WGS-based typing methods and nomenclature

Egle Kudirkiene

2nd multidisciplinary training workshop

October 2023

Content

Introduction to WGS-based bacteria genotyping

SNP: reference - based single nucleotide polymorphism analysis

    CSI Phylogeny tool
    open source and easy to use

cgMLST: core genome multi-locus sequence typing

    SeqSphere and EnteroBase tools
    potential for standartisation and result sharing between the laboratories

Introduction

Bacteria genotyping

Genotyping is the process of determining differences in the **genetic make-up (genotype) of an isolate** by examining its DNA sequence by comparing it with another isolates sequence or a reference sequence.

Clone

Isolates of bacterial species that are **indistinguishable in genotype are assigned as a clone**

Cluster

Instead, in outbreak investigations we use **clusters of isolates with nearly identical genomes** to consider pathogen mutation rates in different hosts/environments and time
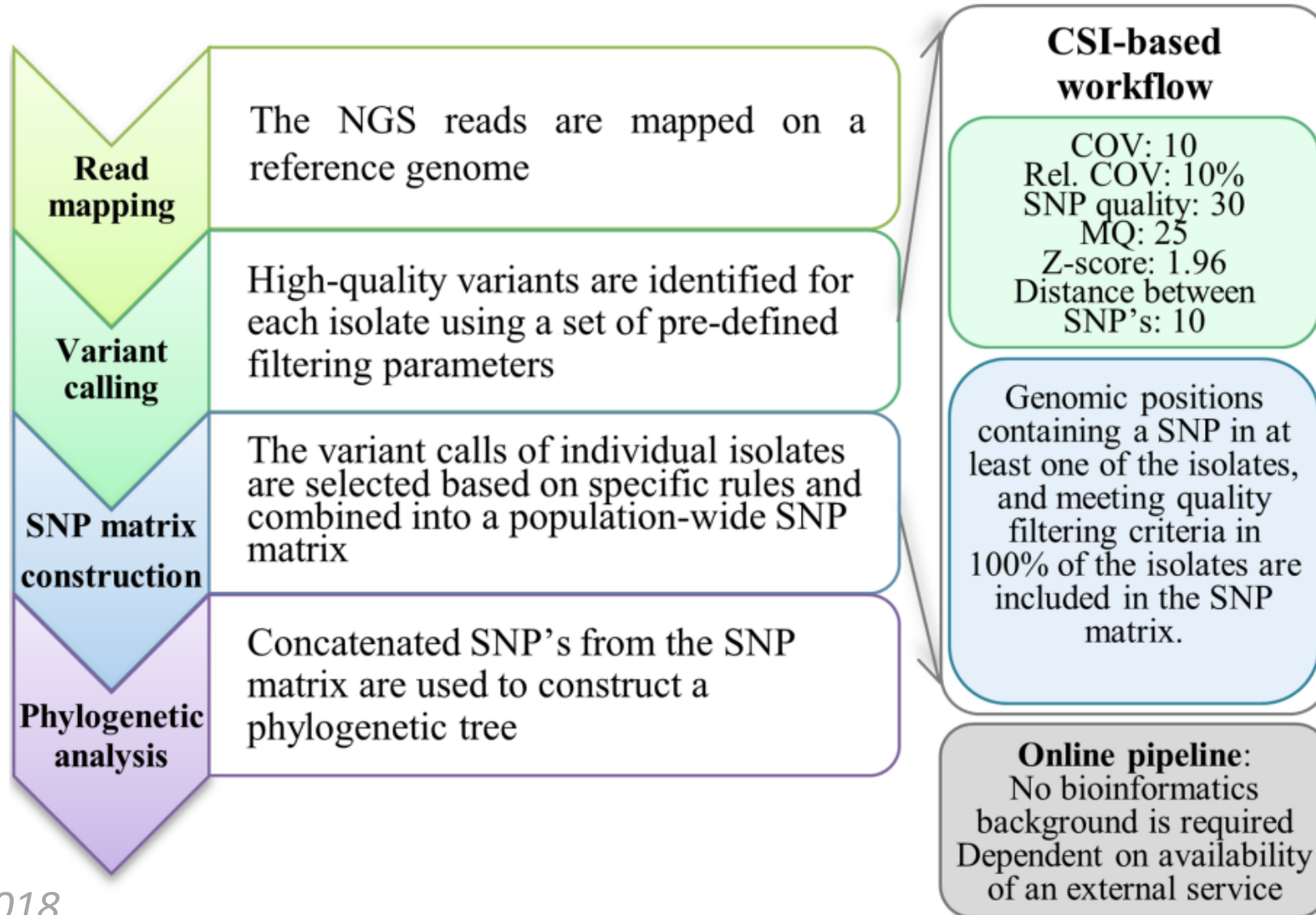
Cluster cut-offs for cgMLST and SNP analyses:
- *Salmonella* - depends on the serovar
  - 0-3 ADs/SNPs in clonal serovars and
  - up to 5 AD/SNP in other serovars
- *Campylobacter* - 5 or less ADs/SNPs

Proposed protocol for whole genome sequencing-based analysis for detection and tracing of epidemic clones of antimicrobial resistant *Salmonella* and *Campylobacter*
- to be used for national surveillance and integrated outbreak investigations by NRLs for public health

8 July 2022

Schematic CSI Phylogeny workflow

FWD AMR·
RefLabCap

Read mapping
The NGS reads are mapped on a reference genome

Variant calling
High-quality variants are identified for each isolate using a set of pre-defined filtering parameters

SNP matrix construction
The variant calls of individual isolates are selected based on specific rules and combined into a population-wide SNP matrix

Phylogenetic analysis
Concatenated SNP's from the SNP matrix are used to construct a phylogenetic tree

CSI-based workflow

COV: 10
Rel. COV: 10%
SNP quality: 30
MQ: 25
Z-score: 1.96
Distance between SNP's: 10

Genomic positions containing a SNP in at least one of the isolates, and meeting quality filtering criteria in 100% of the isolates are included in the SNP matrix.

Online pipeline:
No bioinformatics background is required
Dependent on availability of an external service

Reference-based SNP analysis using CSI Phylogeny

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

Read mapping to the reference genome
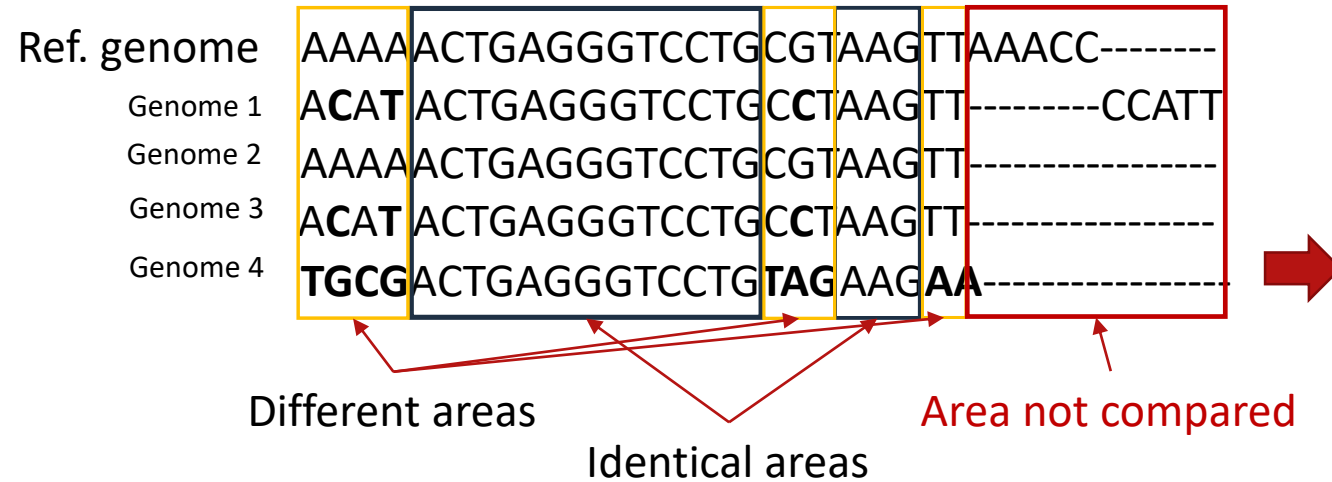
FWD AMR·
RefLabCap

Ref. genome AAAA ACTGAGGGTCCTG CGT AAG TT AAACC--------

Genome 1 A**CAT** ACTGAGGGTCCTG C**C**T AAG TT ---------CCATT

Genome 2 AAAA ACTGAGGGTCCTG CGT AAG TT ----------------

Genome 3 A**CAT** ACTGAGGGTCCTG C**C**T AAG TT ----------------

Genome 4 **TGCG** ACTGAGGGTCCTG **TAG** AAG **AA** ----------------

Different areas

Identical areas

Area not compared

Pairwise similarity matrix (.txt format)

	Ref. Genome	Genome 2	Genome 1	Genome 3	Genome 4
Ref. Genome		0	0	3	3	9
Genome 2	0	0	3	3	9
Genome 1	3	3	0	0	9
Genome 3	3	3	0	0	9
Genome 4	9	9	9	9	0

Percentage of reference genome covered by all isolates 71.4734023710814
3504699 positions was found in all analyzed genomes.
Size of reference genome: 4903501

Below is listed the number of positions that are shared and t

> **false lower number of SNPs** if you choose a distant reference = false assignment to the same genotype

File	Valid positions	Pct. of reference
TC2021-05_R1.ignored_snps	3978591	81.137762590443
TC2021-12_R1.ignored_snps	4307863	87.852801498358
TC2021-02_R1.ignored_snps	4039549	82.3809151869246

How to choose the reference

The reference should be somewhat similar to the isolates you test (>90%, if possible):

- well annotated, high quality, closed genome of same or similar ST (7 gene MLST)

    - use KmerFinder (CGE Tools) to search for ref. genome

- draft genome sequence from own dataset (e.g. representing the first case)

- consensus sequence of all genomes in own dataset

Genotyping using SNP (e.g. CSI Phylogeny)

Pairwise similarity matrix (.txt format)

	Ref. Genome	Genome 2	Genome 1	Genome 3	Genome 4
Ref. Genome	0	0	3	3	9
Genome 2	0	0	3	3	9
Genome 1	3	3	0	0	9
Genome 3	3	3	0	0	9
Genome 4	9	9	9	9	0

	Ref. Genome	Genome 2	Genome 1	Genome 3	Genome 4
Ref. Genome	0	0	3	3	9
Genome 2	0	0	3	3	9
Genome 1	3	3	0	0	9
Genome 3	3	3	0	0	9
Genome 4	9	9	9	9	0

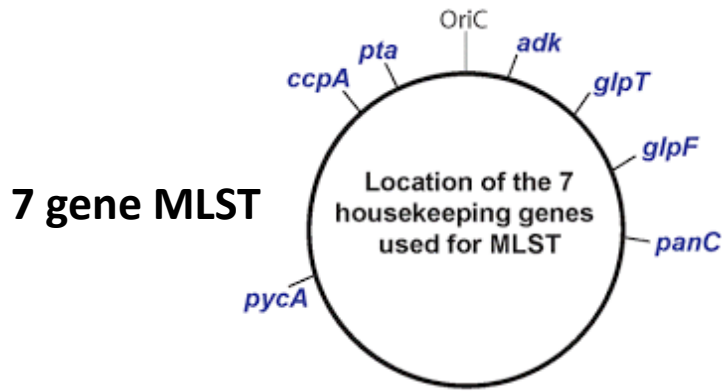**Phylogenetic tree (.newick file format)**
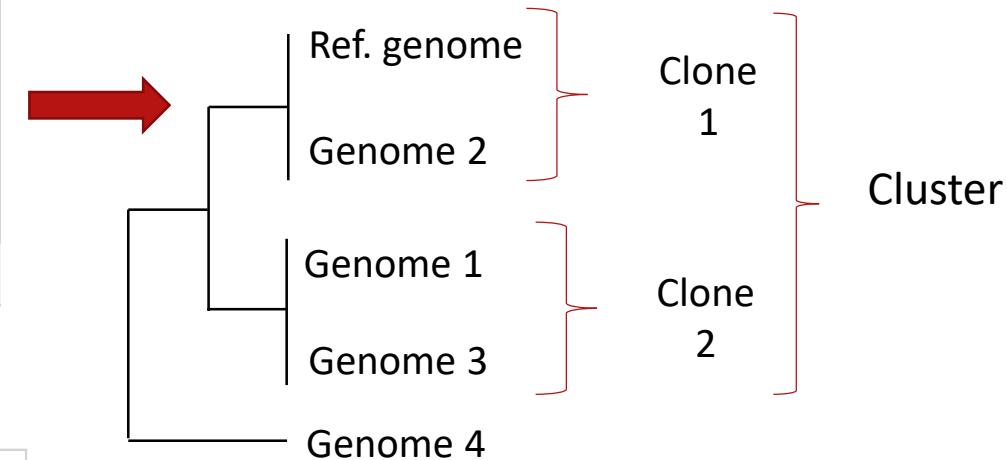
The user defines clusters based on selected SNP thresholds, e.g. 0-3 SNPs

cgMLST (SeqSphere and Enterobase)

Core genome multi-locus sequence typing (cgMLST)

Each gene variant has an allele number

7 gene MLST

Sequence Type: *19*

Locus	Identity	Coverage	Alignment Length	Allele Length	Gaps	Allele
aroC	100	100	501	501	0	aroC_10
dnaN	100	100	501	501	0	dnaN_7
hemD	100	100	432	432	0	hemD_12
hisD	100	100	501	501	0	hisD_9
purE	100	100	399	399	0	purE_5
sucA	100	100	501	501	0	sucA_9
thrA	100	100	501	501	0	thrA_2
Gene08						
Gene09						
Gene10						
Gene11						
Gene12						
Gene13						
Gene14						
Gene15						
Gene16						
Gene17						
Gene18						
Gene19						
Gene20						
Gene21						
Gene22						
Gene23						

Identical allele pattern = clone

cgMLST contains >1000 genes

cgMLST schemes and allele calling methods (most common)

cgMLST schemes	No. of alelles
Salmonella: EnteroBase (Atchman et al., 2021)	3002
Campylobacter: Oxford (Cody et al. 2017)	1343

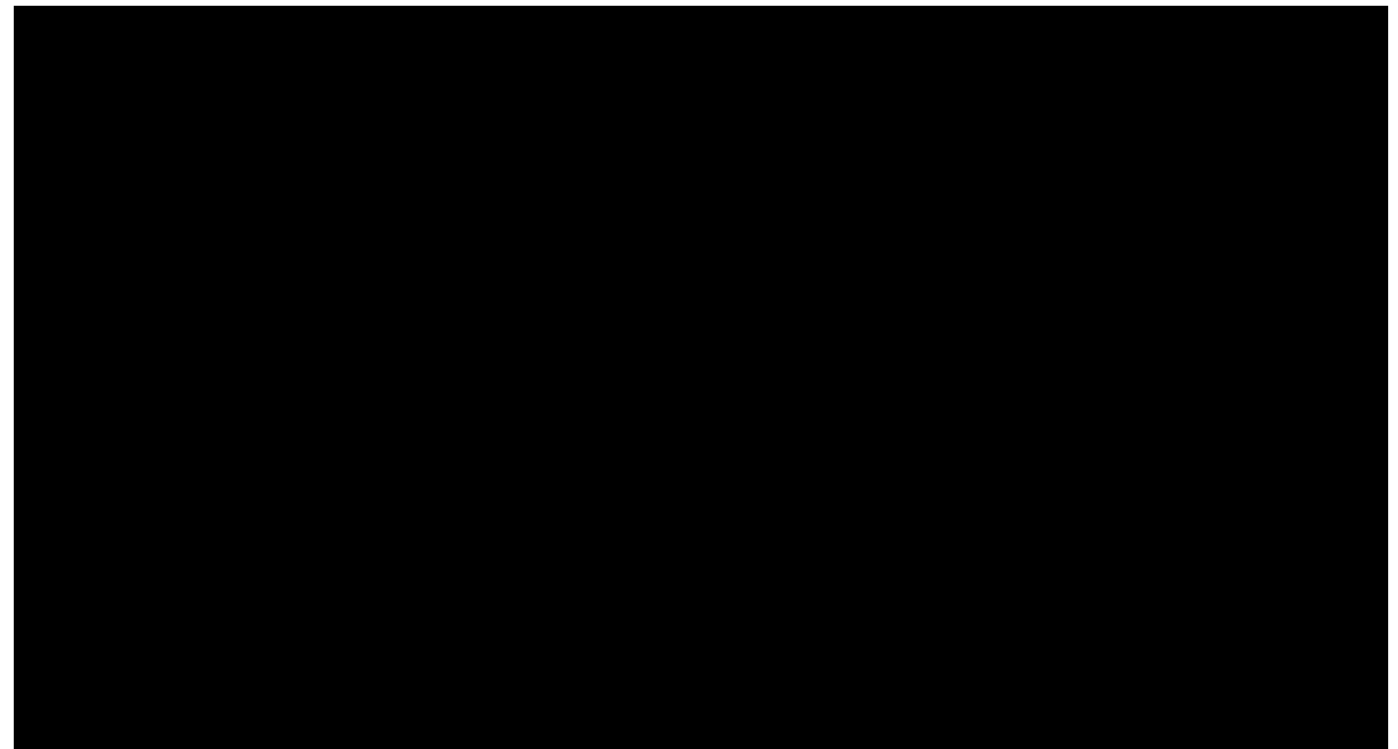**Allele calling workflow in SeqSphere**

Different allele calling methods in use:

EnteroBase, SeqSphere, BioNumerics, chewBACCA, MentaLIST

May result in different allele profiles

Cluster detection using cgMLST

Clusters in SeqSphere
- Default cut-offs

HierCC in EnteroBase scheme for Salmonella using default cut-offs
- 0, 2, 5, 10, 20, 50, 100, etc.

Same thresholds can be used for cluster detection

HOWEVER

To assure proper outbreak definitions, the users may and often choose different cluster cut-offs depending on the species/serovar and in accordance to the epidemiological data of time, place and person.

More about that in the lecture from Pernille

Assignment to clusters by SeqSphere

Default cluster cut-off is ≤6

Clusters of samples with **maximum six alleles distance** are shaded

AD of ≤6 is **not suitable** to recognise an outbreak of RT 181

Assignment to HierCC by EnteroBase

The clusters are assigned **stable cluster group numbers** at different, fixed cgMLST allele distances. *Salmonella* for instance, has cut-offs such as 0, 2, 5, 10, 20, 50, 100, etc.

	ST	HC0 (indistinguis...	HC2	HC5	HC10	HC20	HC50	HC100
👁	35462	35462	35462	719	52	25	25	25
👁	6020	6020	6020	6020	52	25	25	25
👁	35014	35014	35014	35014	15910	15910	25	25
👁	34712	34712	34712	34712	34712	15406	25	25
👁	34679	34679	16899	719	52	25	25	25
👁	33958	33958	33958	33958	33958	33958	25	25
👁	33956	33956	16899	719	52	25	25	25
👁	33752	33752	33752	33752	33752	33752	33752	33752
👁	32538	32538	32538	32538	52	25	25	25
👁	31161	31161	16899	719	52	25	25	25
👁	31161	31161	16899	719	52	25	25	25
👁	23212	23212	23212	23212	23212	17337	25	25
👁	19728	19728	19728	15406	15406	15406	25	25
👁	19652	19652	15751	15751	15751	13072	25	25
👁	19648	19648	19648	19648	19648	13072	25	25
👁	19647	19647	15751	15751	15751	13072	25	25
👁	19637	19637	19637	19637	15910	15910	25	25

Workspace:Dublin_global_ST10   Rows Total:1761 Filtered:1696
Edit Mode: ☐   Experimental Data cgMLST V2 + HierCC V1

Cluster based on H10, but not on H5, H2, H0.

The user decides which threshold to use for recognizing the same outbreak/cluster group

Cluster! And even a clone!

Same ST and cluster group at any cut-off

Summary

Reference-based SNP-based typing

- Species non-specific

- Require user defined reference genome

- The genetic diversity of the reference and dataset influence the discriminatory power of the method

- SNPs in the whole genome, both coding and non-coding sequences

- User defined genotypes. User defined cluster cut-offs

cgMLST-based typing

- Species specific

- No reference genome required. Requires database that is updated continuesly

- Different allele calling methods may result in different allelic profile

- Allelic differences between the genes under the analysis

- The unique combination of alleles is the sequence type (ST1-STXXX…). Default and user defined cluster cut-offs