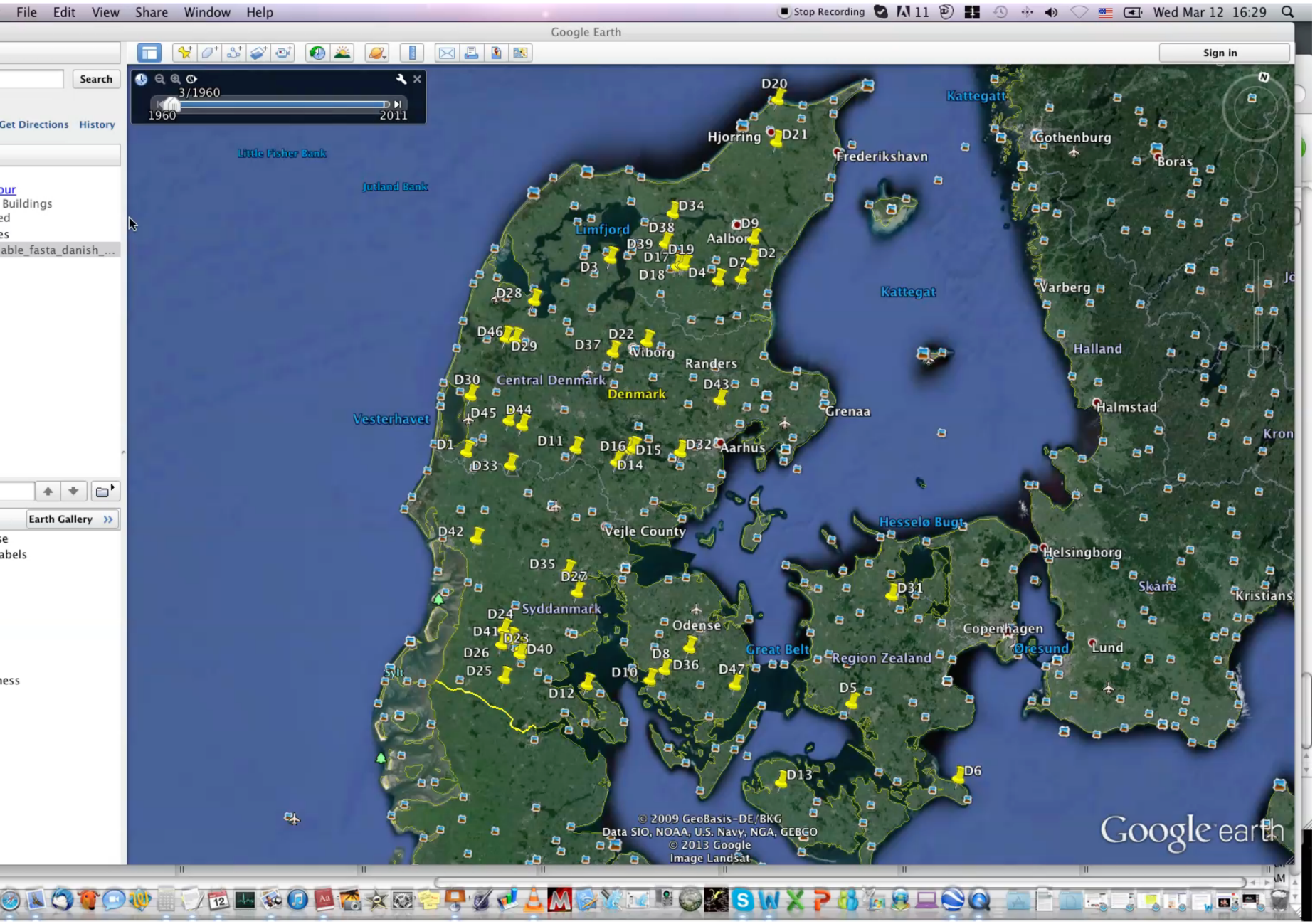


Demographic history of Danish MDR DT104

- The program aimed to eradicate MDR DT104 from infected pig herds by depopulation of pig herds, cleaning and disinfection of building before repopulation with pigs free from DT104

- The evidence of the success of the eradication program implemented by the Federation of Danish Pig Producers and Slaughterhouse in collaboration with the Danish Veterinary Service




File Edit View Share Window Help
Stop Recording
Wed Mar 12 16:29
Google Earth
Sign in
Search
Get Directions
History
Earth Gallery
3/1960
1960
2011
Little Fisher Bank
Jutland Bank
Hjorring
Frederikshavn
Kattegatt
Gothenburg
Boras
Limfjord
Aalborg
Kattegat
Varberg
Halland
Halmstad
Central Denmark
Denmark
Randers
Viborg
Grenaa
Vesterhavet
Aarhus
Vejle County
Hesselø Bugt
Helsingborg
Skane
Kristians
Syddanmark
Odense
Copenhagen
Oresund
Lund
Great Belt
Region Zealand
Sylt
D20
D21
D34
D38
D9
D39
D19
D17
D2
D3
D18
D4
D7
D28
D46
D29
D37
D22
D30
D43
D45
D44
D1
D11
D16
D15
D32
D33
D14
D42
D35
D27
D31
D24
D41
D23
D40
D26
D8
D36
D47
D25
D10
D12
D5
D13
D6
© 2009 GeoBasis-DE/BKG
Data SIO, NOAA, U.S. Navy, NGA, GEBCO
© 2013 Google
Image Landsat
Google earth