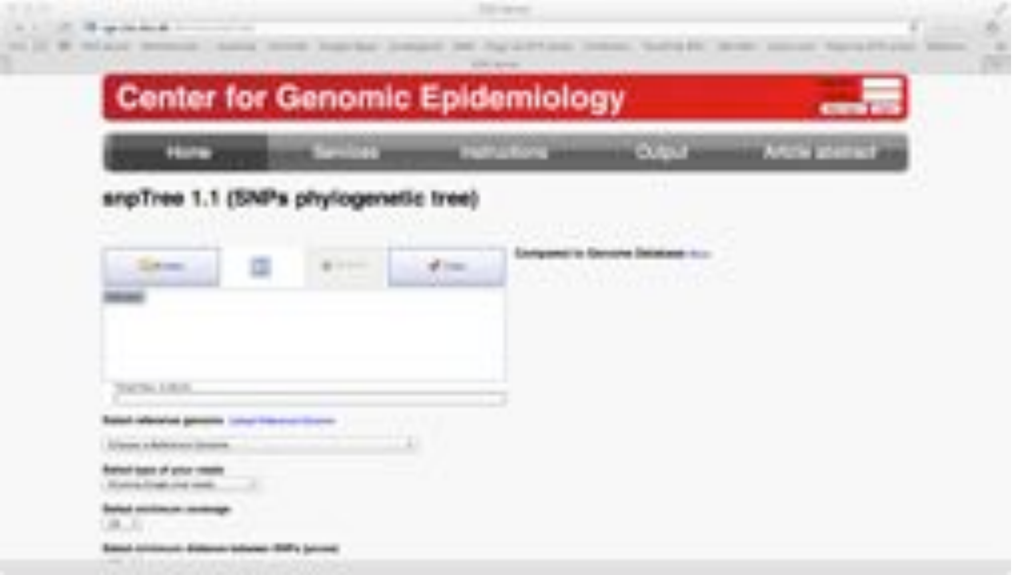
snpTree

- First online webserver for constructing phylogenetic trees based on whole genome sequencing



Center for Genomic Epidemiology

snpTree 1.1 (SNPs phylogenetic tree)

Compared to Genome Database

Select reference genome

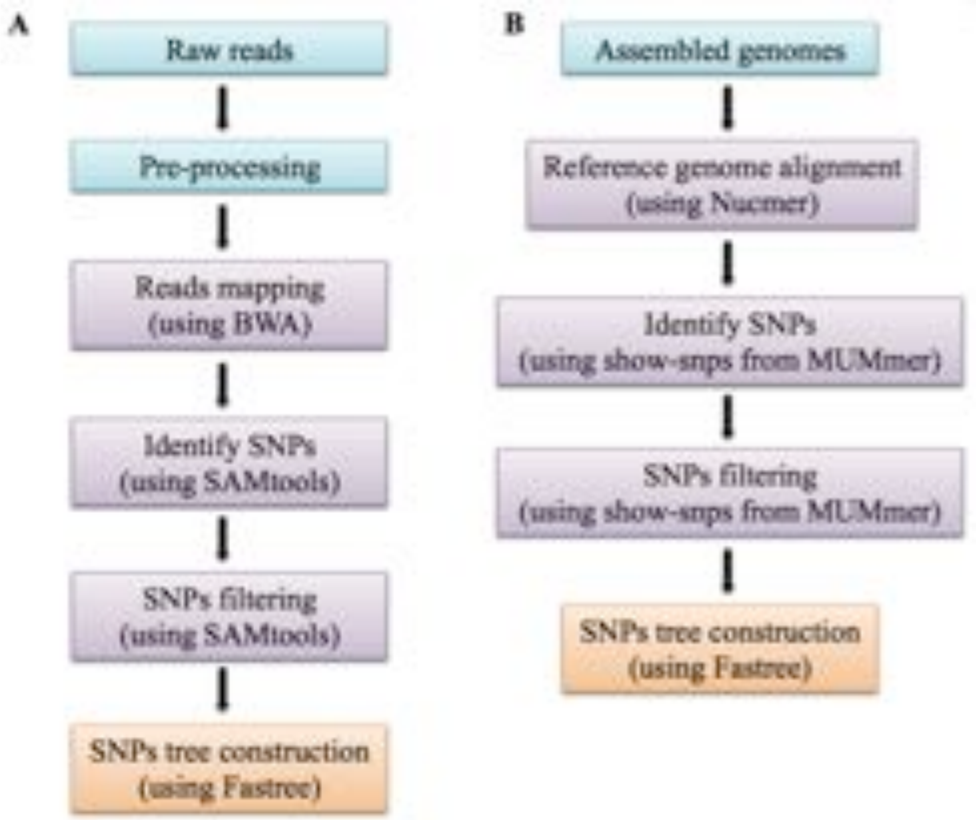
Select type of your reads

Select minimum coverage

Select minimum distance between SNPs (prune)

snpTree--a web-server to identify and construct SNP trees from whole genome sequence data. Leekitcharoenphon P, Kaas RS, Thomsen MC, Friis C, Rasmussen S, Aarestrup FM. BMC Genomics. 2012;13 Suppl 7:S6.

snpTree flow



A

Raw reads

Pre-processing

Reads mapping
(using BWA)

Identify SNPs
(using SAMtools)

SNPs filtering
(using SAMtools)

SNPs tree construction
(using Fastree)

B

Assembled genomes

Reference genome alignment
(using Nucmer)

Identify SNPs
(using show-snps from MUMmer)

SNPs filtering
(using show-snps from MUMmer)

SNPs tree construction
(using Fastree)

CSI Phylogeny

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

- SNP identification same as snpTree
- 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

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

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

CSI Phylogeny

Output

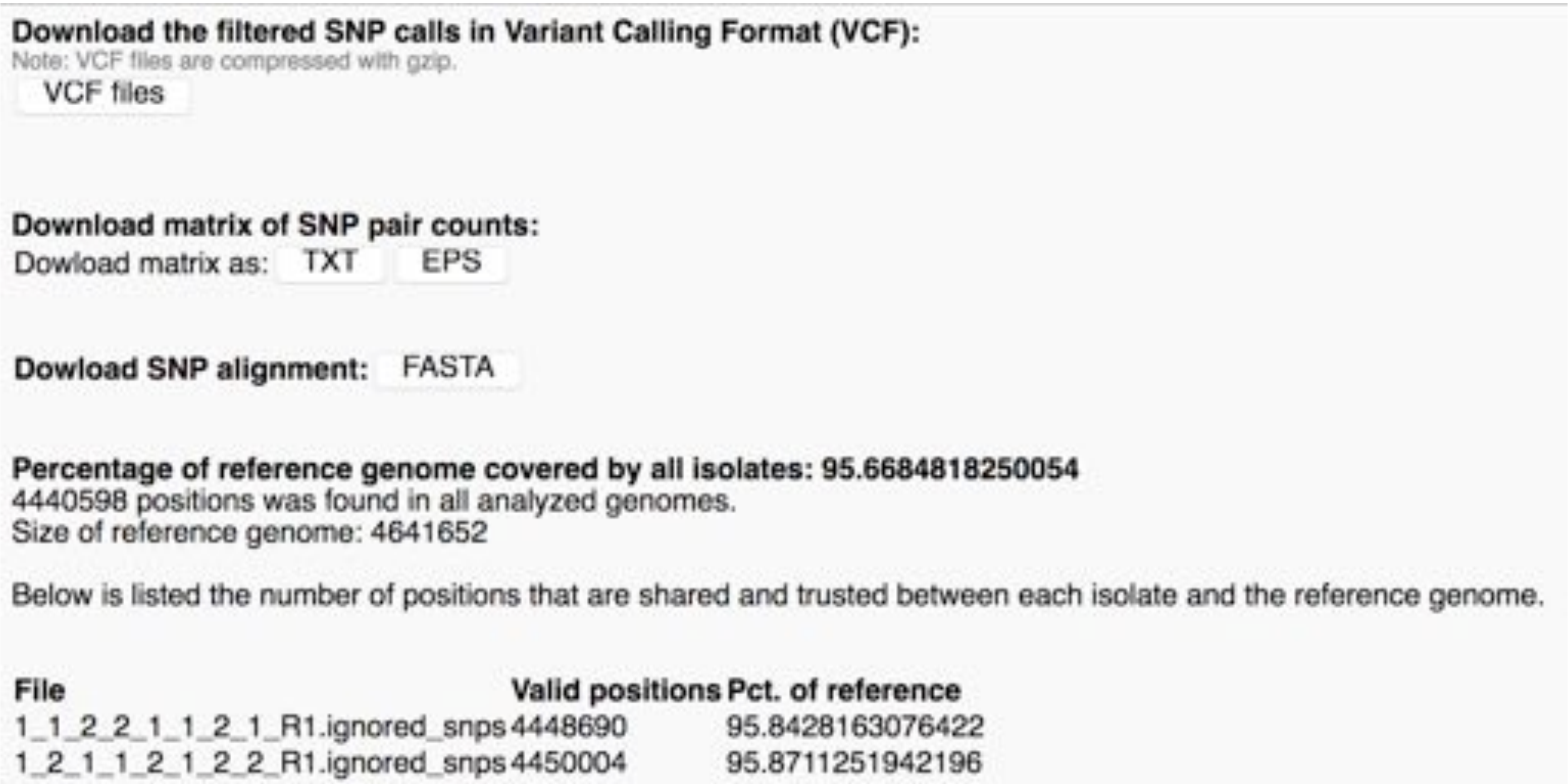
Tree build by FastTree algorithm, in Newick format

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

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

- Diagonal SNP matrix

CSI Phylogeny



Download the filtered SNP calls in Variant Calling Format (VCF):

Note: VCF files are compressed with gzip.

Download matrix of SNP pair counts:

Dowload matrix as:

Dowload SNP alignment:

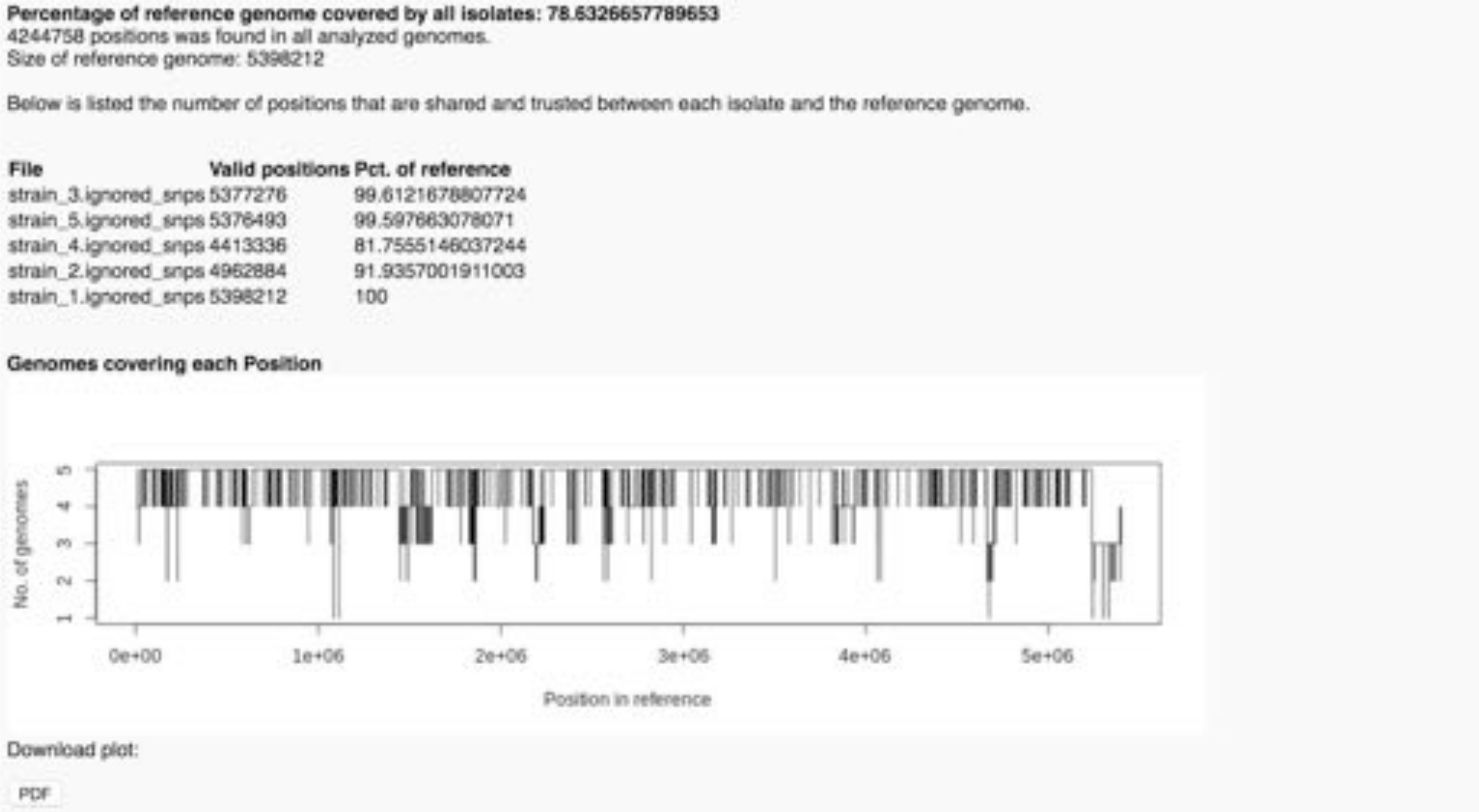
Percentage of reference genome covered by all isolates: 95.6684818250054

4440598 positions was found in all analyzed genomes.

Size of reference genome: 4641652

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
1_1_2_2_1_1_2_1_R1.ignored_snps	4448690	95.8428163076422
1_2_1_1_2_1_2_2_R1.ignored_snps	4450004	95.8711251942196



Percentage of reference genome covered by all isolates: 78.6326657789653
4244758 positions was found in all analyzed genomes.
Size of reference genome: 5398212

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
strain_3.ignored_snps	5377276	99.6121678807724
strain_5.ignored_snps	5376493	99.597663078071
strain_4.ignored_snps	4413336	81.7555146037244
strain_2.ignored_snps	4962884	91.9357001911003
strain_1.ignored_snps	5398212	100

Genomes covering each Position

Download plot:

NDtree

https://cge.cbs.dtu.dk/services/NDtree/

Nucleotide calling

- A different approach where the main distinction is not between if a SNP should be called or not, but between whether or not there is solid evidence for the nucleotide at the given position.

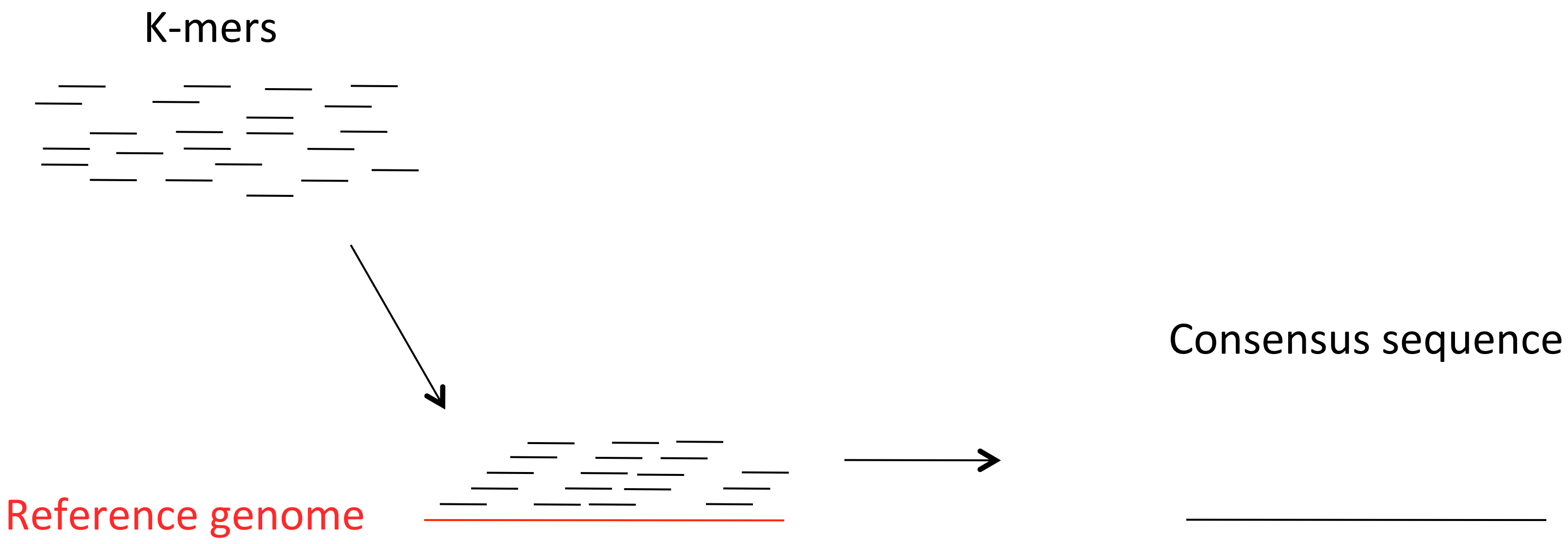
Real-Time Whole-Genome Sequencing for Routine Typing, Surveillance, and Outbreak Detection of Verotoxigenic Escherichia coli. Joensen KG, Scheutz F, Lund O, Hasman H, Kaas RS, Nielsen EM, Aarestrup FM. J Clin Microbiol. 2014 May;52(5):1501-10.

NDtree

Simple mapping approach

- Cuts all reads into K-mers
- Maps all K-mers to reference genome
- Makes an ungapped consensus sequences of equal lengths

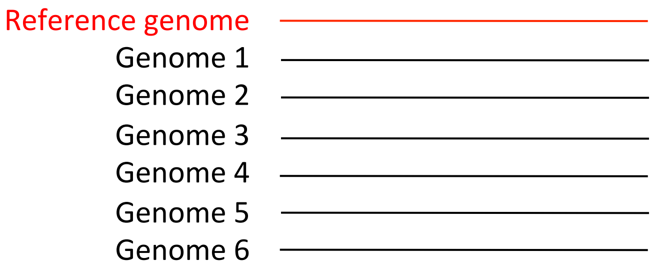
Mapping



K-mers

Consensus sequence

Reference genome



Reference genome

Genome 1

Genome 2

Genome 3

Genome 4

Genome 5

Genome 6

NDtree

Nucleotide calling

- When all reads have been mapped the significance of the base call at each position was evaluated by calculating the number of reads X having the most common nucleotide at that position, and the number of reads Y supporting other nucleotides.

A Z-score threshold is calculated

$$Z = \frac{X - Y}{\sqrt{\mathrm{X+Y}}} \quad > 1.96 \text{ (or } 3.29\text{)}$$

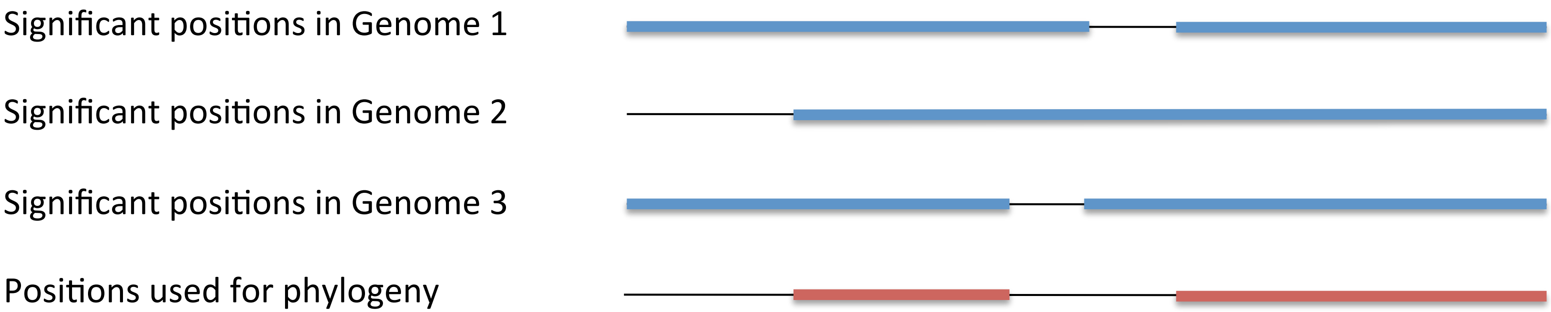
>90% of reads supporting the same base

NDtree

Count nucleotide differences

- **Method 1:** Each pair of sequences was compared and the number of nucleotide differences in positions called in **all** sequences was counted.
  - More accurate (Z=1.96 is used as threshold)
- **Method 2:** Each pair of sequences was compared and the number of nucleotide differences in positions called in **both** sequences was counted.
  - More robust (Z=3.29 is used as threshold)

Method 1 – all called



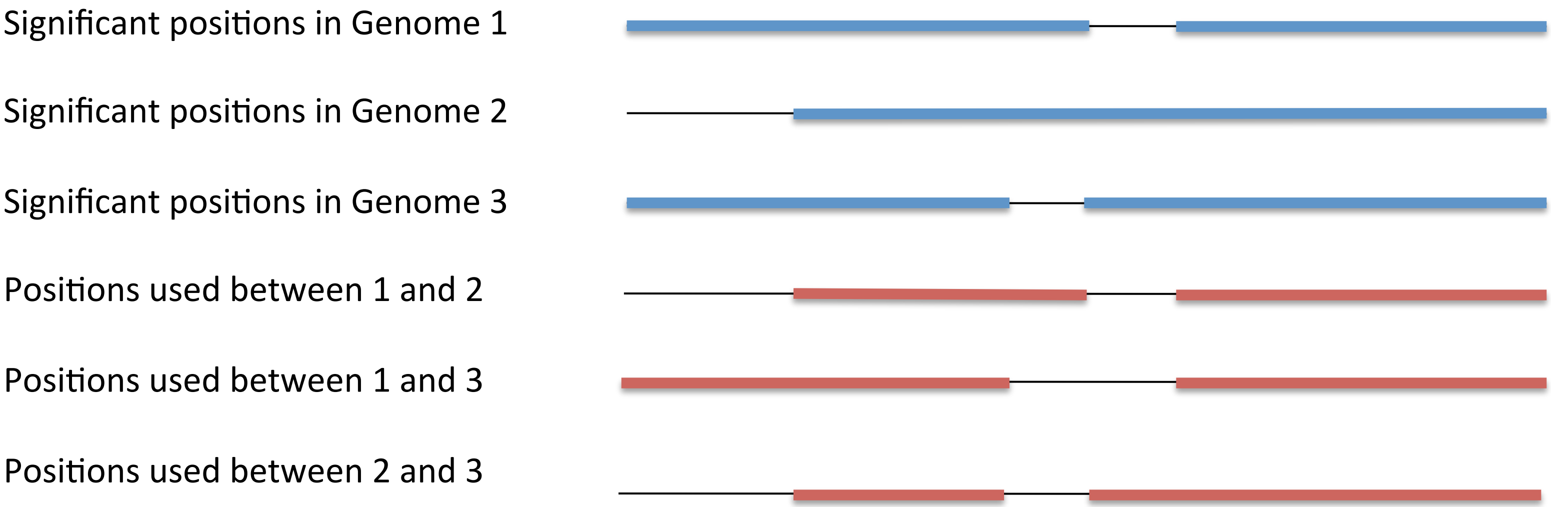
Significant positions in Genome 1

Significant positions in Genome 2

Significant positions in Genome 3

Positions used for phylogeny

Method 2 – pairwise significance



Significant positions in Genome 1

Significant positions in Genome 2

Significant positions in Genome 3

Positions used between 1 and 2

Positions used between 1 and 3

Positions used between 2 and 3

NDtree

Uses two different algorithms to make two different trees

- UPGMA
- Neighbor Joining

Both algorithms are part of the PHYLIP Neighbor program package and make trees from distance matrices

UPGMA vs. Neighbor Joining

- UPGMA works when samples have been taken the same time

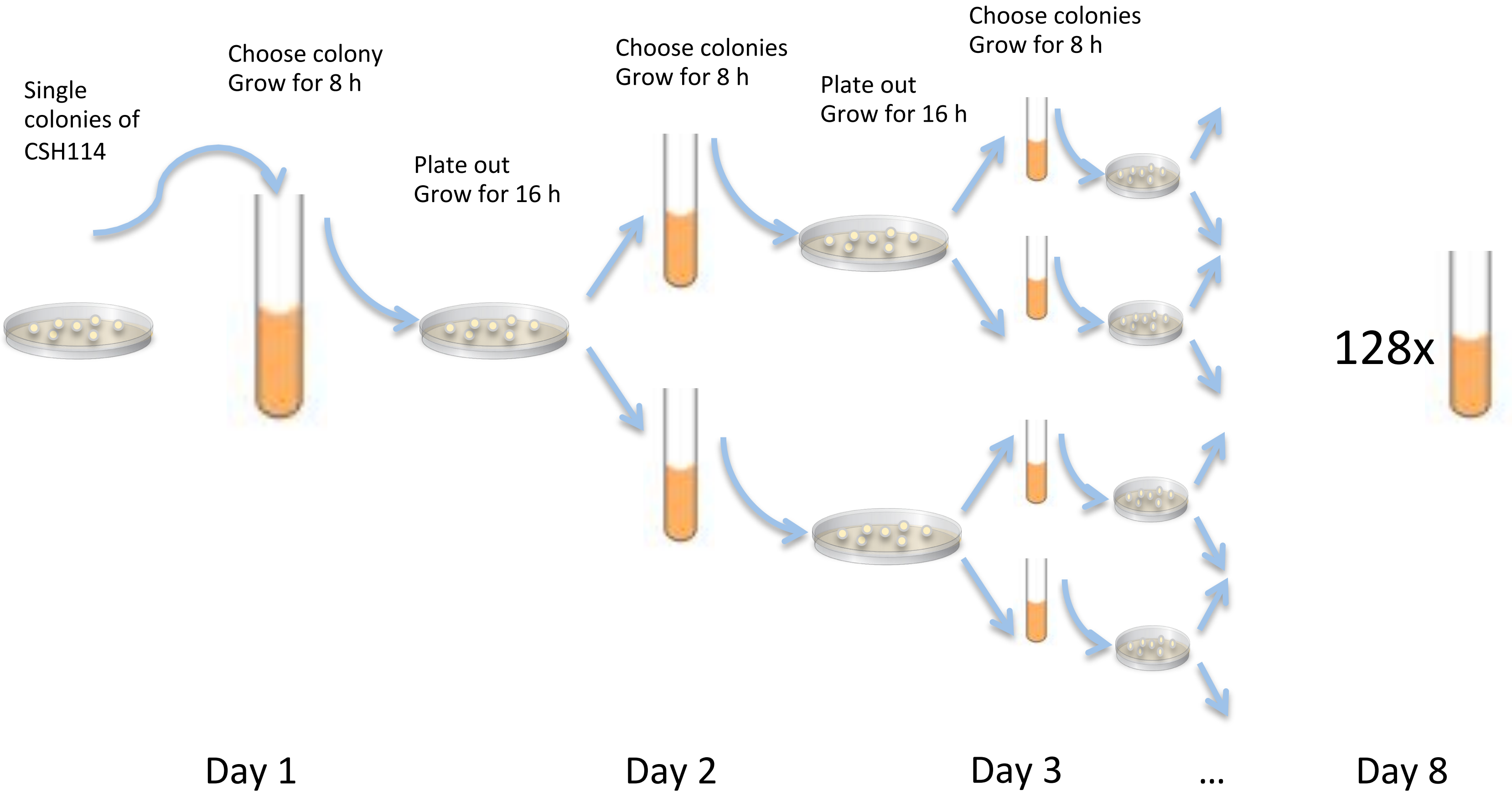
- Neighbor Joining is better when samples have been taken at different times

NDtree

Output

- **distance.txt**: Distance matrix - tab separated
- **dist.mat**: Distance matrix - PHYLIP format
- **tree.nj.newick**: Neighbor Joining tree - Newick format
  - Branch lengths is number of Nucleotide Differences
- **tree.upgma.newick**: UPGMA tree – Newick format
  - Branch lengths is number of Nucleotide Differences

Controlled Evolution study



For each 8 hour culture a sample was saved for DNA sequencing

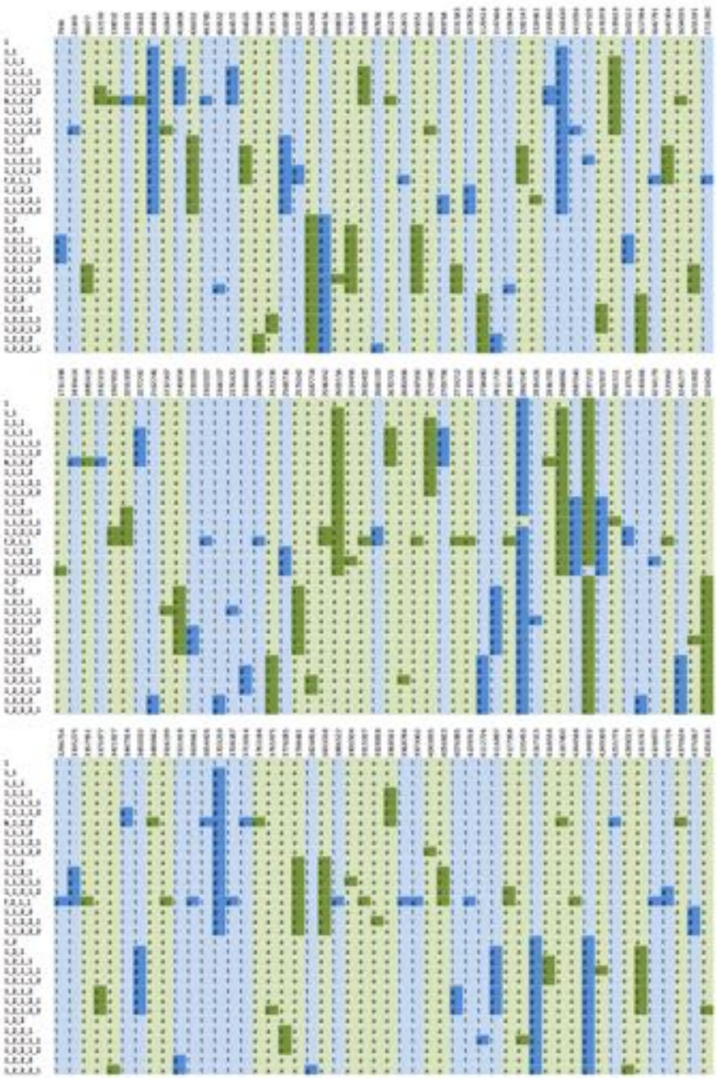
DTU Bioinformatics
Department of Bio and Health Informatics

J. Ahrenfeldt, C. Skaarup, H. Hasman, A. G. Pedersen, F. M. Aarestrup and O. Lund. Bacterial whole genome-based phylogeny: construction of a new benchmarking dataset and assessment of some existing methods. BMC Genomics (2017) 18:19

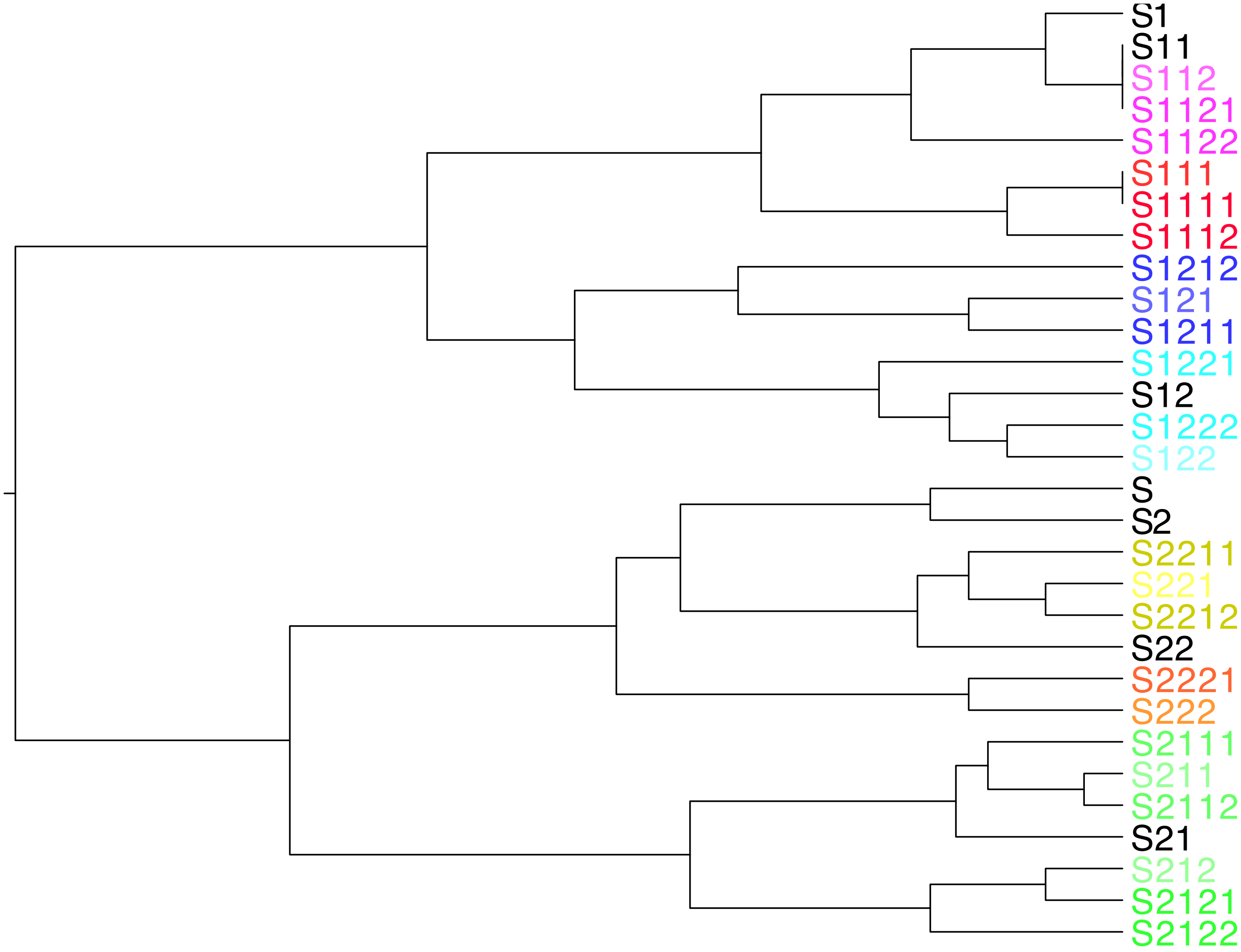
Naming the descendants

Day 1	Day 2	Day 3	Day 4	Day 5
S	**S1**	**S11**	**S111**	**S1111**
				S1112
			S112	**S1121**
				S1122
		S12	**S121**	**S1211**
				S1212
			S122	**S1221**
				S1222
	S2	**S21**	**S211**	**S2111**
				S2112
			S212	**S2121**
				S2122
		S22	**S221**	**S2211**
				S2212
			S222	**S2221**
				S2222

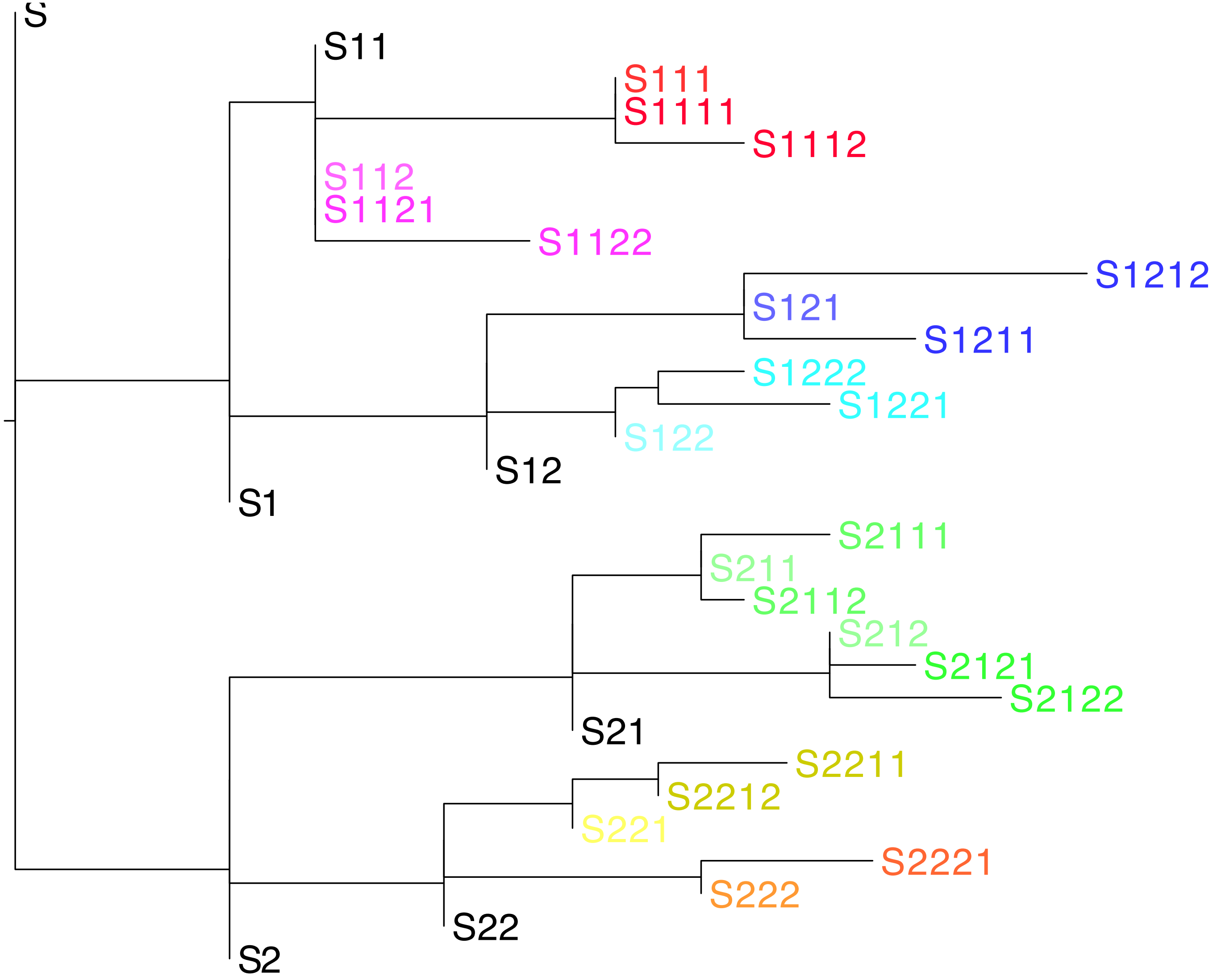
Mutations



Phylogenetic tree using NDtree (UPGMA)



Phylogenetic tree using NDtree (Neighbor Joining)

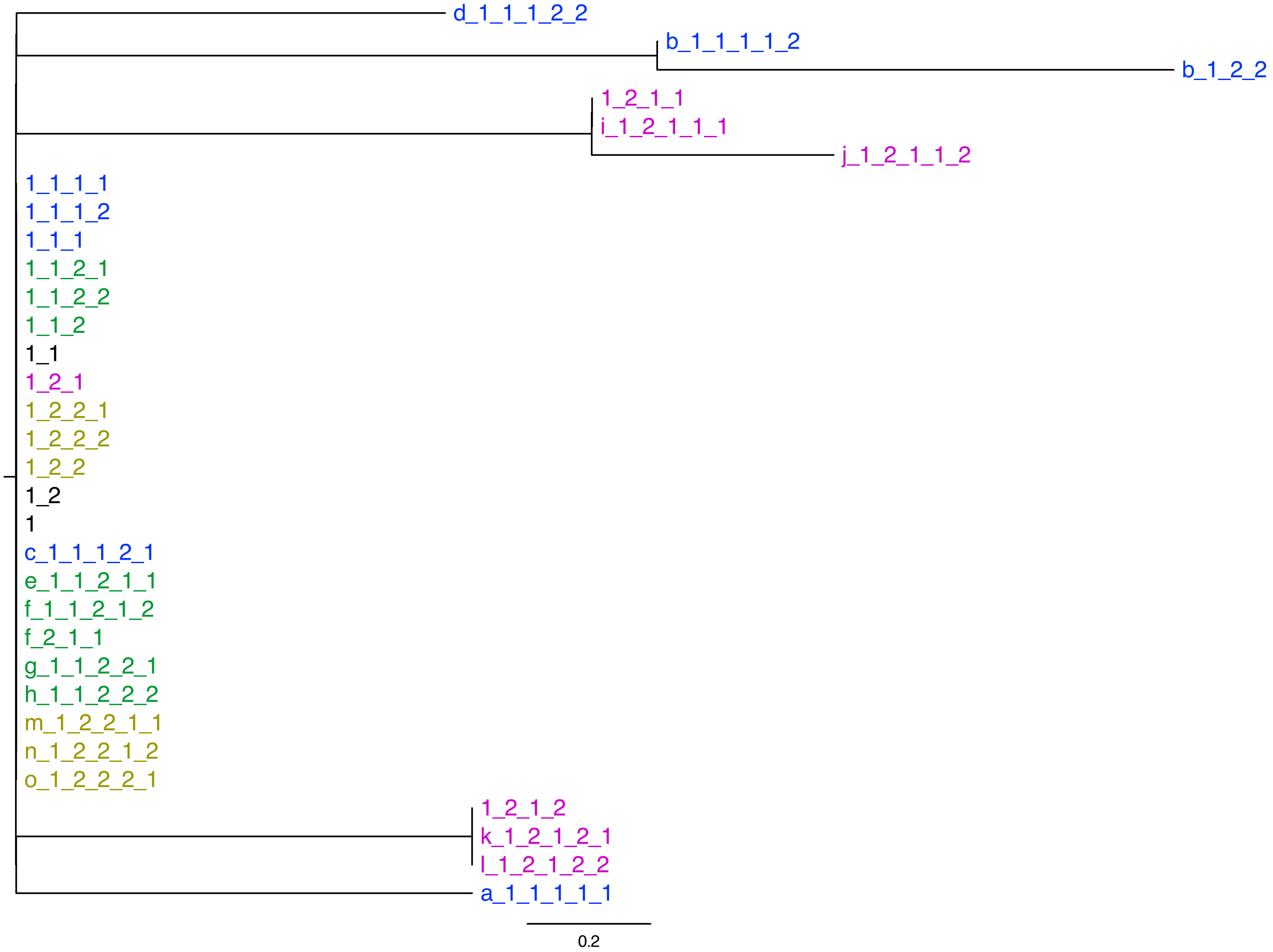


UPGMA vs. Neighbor Joining

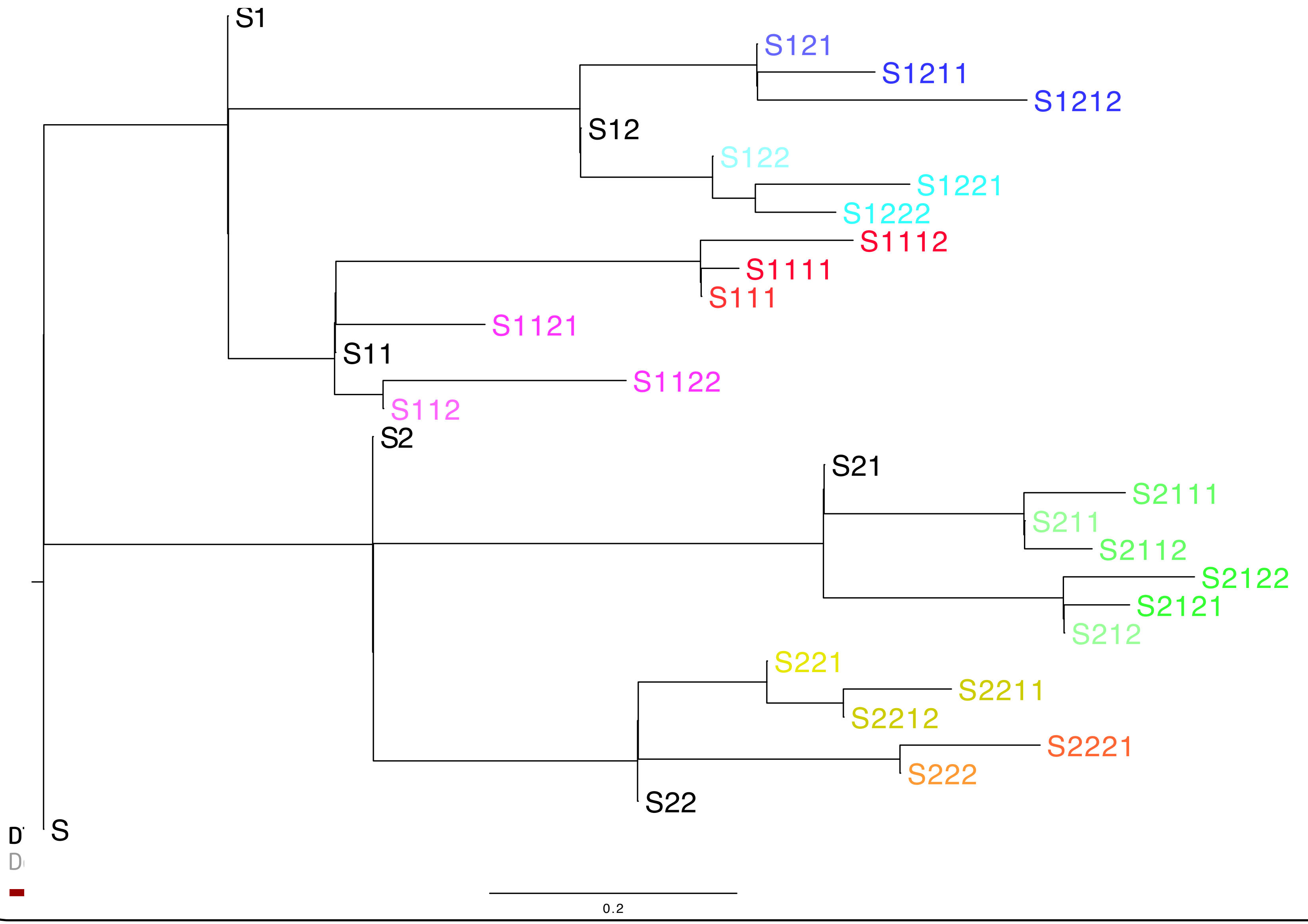
- UPGMA works when samples have been taken the same time

- Neighbor Joining is better when samples have been taken at different times

CSI Phylogeny – Default settings



CSI Phylogeny – Pruning disabled



So... What should I use when?

CSI Phylogeny

- Has very good statistics and a good graphical overview.
- Advantageous to use when you expect the differences between the isolates to be larger than 5-10 mutations.
- Is faster

NDtree

- Is able to find very small differences.
- Does not take recombination into consideration.
- Works best on raw reads. If given assembled genomes, it simulates reads.

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.

What defines an outbreak

- We can’t tell for certain
- It depends on the species
- But a rule of thump is:
  - Within 10 SNPs it is definitely an outbreak
  - Within 30 SNPs it might be an outbreak
  - Above 60 SNPs it is most likely not an outbreak

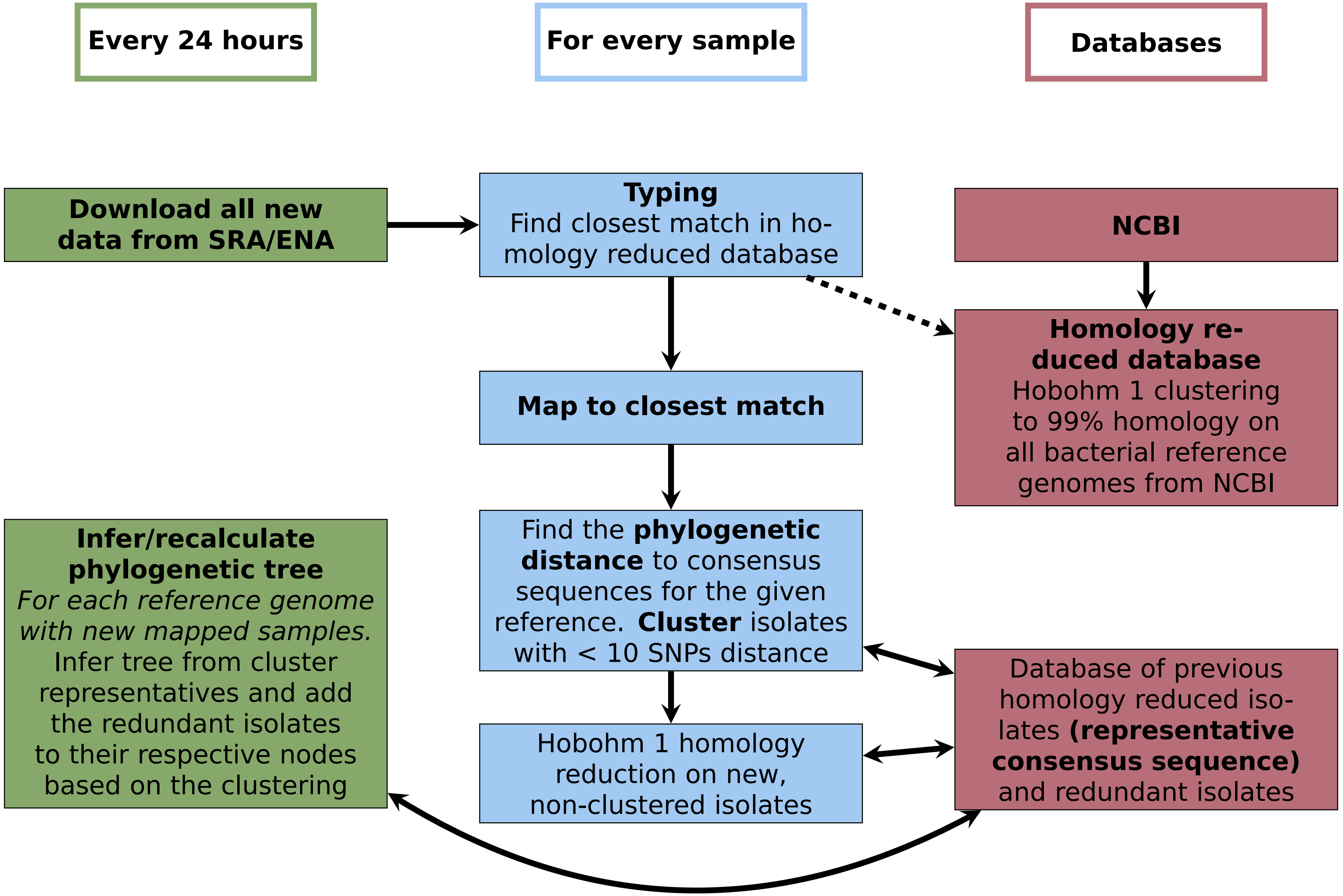
And now a little advertisement for a cool project we are working on

Evergreen

- SNP trees continuously updated with all new SRA/ENA entries for selected species daily
- Pilot
  - Coli, Campy, Shigella, Salmonella, and Listeria
  - 2017
- Species, “from data” can be user selected

Judit Szarvas, Johanne Ahrenfeldt

Evergreen - flowchart



Every 24 hours

For every sample

Databases

Download all new data from SRA/ENA

Typing
Find closest match in homology reduced database

NCBI

Homology reduced database
Hobohm 1 clustering to 99% homology on all bacterial reference genomes from NCBI

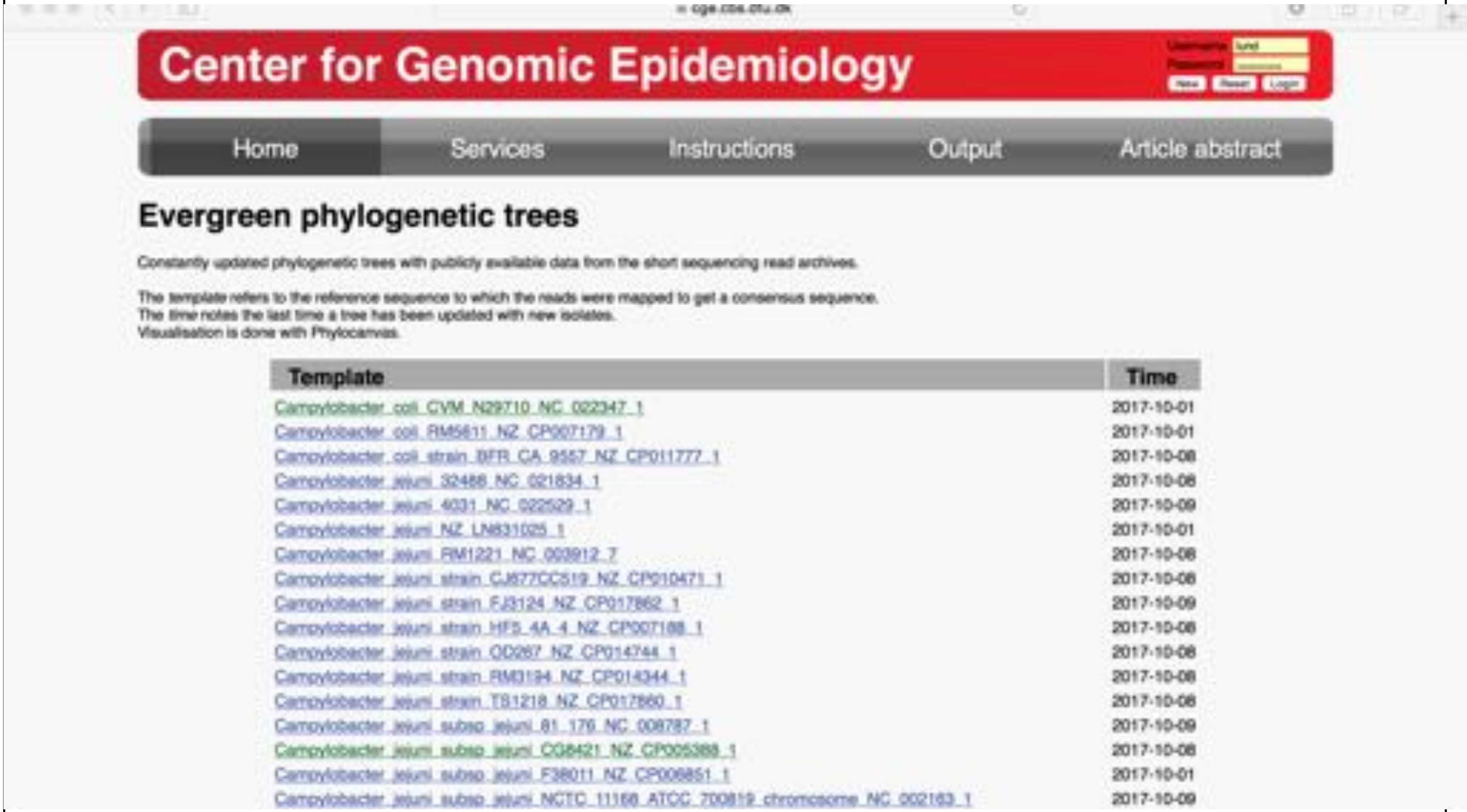
Map to closest match

Find the **phylogenetic distance** to consensus sequences for the given reference. **Cluster** isolates with < 10 SNPs distance

Infer/recalculate phylogenetic tree
For each reference genome with new mapped samples.
Infer tree from cluster representatives and add the redundant isolates to their respective nodes based on the clustering

Hobohm 1 homology reduction on new, non-clustered isolates

Database of previous homology reduced isolates **(representative consensus sequence)** and redundant isolates

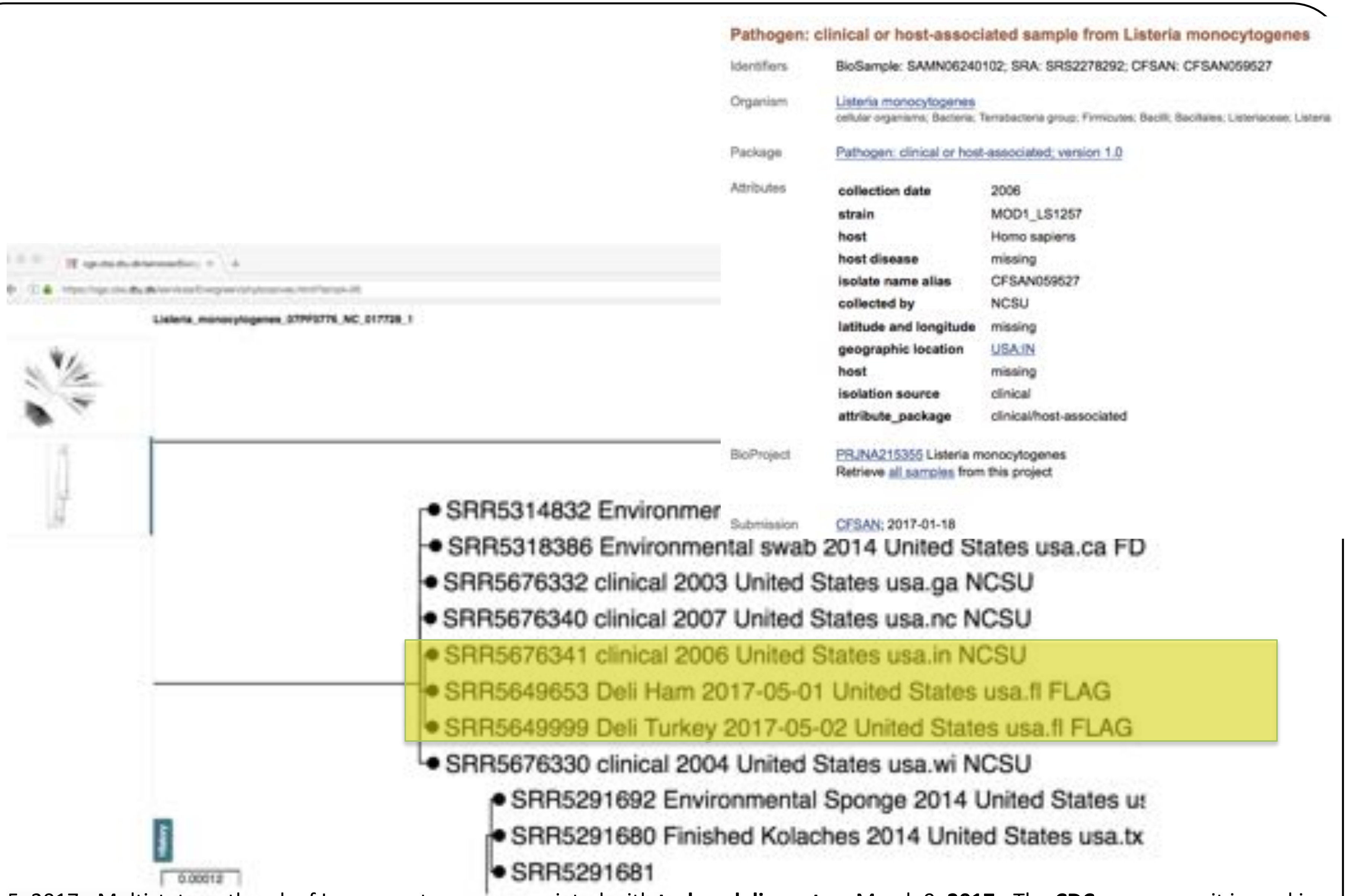


Evergreen phylogenetic trees

Constantly updated phylogenetic trees with publicly available data from the short sequencing read archives.

The template refers to the reference sequence to which the reads were mapped to get a consensus sequence.
The time notes the last time a tree has been updated with new isolates.
Visualisation is done with Phylocanvas.

Template	Time
Campylobacter_coli_CVM_N29710_NC_022347_1	2017-10-01
Campylobacter_coli_RM5611_NZ_CP007179_1	2017-10-01
Campylobacter_coli_strain_BFR_CA_9557_NZ_CP011777_1	2017-10-08
Campylobacter_jejuni_32488_NC_021834_1	2017-10-08
Campylobacter_jejuni_4031_NC_022529_1	2017-10-09
Campylobacter_jejuni_NZ_LN831025_1	2017-10-01
Campylobacter_jejuni_RM1221_NC_003912_7	2017-10-08
Campylobacter_jejuni_strain_CJ677CC519_NZ_CP010471_1	2017-10-08
Campylobacter_jejuni_strain_FJ3124_NZ_CP017862_1	2017-10-09
Campylobacter_jejuni_strain_HF5_4A_4_NZ_CP007188_1	2017-10-08
Campylobacter_jejuni_strain_OD267_NZ_CP014744_1	2017-10-08
Campylobacter_jejuni_strain_RM3194_NZ_CP014344_1	2017-10-08
Campylobacter_jejuni_strain_TS1218_NZ_CP017860_1	2017-10-08
Campylobacter_jejuni_subsp_jejuni_81_176_NC_008787_1	2017-10-09
Campylobacter_jejuni_subsp_jejuni_CG8421_NZ_CP005388_1	2017-10-08
Campylobacter_jejuni_subsp_jejuni_F38011_NZ_CP006851_1	2017-10-01
Campylobacter_jejuni_subsp_jejuni_NCTC_11168_ATCC_700819_chromosome_NC_002163_1	2017-10-09



Pathogen: clinical or host-associated sample from Listeria monocytogenes

Identifiers: BioSample: SAMN06240102; SRA: SRS2278292; CFSAN: CFSAN059527

Organism: Listeria monocytogenes
cellular organisms; Bacteria; Terrabacteria group; Firmicutes; Bacilli; Bacillales; Listeriaceae; Listeria

Package: Pathogen: clinical or host-associated; version 1.0

Attributes:

collection date	2006
strain	MOD1_LS1257
host	Homo sapiens
host disease	missing
isolate name alias	CFSAN059527
collected by	NCSU
latitude and longitude	missing
geographic location	USA:IN
host	missing
isolation source	clinical
attribute_package	clinical/host-associated

BioProject: PRJNA215355 Listeria monocytogenes
Retrieve all samples from this project

Submission: CFSAN; 2017-01-18

Listeria_monocytogenes_07PF0776_NC_017728_1

SRR5314832 Environmer
SRR5318386 Environmental swab 2014 United States usa.ca FD
SRR5676332 clinical 2003 United States usa.ga NCSU
SRR5676340 clinical 2007 United States usa.nc NCSU
SRR5676341 clinical 2006 United States usa.in NCSU
SRR5649653 Deli Ham 2017-05-01 United States usa.fl FLAG
SRR5649999 Deli Turkey 2017-05-02 United States usa.fl FLAG
SRR5676330 clinical 2004 United States usa.wi NCSU
SRR5291692 Environmental Sponge 2014 United States us
SRR5291680 Finished Kolaches 2014 United States usa.tx
SRR5291681

Jun 5, 2017 - Multistate outbreak of L. monocytogenes associated with **turkey deli meat**. ... March 9, **2017** - The **CDC** announces it is working with the FDA to ... were reported in four states - Connecticut, **Florida**, New York, and Vermont.

Thank you for listening

- Questions?