Overview of Services

Phenotyping

- ResFinder
- PathogenFinder
- VirulenceFinder
- Restriction-ModificationFinder

Typing

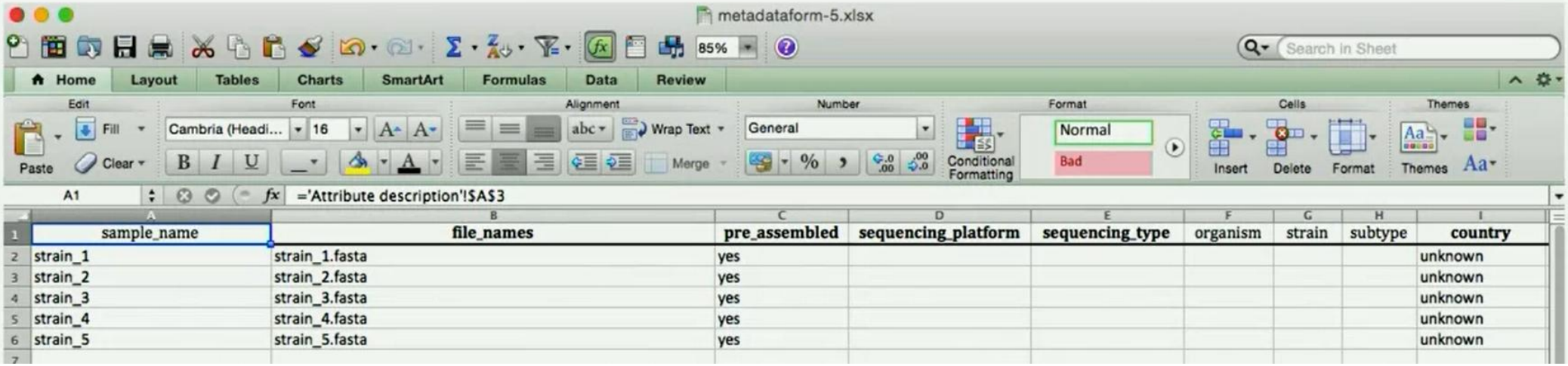
- MLST
- pMLST *
- PlasmidFinder *
- KmerFinder
- SpeciesFinder *
- Reads2Type
- Tapir (OBS! Only works for Firefox)
- TaxonomyFinder *
- SeroTypeFinder

Phylogeny

- snpTree
- CSI Phylogeny
- NDtree

Other

- PanFunPro
- MyDbFinder



	A	B	C	D	E	F	G	H	I
1	sample_name	file_names	pre_assembled	sequencing_platform	sequencing_type	organism	strain	subtype	country
2	strain_1	strain_1.fasta	yes						unknown
3	strain_2	strain_2.fasta	yes						unknown
4	strain_3	strain_3.fasta	yes						unknown
5	strain_4	strain_4.fasta	yes						unknown
6	strain_5	strain_5.fasta	yes						unknown
7									

Phylogeny inferred from SNPs

DTU

Assumption: Random + Independent

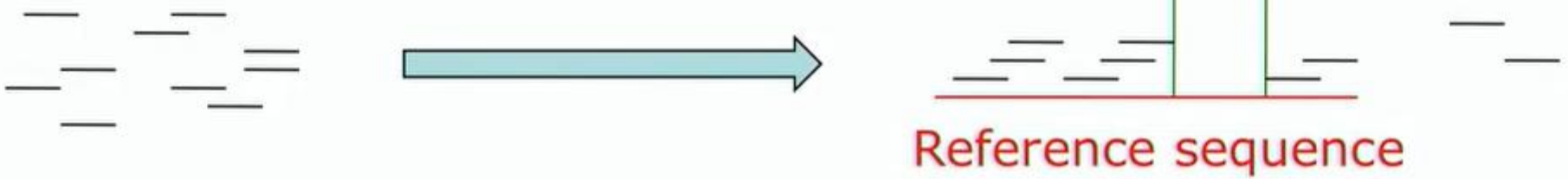
1. Find differences (SNP calling), compared to a reference sequence.
   - Close reference is better
2. Make pseudo-alignment (independent assumption)
3. Infer phylogeny

Phylogeny inferred from SNPs

CSI Phylogeny

Mappers:
BWA
Bowtie
MAQ

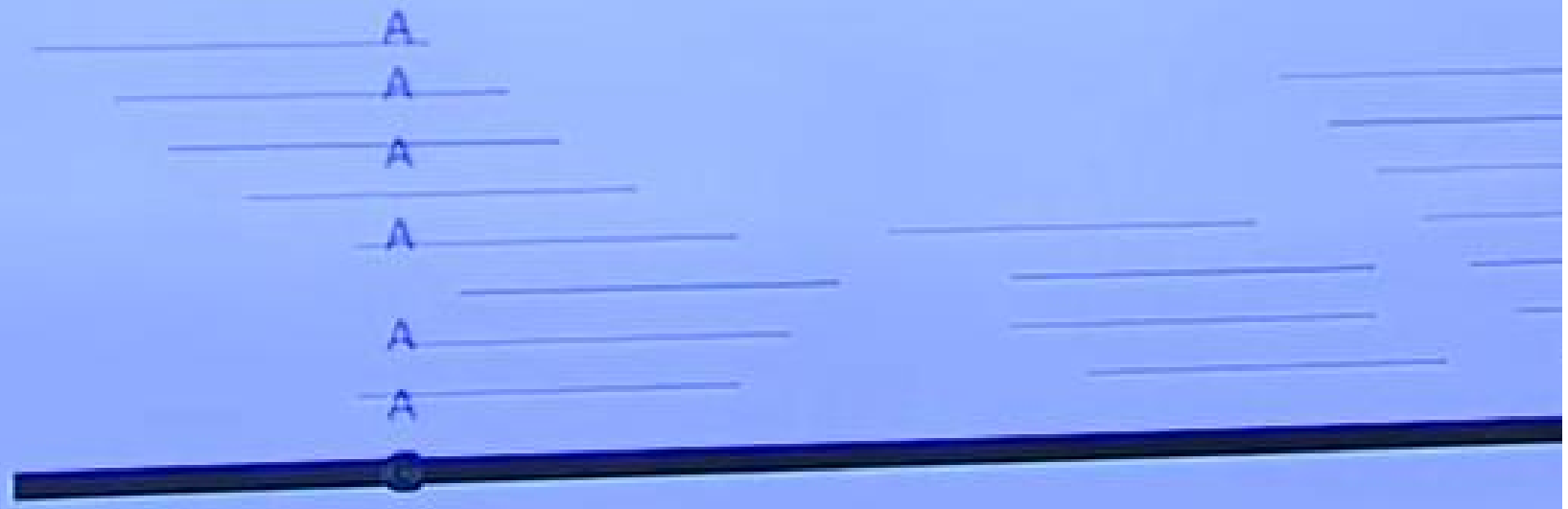
raw reads



Phylogeny inferred from SNPs

SNP filtering:

- Repeat sequence
- Mobile elements

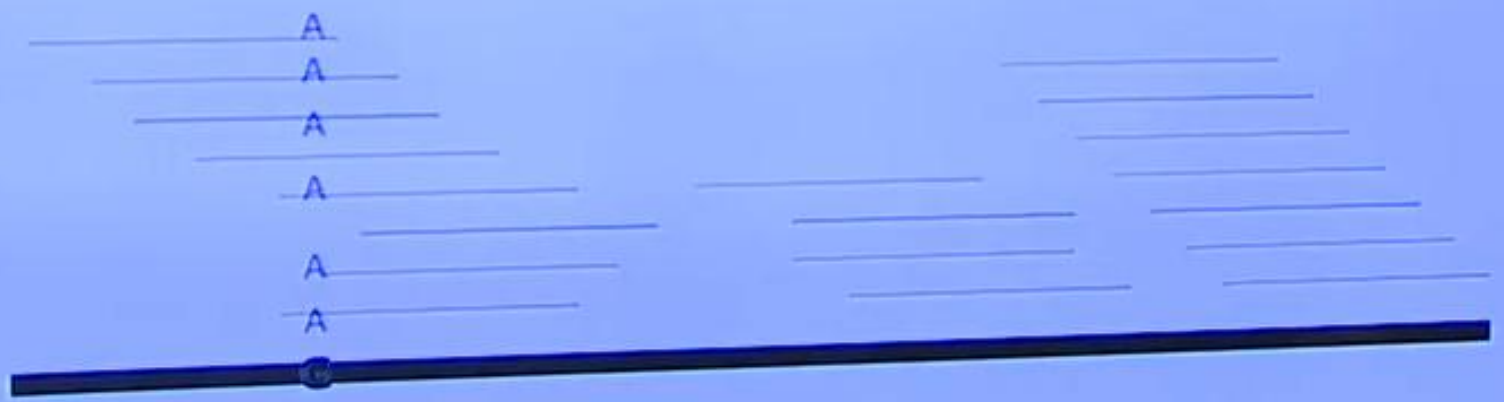


Phylogeny inferred from SNPs

DTU

SNP filtering:
- Repeat sequence
- Mobile elements

What is the purpose?
- Find mutations (Less conservative filtering)
- Study relations (Conservative filtering)



then you want to use a more

Phylogeny inferred from SNPs

1. Call SNPs for each isolate, using the same reference
2. Concatenate SNPs into "SNP sequences", one per isolate

Example:
Isolate 1: ACGTTTACC
Isolate 2: ACCTTTGCC
Isolate 3: ACCTTAGCC

3. Create tree using a chosen algorithm.

CSI Phylogeny method/pipeline

Raw reads

1. Map reads to reference (BWA)
2. Call all possible SNPs (Samtools)
3. Filter positions and SNPs using: coverage, quality, and z-score
   - Z-score = (X-Y)/sqrt(X+Y)
4. Prune SNPs.

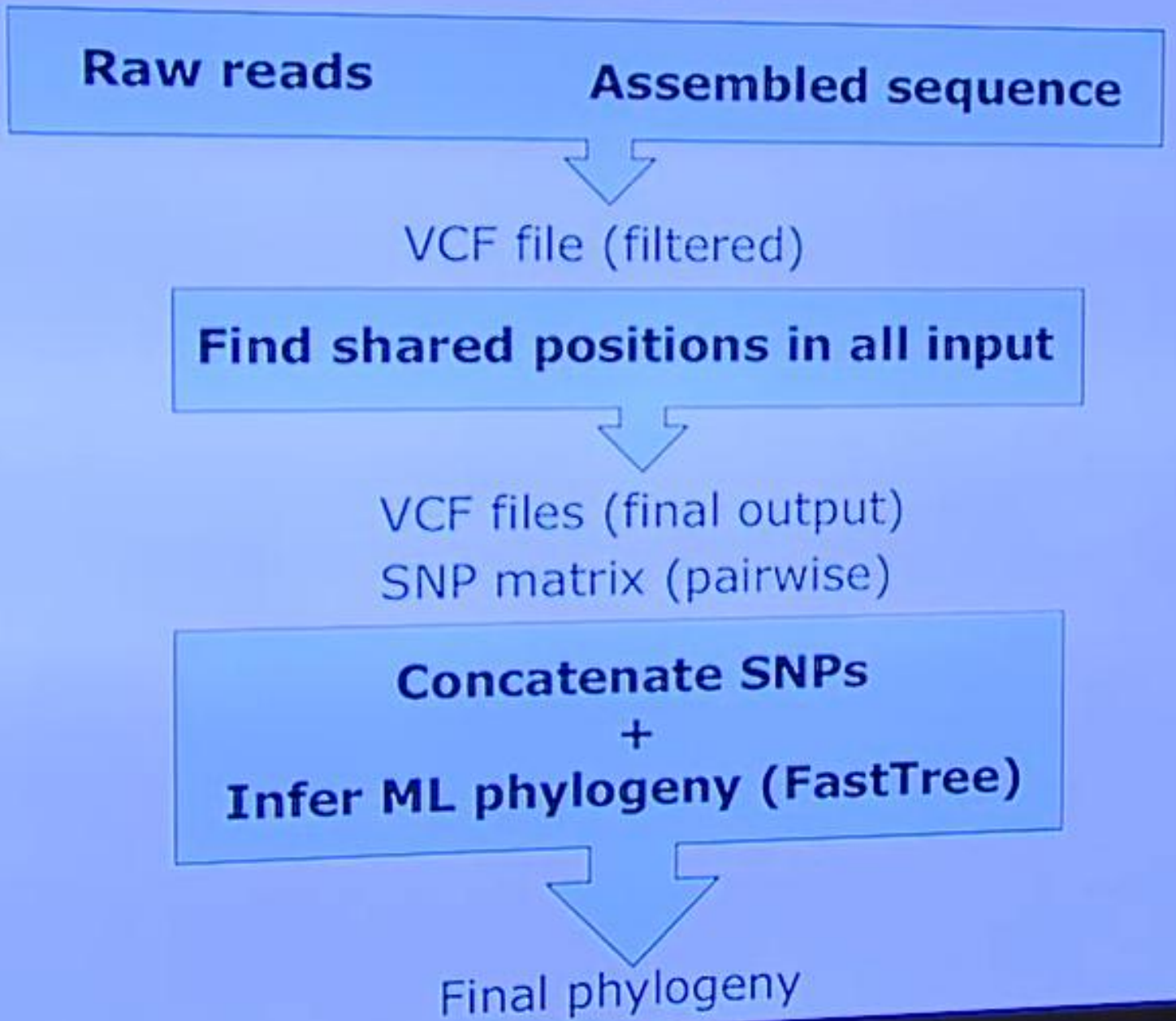
Output: VCF file (filtered)

Assembled sequence

1. NUCMER (Part of MUMMER)
   - Only pruning

Output: VCF file (filtered)

CSI Phylogeny method/pipeline



Output results

- Phylogeny (Newick, PDF, png)
  - Use FigTree or similar to open Newick files
- SNP Matrix (txt, pdf)
  - txt file is tab seperated text file.
- Pseudo-alignment (fasta)
  - Create tree using other algorithms
- Quality Control
  - Percentage of reference covered
  - Plot of ignored SNPs
  - Plot of heterozygous SNPs

Genomic epidemiology

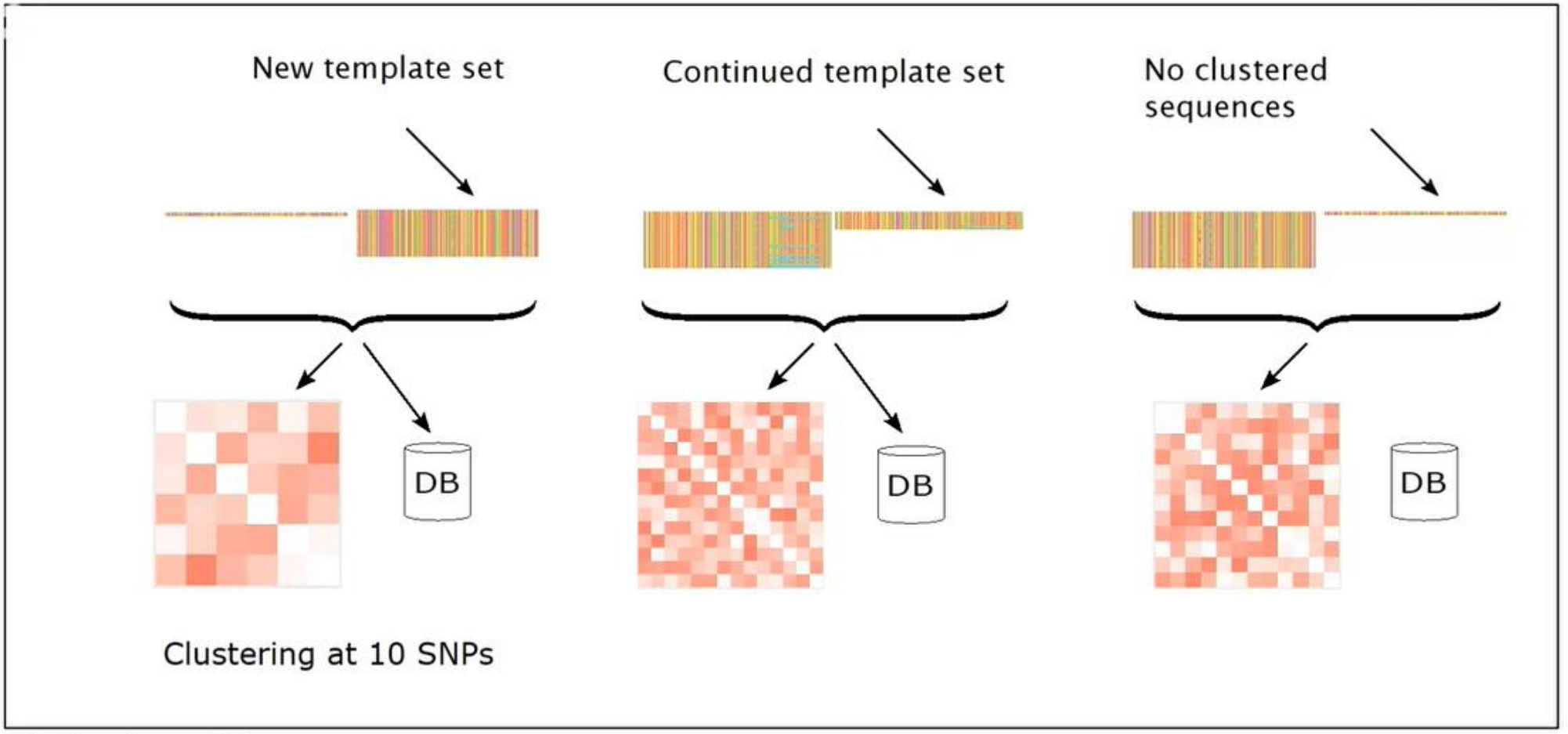
- Use of genomic data in epidemiological investigations to track infectious diseases.
- Samples, that differ only in a few single nucleotide polymorphisms (SNPs), could be related and have a common source.
- Phylogenetic trees are useful for visualizing the relationships between samples, and clustering related samples.
- Evergreen Online aims to discover foodborne disease outbreaks and connect clinical samples to sources of the infection.

Which of the following sentences is TRUE?

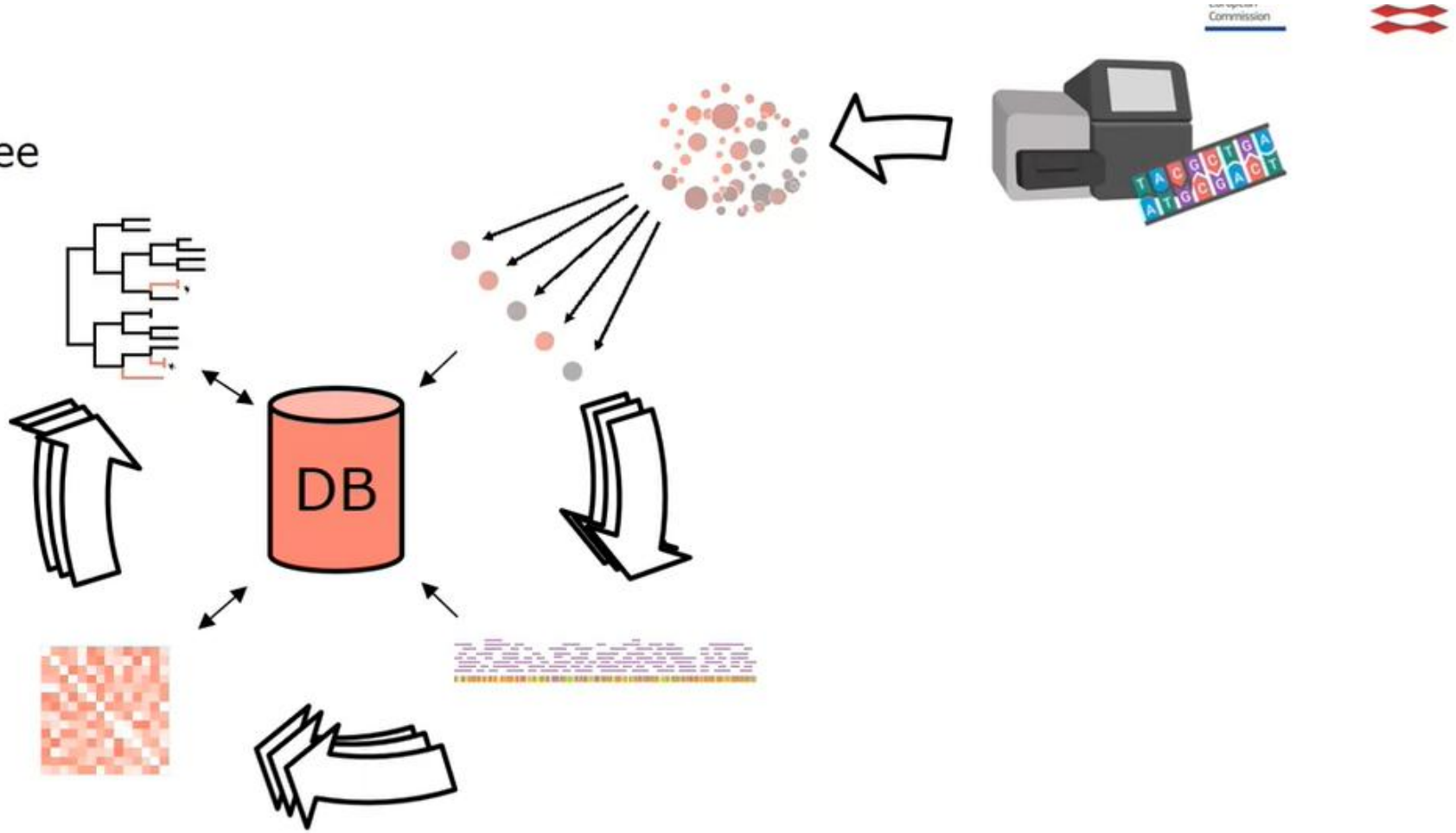
- ○ Genomic epidemiology is the genetic analysis of the host of the infectious disease.
- ○ Phylogenetic trees can not show evolutionary relatedness between samples.
- ◉ Genomic epidemiology is the epidemiological investigation of infectious disease cases, supplemented by genomic data.
- ○ Evergreen Online aims to detect airborne disease outbreaks.

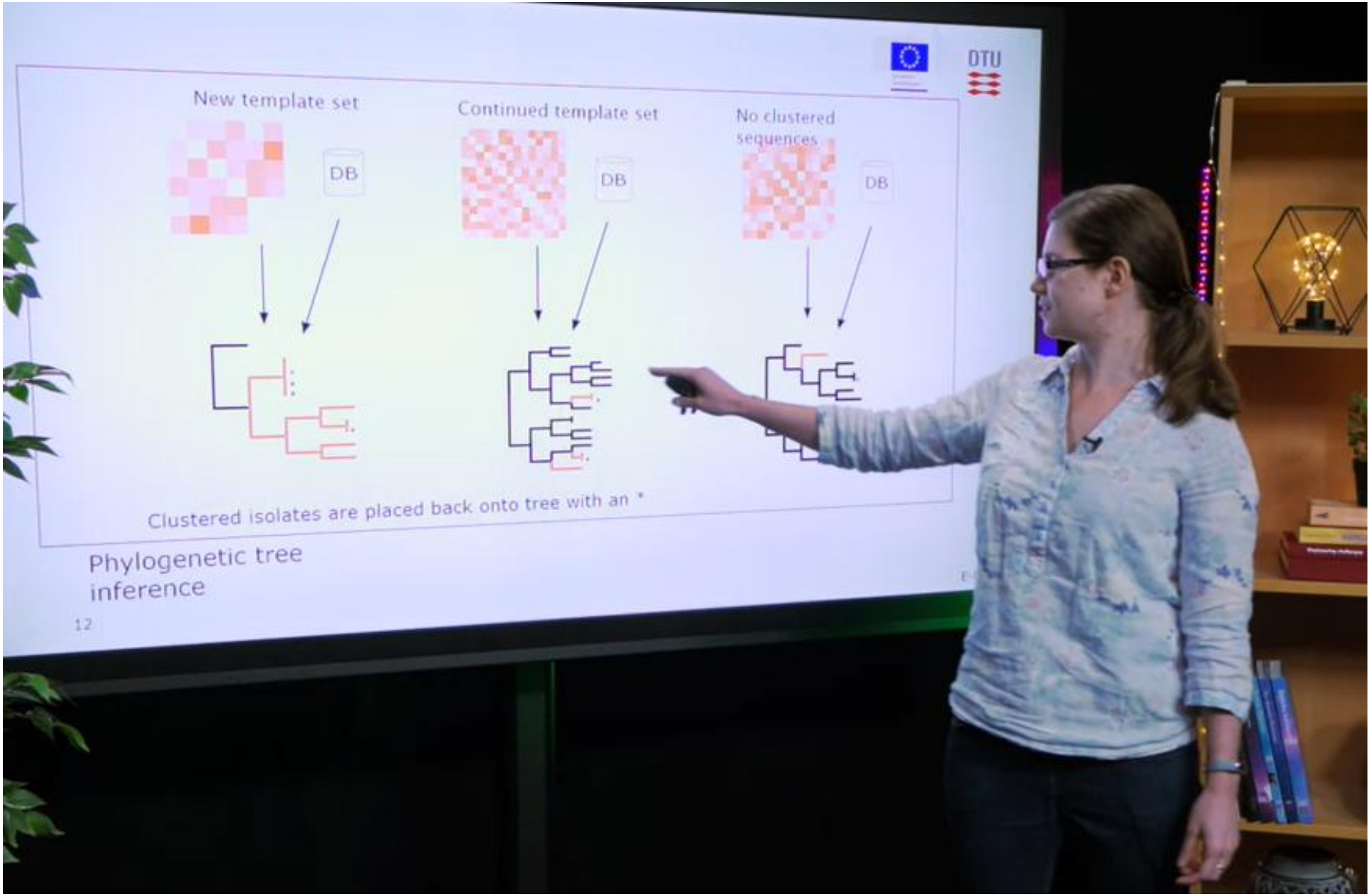
The pipeline behind Evergreen Online

- Downloads raw sequencing data from public repositories, uploaded by public health and food-safety laboratories, etc.
- Divides the data by subtyping
- Calls high-quality SNPs by aligning to references, calculates genetic distances
- Infers phylogenies
- Displays trees and putative clusters


New template set
Continued template set
No clustered sequences
DB
DB
DB
Clustering at 10 SNPs

Phylogenetic tree inference




DTU
New template set
DB
Continued template set
DB
No clustered sequences
DB
Clustered isolates are placed back onto tree with an *
Phylogenetic tree inference
12

Which of the following sentences is TRUE ?

- ○ The bioinformatic pipeline downloads assembled genomes from public repositories.
- ○ All columns with unknown bases are removed before genetic distance calculation.
- ○ The pipeline re-calculates all output each time it runs.
- ◉ Phylogenetic trees are part of the output of the pipeline.

What is the format for your own sequence database ?

- (●) Nucletotide sequences in FASTA
- ( ) Nucleotide sequences in Genbank
- ( ) Protein sequences in FASTA
- ( ) Protein sequences in Genbank