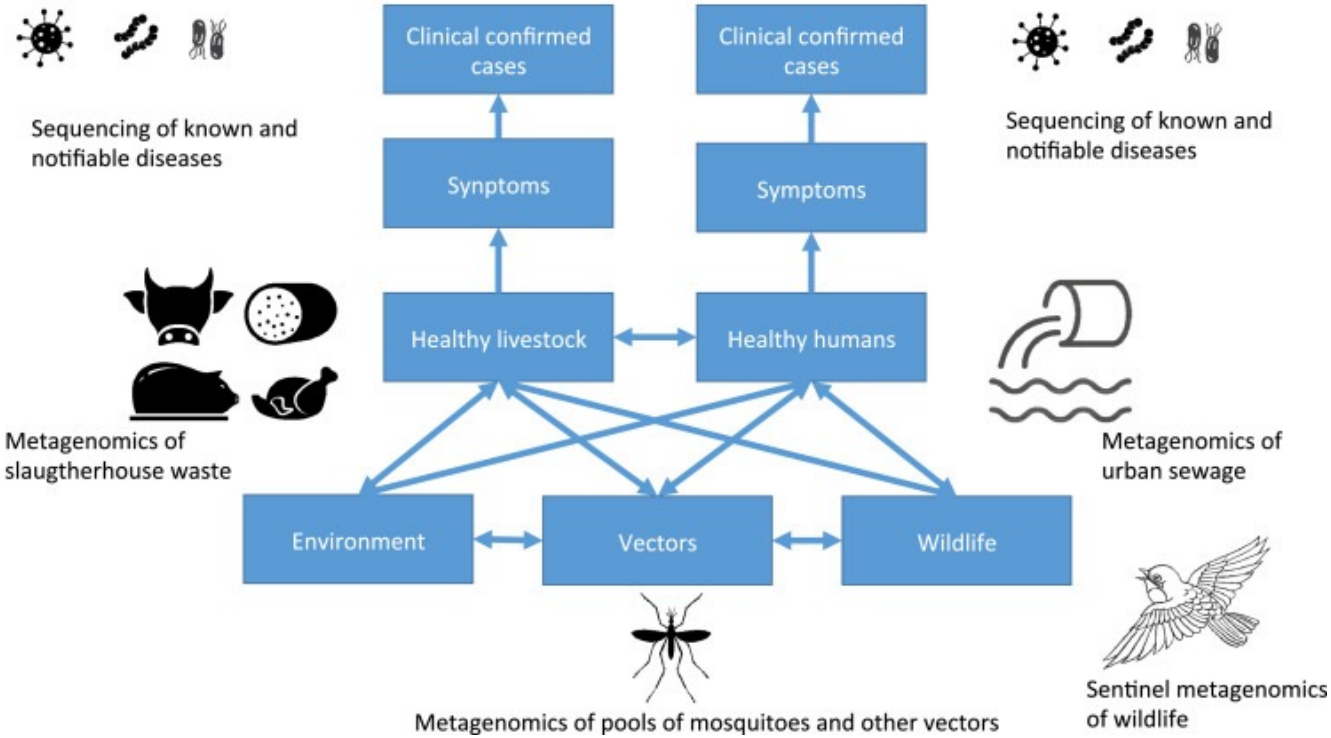
Judit Szarvas, postdoc

Infectious disease bioinformatics, March 2022

Large-scale genomic surveillance of pathogens

Genomic surveillance

- Supplementing disease surveillance with genomics data
- One Health approach:
  - Human and livestock clinical samples whole-genome or amplicon sequencing
  - Environmental metagenomics



https://doi.org/10.1016/j.lanepe.2021.100210

Genomic surveillance

Published: 13 November 2017

Towards a genomics-informed, real-time, global pathogen surveillance system

Jennifer L. Gardy & Nicholas J. Loman

Nature Reviews Genetics **19**, 9–20 (2018) | Cite this article

COMMENT | 07 June 2018

Pandemics: spend on surveillance, not prediction

Trust is undermined when scientists make overblown promises about disease prevention, warn Edward C. Holmes, Andrew Rambaut and Kristian G. Andersen.

Edward C. Holmes, Andrew Rambaut & Kristian G. Andersen

Changing the paradigm for hospital outbreak detection by leading with genomic surveillance of nosocomial pathogens

Sharon J. Peacock[1,2,3], Julian Parkhill[3], Nicholas M. Brown[4]

View Affiliations

First Published: 27 July 2018 | **https://doi.org/10.1099/mic.0.000700**

Genomic surveillance

The use of next generation sequencing for improving food safety: Translation into practice

Balamurugan Jagadeesan [a], Peter Gerner-Smidt [b], Marc W. Allard [c], Sébastien Leuillet [d], Anett Winkler [e], Yinghua Xiao [f], Samuel Chaffron [g], Jos Van Der Vossen [h], Silin Tang [i], Mitsuru Katase [j], Peter McClure [k], Bon Kimura [l], Lay Ching Chai [m], John Chapman [n], Kathie Grant [o]

Show more

+ Add to Mendeley Share Cite

https://doi.org/10.1016/j.fm.2018.11.005

Get rights and content

Open access

Operational burden of implementing *Salmonella* Enteritidis and Typhimurium cluster detection using whole genome sequencing surveillance data in England: a retrospective assessment

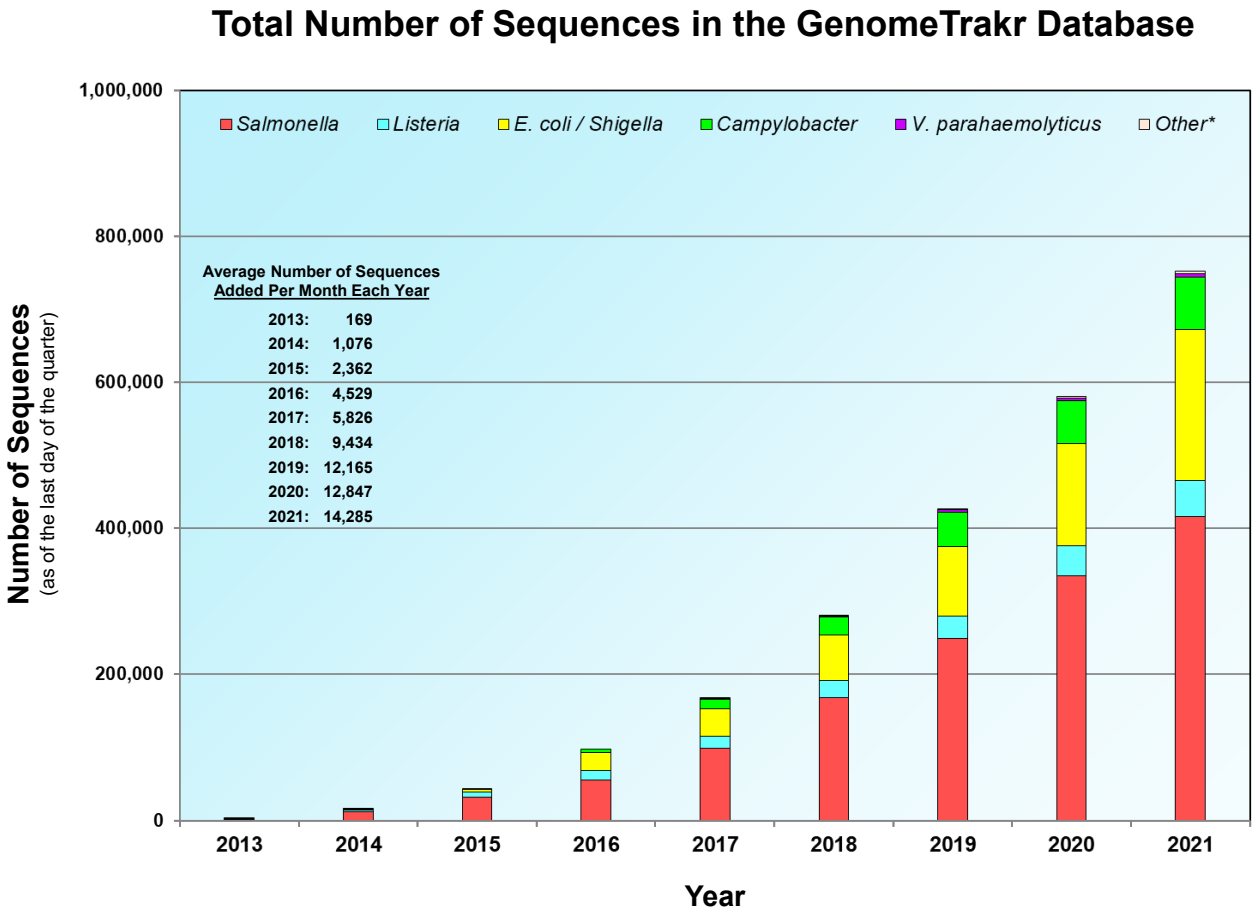
Published online by Cambridge University Press: **02 July 2018**

Piers Mook, Daniel Gardiner, Neville Q. Verlander, Jacquelyn McCormick, Martine Usdin, Paul Crook, Claire Jenkins and Timothy J. Dallman

Show author details

Whole-genome sequencing projects for food safety

- National and international initiatives for WGS of foodborne bacteria

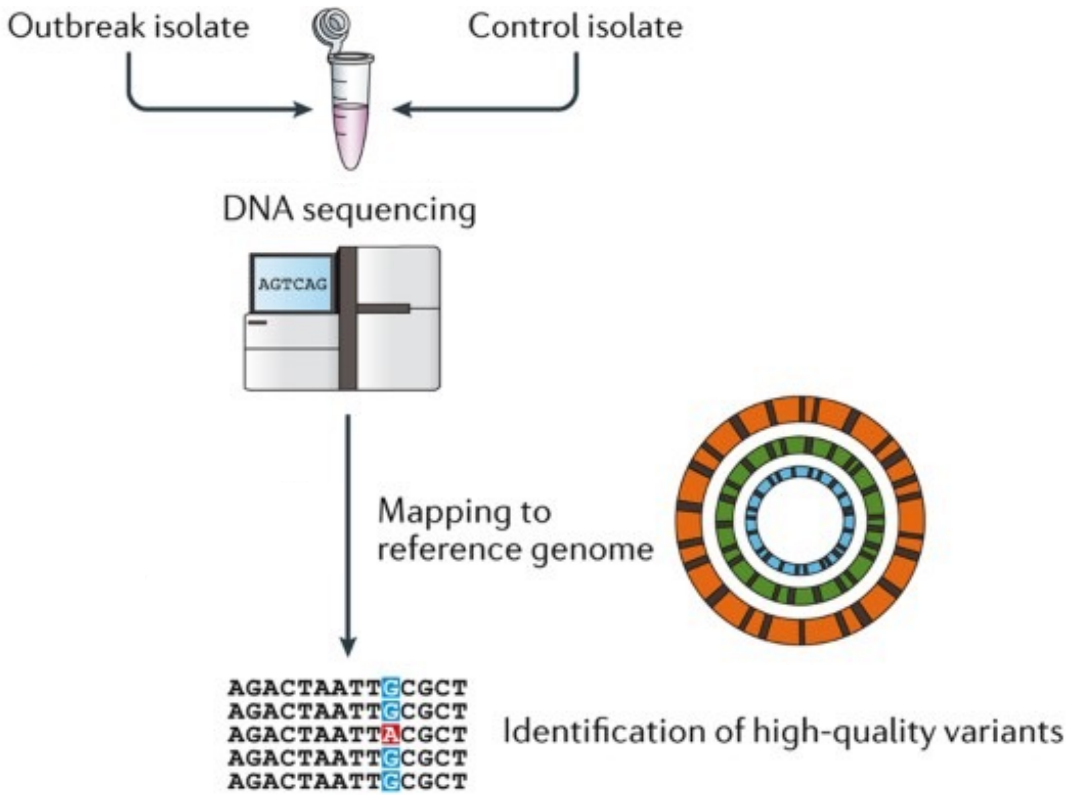


First sequences uploaded in February 2013

* Other pathogens: *Cronobacter, V. vulnificus, C. botulinum, and C. perfringens*

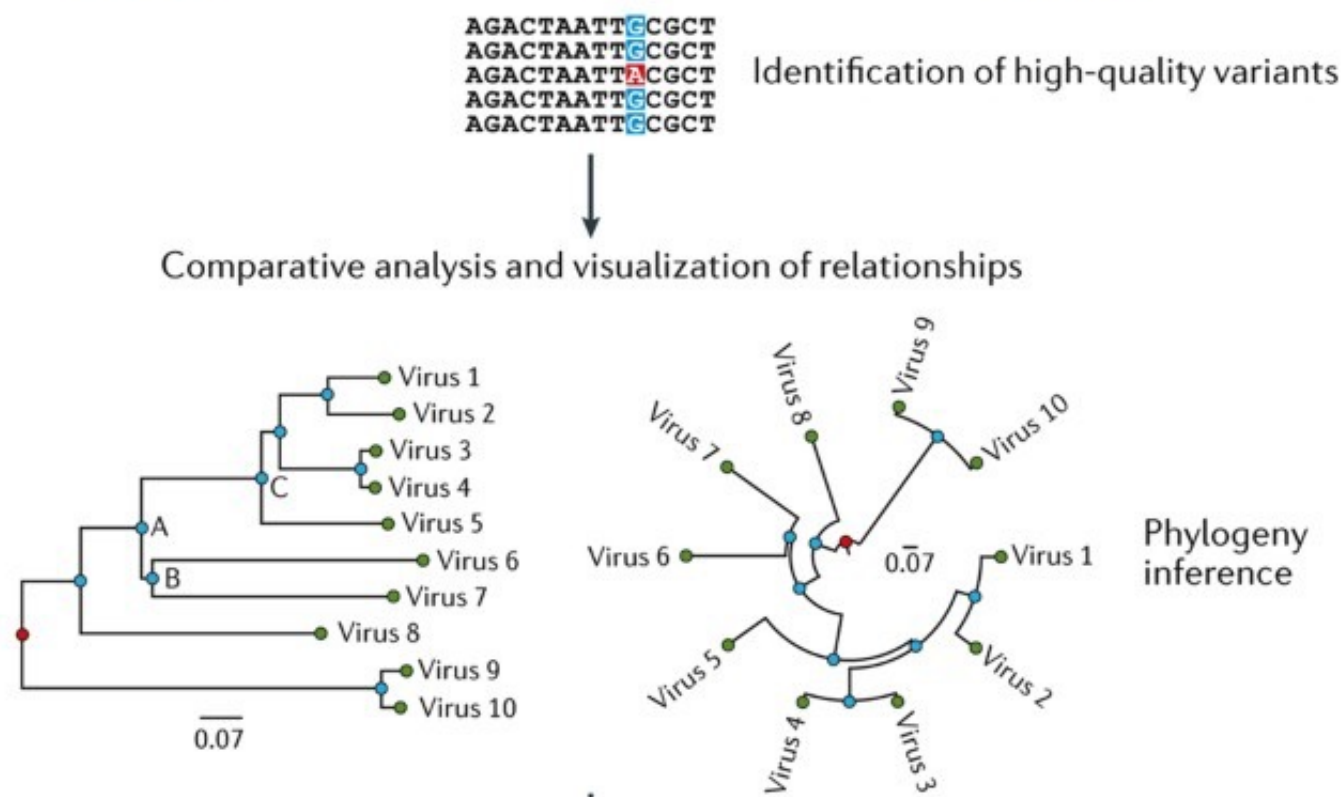
https://www.fda.gov/food/whole-genome-sequencing-wgs-program/genometrakr-fast-facts

SNP-based phylogeny recap



https://doi.org/10.1038/nrg.2017.88

SNP-based phylogeny recap



What if there is a long-time, ongoing outbreak?

+ Historical samples
+ Environmental samples

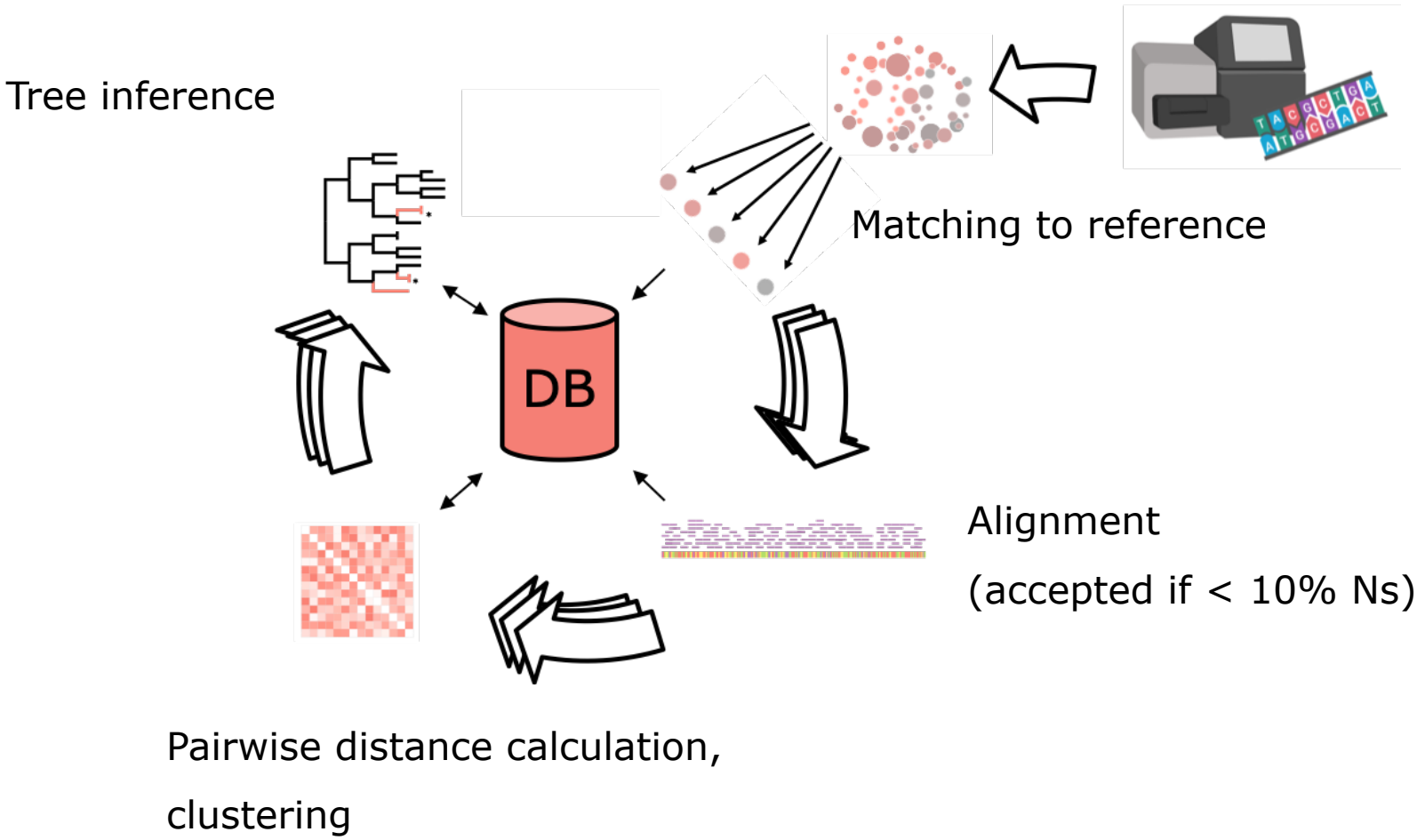
https://doi.org/10.1038/nrg.2017.88

Large-scale genomic surveillance design considerations

- Design a system for large-scale, cyclic analysis of sequencing data
- Reference-based:
  - Allow SNP-level resolution
- Homology reduction within the sets by clustering at x genomic threshold:
  - Connect samples that could be epidemiologically linked

Cyclic analysis of sequencing data

- Intermediate data and information kept between runs
- Eliminate need to re-compute for older samples



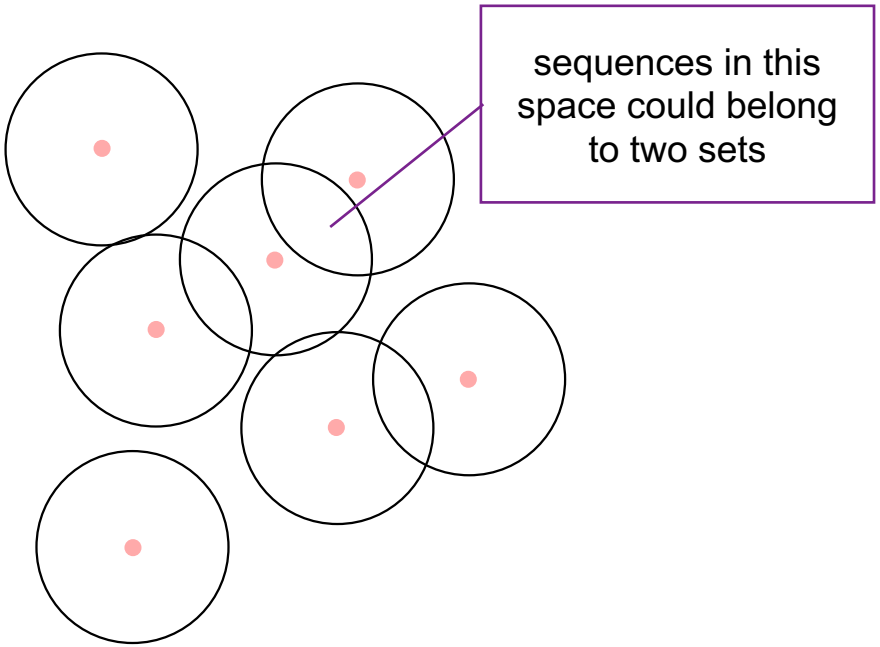
https://doi.org/fmicb.2021.636608

Reference based

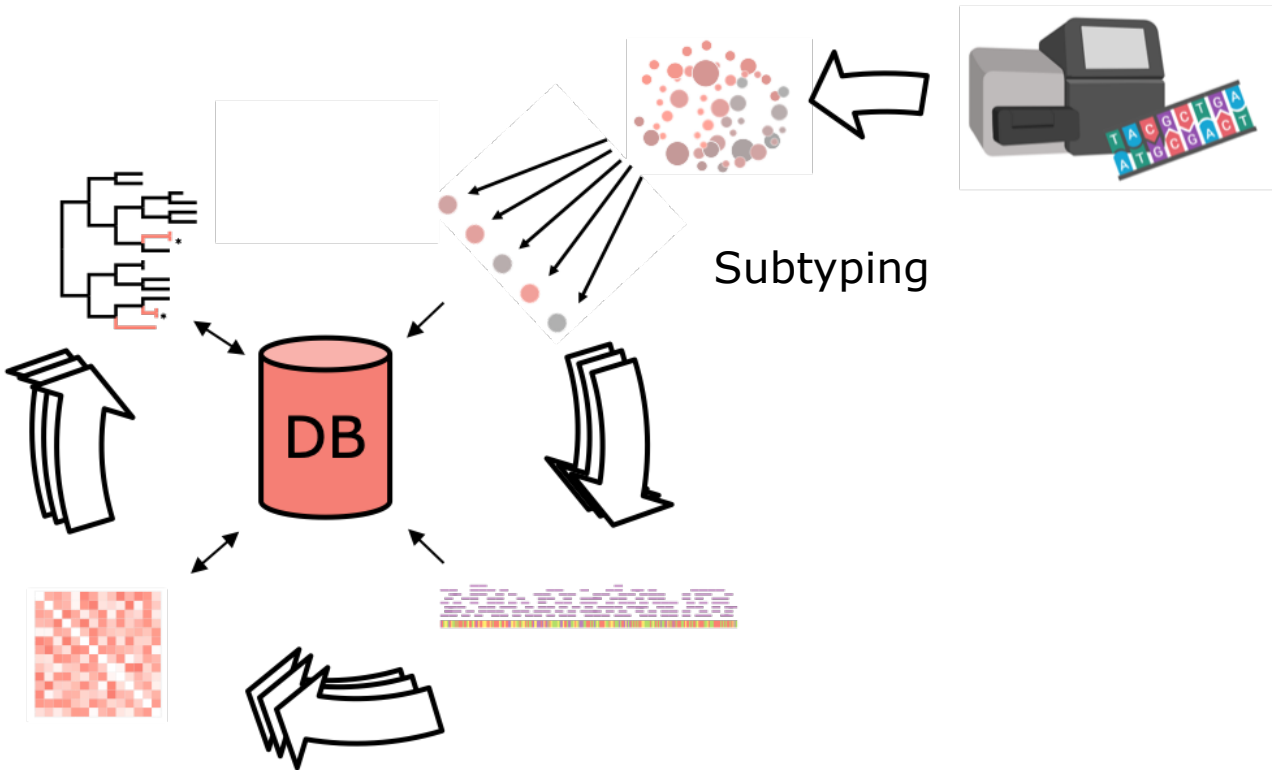
- Allow SNP-level resolution, without the maintenance of allelic-schemes
- Quality controlled references can be supplied by user, flexible for each analysis
- In theory, each sample would be matched to a very close reference, which is required for high resolution SNP analysis
- Divide the data into smaller sets:
  - divide and conquer approach: solve the problem in smaller data sets, reducing the complexity of it

Reference database supplied

- NCBI RefSeq complete bacterial genomes, with plasmids removed
- Homology reduced (usually to 99.0id%):
  - reference sequences, that are more similar to each other, are removed from the created template database



Sequences in sequence space



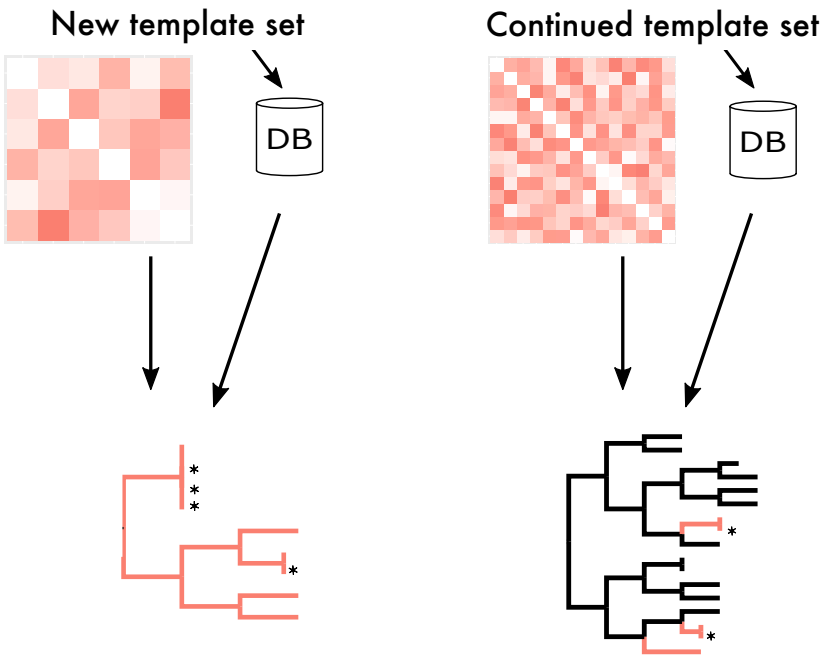
Pairwise distance calculation,

clustering

https://doi.org/fmicb.2021.636608

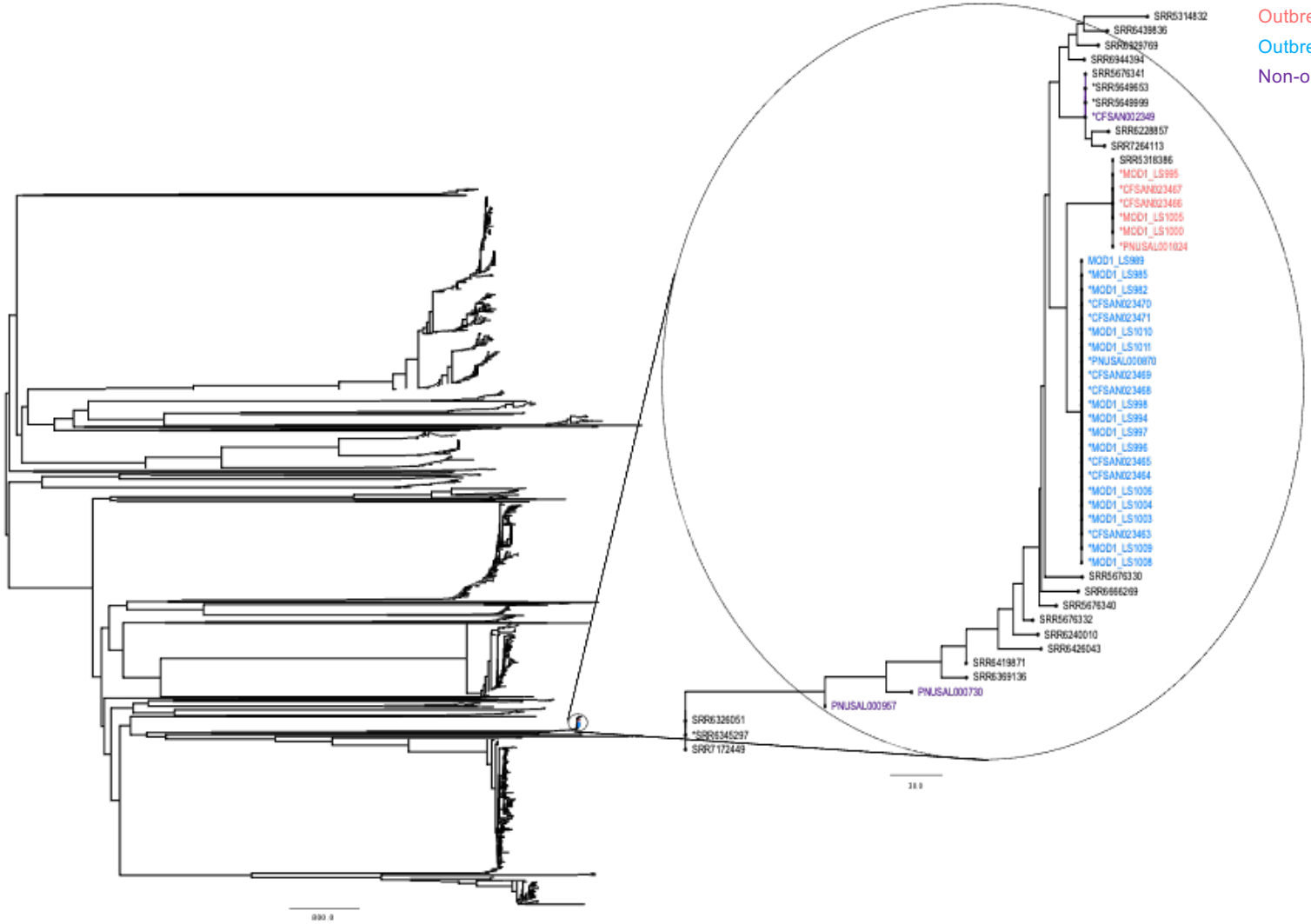
Homology reduction within a set

- Clustering at 10 SNP threshold, keeping only one representative of the cluster
- Further reduce the computational burden by limiting the redundancy in the set
  - 25-40% of samples can be redundant in a set
- Connect samples that could be epidemiologically linked



Clustered isolates are placed back onto tree with an *

Listeria monocytogenes benchmarking



Timme et al. Benchmark datasets for phylogenomic pipeline validation, applications for foodborne pathogen surveillance. PeerJ. 2017 Oct 6;5:e3893.

Implementation

- Python2/3, sqlite database
- Dependencies:
  - KMA
  - IQtree (maximum likelihood inference)
  - Neighbor (Neighbor-joining)

- Bacteria: https://bitbucket.org/genomicepidemiology/evergreen/src/COMPARE/
- Viruses (general use): https://bitbucket.org/jszarvas/viral_surveillance/src/master/

Implementation

- Example command

```
vu_pipeline.py -b /home/user/project/coronaviruses \
-o /home/user/project/coronaviruses/analysis.7.43 \
-f /home/user/project/coronaviruses/input/coronavirus_data.iso \
-g /home/user/project/coronaviruses/input/coronavirus_metadata.tsv \
-d /home/user/project/coronaviruses/coronavirus.7.43.db
-r /home/user/project/coronaviruses/references/coronavirus_kma_7.43 \
-t 85.0 –ml -pairwise –ebi
```

- Example input .iso

```
S000001    /path/to/read_01_1.fastq.gz,/path/to/read_01_2.fastq.gz
S000002    /path/to/read_02_1.fastq.gz,/path/to/read_02_2.fastq.gz
```

Implementation

- Logging is printed to stdout, error messages to stderr
- Output files for each reference, that collected more than 2 samples
  - *.newick: phylogenetic tree (dist and ml for neighbor-joining and maximum likelihood) with metadata appended to taxa labels
  - *.nwk and *.tsv: Microreact compatible output
  - *.mat: phylip distance matrix for the non-redundant samples
  - *.aln: optional multiple-alignment

Exercise

- 3 parts:
  - 1st: input files are prepared for you, and you need to write the job-script
  - 2nd: you need to create the correct input files, and run a second analysis cycle
  - 3rd: transfer the outputs from the two rounds to your computer and answer the questions

Other solutions

- K-mers and wgMLST:
  - https://www.ncbi.nlm.nih.gov/pathogens/about
- cgMLST:
  - https://pathogen.watch
- cgMLST, wgMLST, rMLST, etc:
  - https://enterobase.warwick.ac.uk/
- SNPs:
  - https://nextstrain.org/pathogens

DTU