Session 4.2 – Gene-by-Gene analysis

Sara Monzón Fernández

BU-ISCIII

Unidades Comunes Científico Técnicas – SGSAFI-ISCIII

16-20 Octubre 2022, 5ª Edición
Programa Formación Continua, ISCIII

Index

Gene-by-Gene analysis and comparison with SNP-based approaches:

- Download wg/cgMLST schemes.
- Software for gene-by-gene analysis.
  - SeqSphere
  - Chewbacca
  - Taranis
- Gene-by-gene vs SNP-based approaches

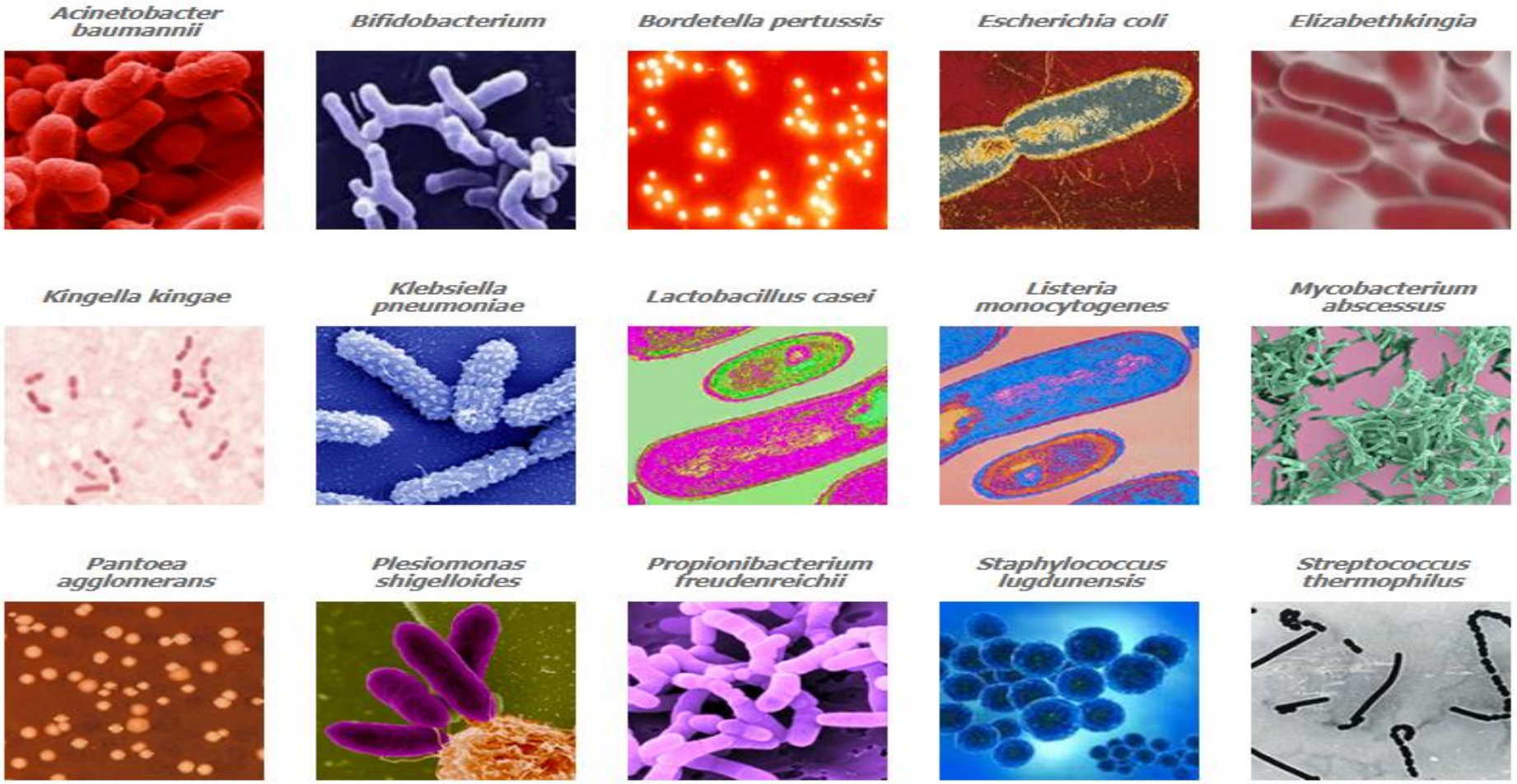
cg/wgMLST schemes available

- PubMLST
- Pasteur bigsDB database
- Enterobase

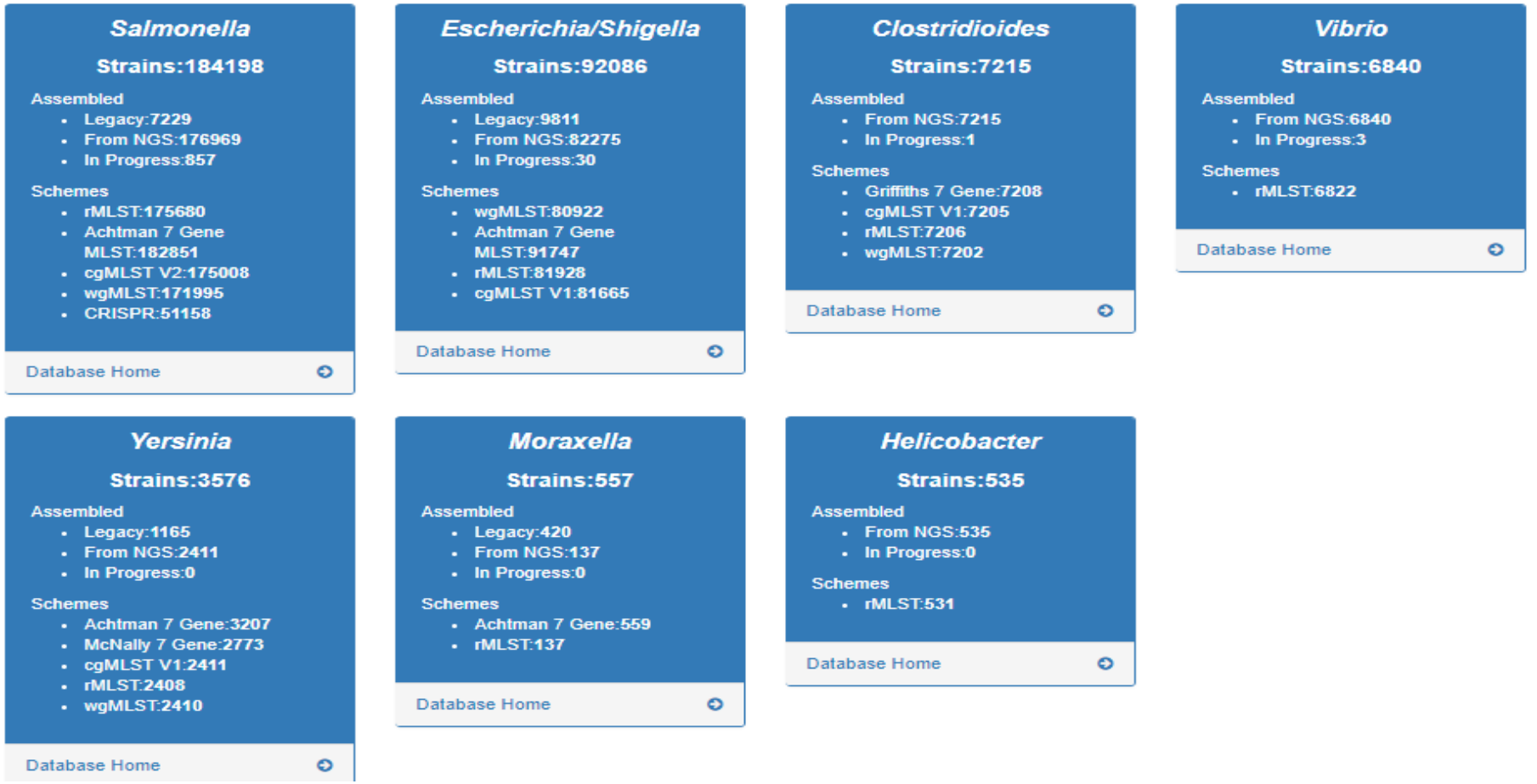
cg/wgMLST schemes available

- *Achromobacter*
- *Acinetobacter baumannii*
- *Aeromonas spp.*
- *Anaplasma phagocytophilum*
- *Arcobacter spp.*
- *Bacillus cereus*
- *Bacillus licheniformis*
- *Bacillus subtilis*
- *Bordetella spp.*
- *Borrelia spp.*
- *Bartonella bacilliformis*
- *Bartonella henselae*
- *Brachyspira spp.*
- *Brucella spp.*
- *Burkholderia cepacia complex*
- *Burkholderia pseudomallei*
- *Campylobacter spp.*
- *Carnobacterium maltaromaticum*
- *Chlamydiales spp.*
- *Citrobacter freundii*
- *Clostridium botulinum*
- *Clostridium difficile*
- *Clostridium septicum*
- *Corynebacterium diphtheriae*
- *Cronobacter spp.*
- *Dichelobacter nodosus*
- *Enterobacter cloacae*
- *Edwardsiella spp.*
- *Enterococcus faecalis*
- *Enterococcus faecium*
- *Escherichia spp.*
- *Flavobacterium psychrophilum*
- *Gallibacterium anatis*
- *Haemophilus influenzae*
- *Haemophilus parasuis*
- *Helicobacter cinaedi*
- *Helicobacter pylori*
- *Helicobacter suis*
- *Klebsiella aerogenes*
- *Klebsiella oxytoca*
- *Lactobacillus salivarius*
- *Leptospira spp.*
- *Macrococcus canis*
- *Macrococcus caseolyticus*
- *Mannheimia haemolytica*
- *Melissococcus plutonius*
- *Mycobacteria spp.*
- *Mycobacterium abscessus complex*
- *Mycoplasma agalactiae*
- *Mycoplasma bovis*
- *Mycoplasma hyopneumoniae*
- *Mycoplasma hyorhinis*
- *Mycoplasma iowae*
- *Mycoplasma pneumoniae*
- *Mycoplasma synoviae*
- *Neisseria spp.*
- *Oral Streptococcus spp.*
- *Orientia tsutsugamushi*
- *Ornithobacterium rhinotracheale*
- *Paenibacillus larvae*
- *Pasteurella multocida*
- *Pediococcus pentosaceus*
- *Photobacterium damselae*
- *Piscirickettsia salmonis*
- *Porphyromonas gingivalis*
- *Propionibacterium acnes*
- *Pseudomonas aeruginosa*
- *Pseudomonas fluorescens*
- *Rhodococcus equi*
- *Riemerella anatipestifer*
- *Sinorhizobium spp.*
- *Salmonella spp.*
- *Staphylococcus aureus*
- *Staphylococcus epidermidis*
- *Staphylococcus haemolyticus*
- *Staphylococcus hominis*
- *Staphylococcus pseudintermedius*
- *Stenotrophomonas maltophilia*
- *Streptococcus agalactiae*
- *Streptococcus bovis/equinus complex*
- *Streptococcus canis*
- *Streptococcus dysgalactiae*
- *Streptococcus gallolyticus*
- *Streptococcus pneumoniae*
- *Streptococcus pyogenes*
- *Streptococcus suis*
- *Streptococcus thermophilus*
- *Streptococcus uberis*
- *Streptococcus zooepidemicus*
- *Streptomyces spp.*
- *Taylorella spp.*
- *Tenacibaculum spp.*
- *Treponema pallidum subsp. pallidum*
- *Ureaplasma spp.*
- *Vibrio spp.*
- *Vibrio cholerae*
- *Vibrio parahaemolyticus*
- *Vibrio tapetis*
- *Vibrio vulnificus*
- *Wolbachia spp.*
- *Xylella fastidiosa*
- *Yersinia pseudotuberculosis* (legacy)
- *Yersinia spp.* (legacy)
- *Yersinia ruckeri*

cg/wgMLST schemes available



cg/wgMLST schemes available



What is a cg/wgMLST schema?

- Set of fasta files with sequence for genes belonging to the core genome or pangenome of a bacterial of interest.
- Moreover all alleles in the population are stored in the database.

Gene 1 fasta file

>Allele 1

>Allele 2

…

> Allele n

Gene 2 fasta file

>Allele 1

>Allele 2

…

> Allele n

……

Gene N fasta file

>Allele 1

>Allele 2

…

> Allele n

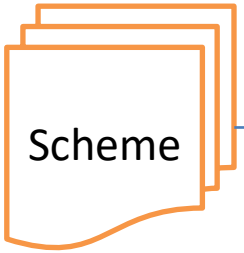
Allele calling

- An allele is a specific sequence variant that occurs at a given locus.
- Given a DNA sequence, the assignment of a putative allele to a locus can be confounded by several factors:
  - Quality of the sequence assembly (influenced by several aspects, such as the sequencing method, the assembler used, etc);
  - If the alleles must correspond to coding sequences (CDSs);
  - Presence of possibly homologous loci (this situation can result in a wrong allele assignment to a given locus given the difficulty in distinguishing closely related homologs

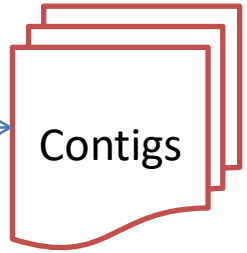
Software

Software	Type
SeqSphere	Commercial
AppliedMaths – Bionumerics	Commercial
ChewBBACA	Free
Taranis – beta	Free

SeqSphere



Reference allele



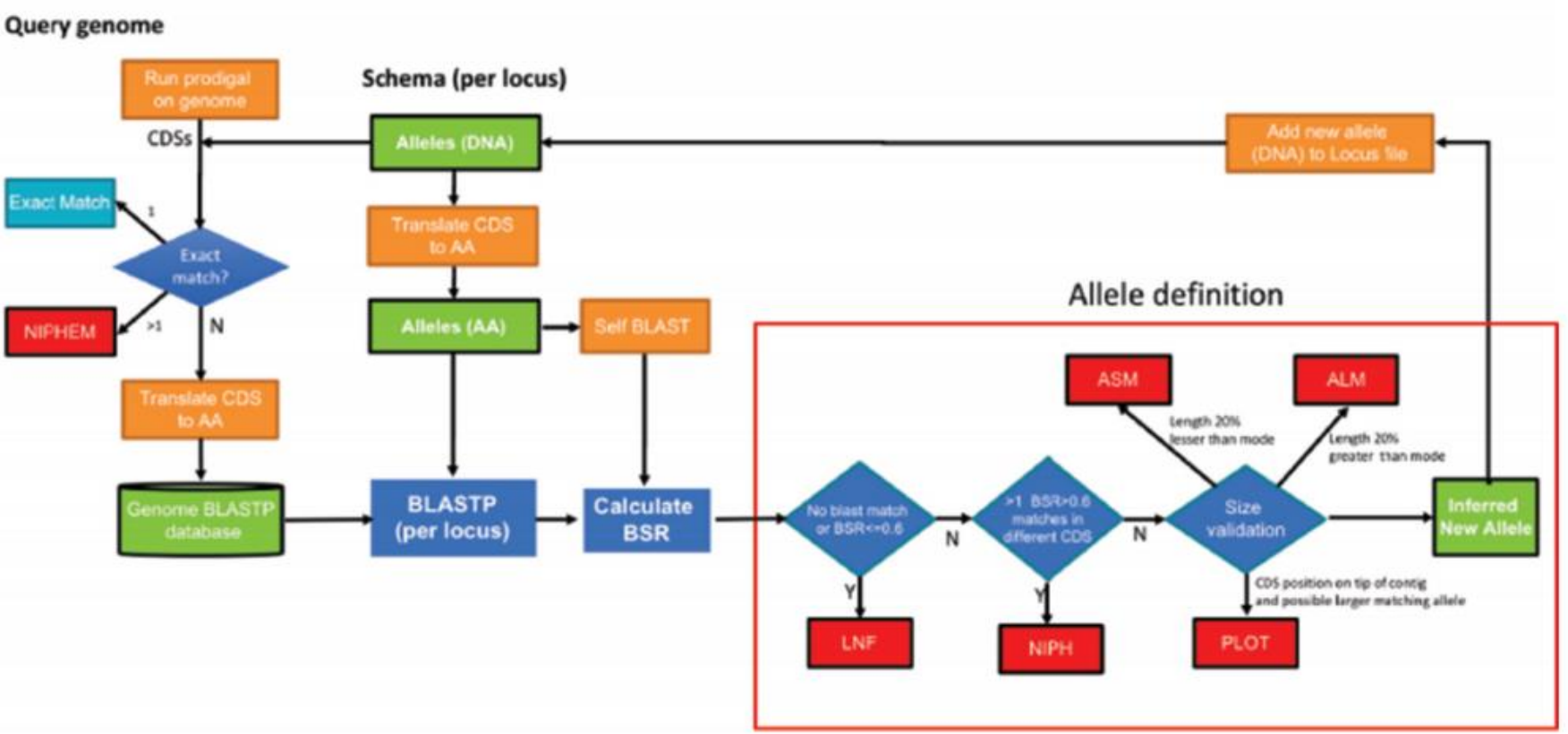
Blast
90% identity
99% coverage

Sample allele
100% identity DB Allele <- Allele nomenclature
Else <- Infere new allele

Sample	Gene1	Gene2
Sample1	1	45
Sample2	1	24

Minimum Spanning tree

ChewBBACA



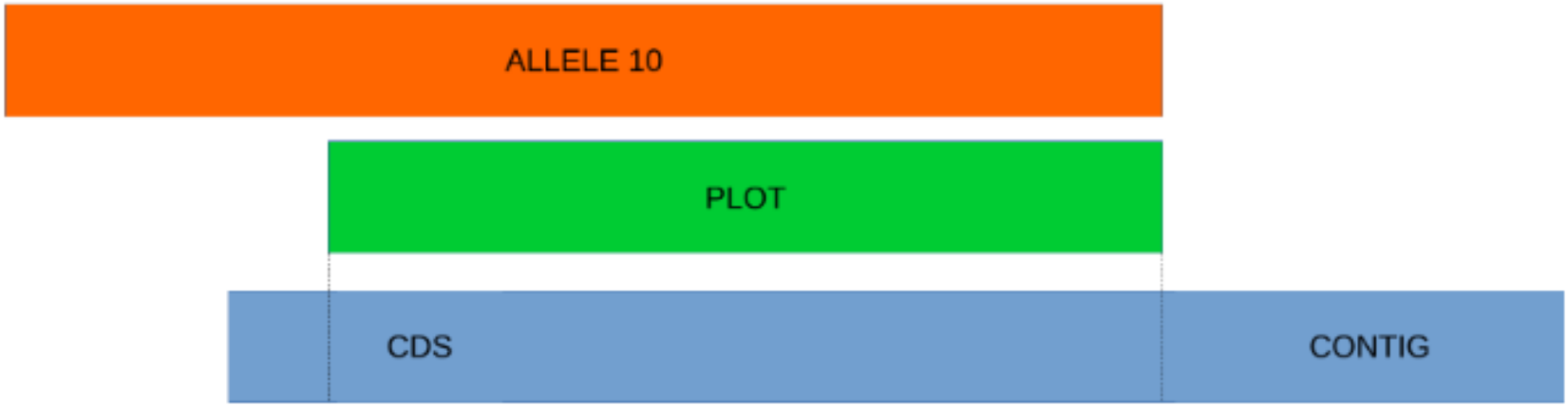
Silva et al. Microbial Genomics. 2018

ChewBBACA

- ***EXC*** - alleles which have exact matches (100% DNA identity) with previously identified alleles
- ***INF*** - inferred new alleles using Prodigal CDS predictions
- ***LNF*** - loci not found. No alleles were found for the number of loci in the schema shown. This means that, for those loci, there were no BLAST hits or they were not within the BSR threshold for allele assignment.
- ***PLOT*** - possible loci on the tip of the query genome contigs (see image below). A locus is classified as *PLOT* when the CDS of the query genome has a BLAST hit with a known larger allele that covers the CDS sequence entirely and the unaligned regions of the larger allele exceeds one of the query genome contigs ends. This could be an artifact caused by genome fragmentation resulting in a shorter CDS prediction by Prodigal. To avoid locus misclassification, loci in such situations are classified as *PLOT*.

ChewBBACA

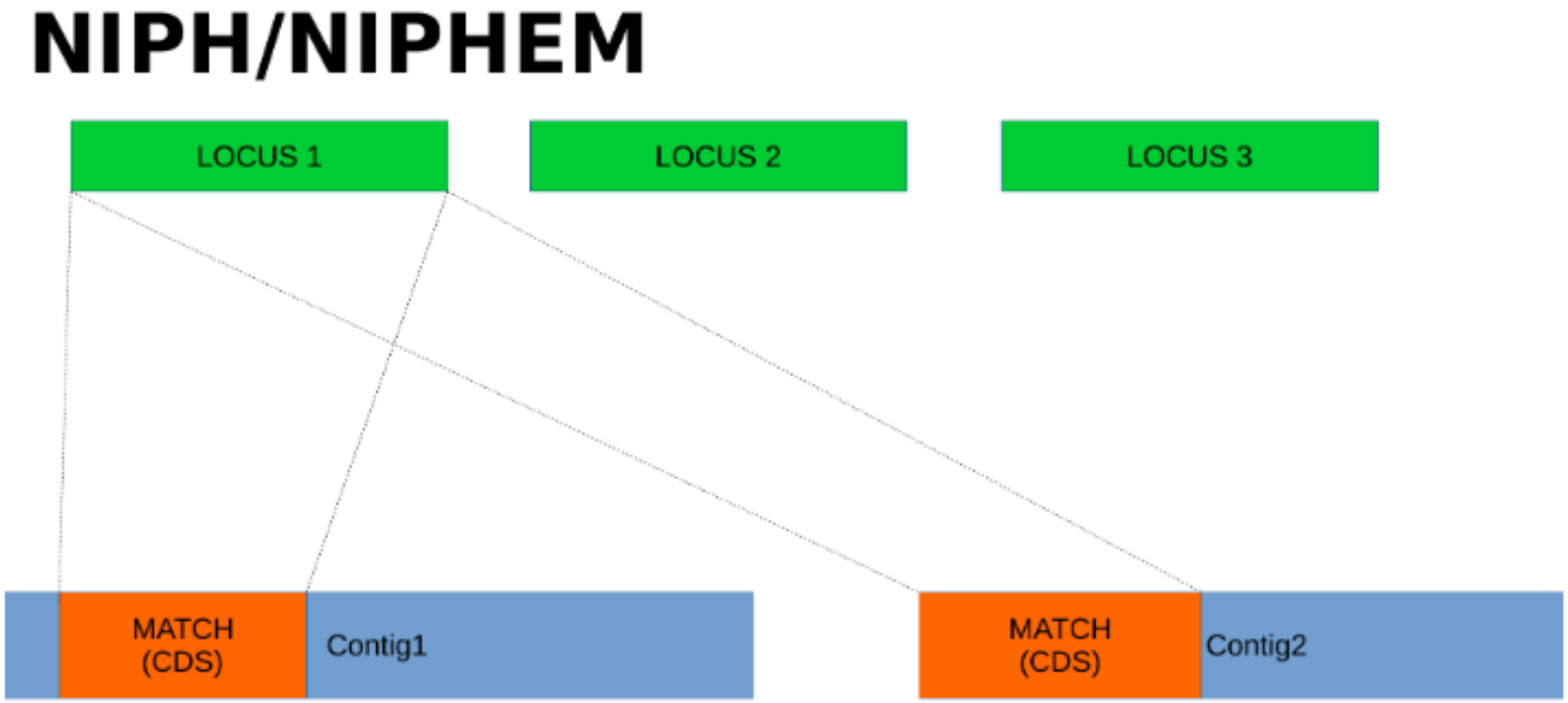
PLOT



Chewbbaca

- ***NIPH*** - non-informative paralogous hit (see image below). When ≥ 2 CDSs in the query genome match one locus in the schema with a BSR > 0.6, that locus is classified as *NIPH*. This suggests that such locus can have paralogous (or orthologous) loci in the query genome and should be removed from the analysis due to the potential uncertainty in allele assignment (for example, due to the presence of multiple copies of the same mobile genetic element (MGE) or as a consequence of gene duplication followed by pseudogenization). A high number of NIPH may also indicate a poorly assembled genome due to a high number of smaller contigs which result in partial CDS predictions. These partial CDSs my contain conserved domains that match multiple loci. This classification takes precedence over *PLOT* classification.
- ***NIPHEM*** - similar to *NIPH* classification (NIPH with exact match), but specifically referring to exact matches. Whenever > 1 CDS matches different alleles of the same locus with 100% DNA similarity during the first DNA sequence comparison, the *NIPHEM* tag is attributed. The loci classified as *NIPHEM* are included in *NIPH* statistics file column, but represent a distinct classification in the MLST profile.

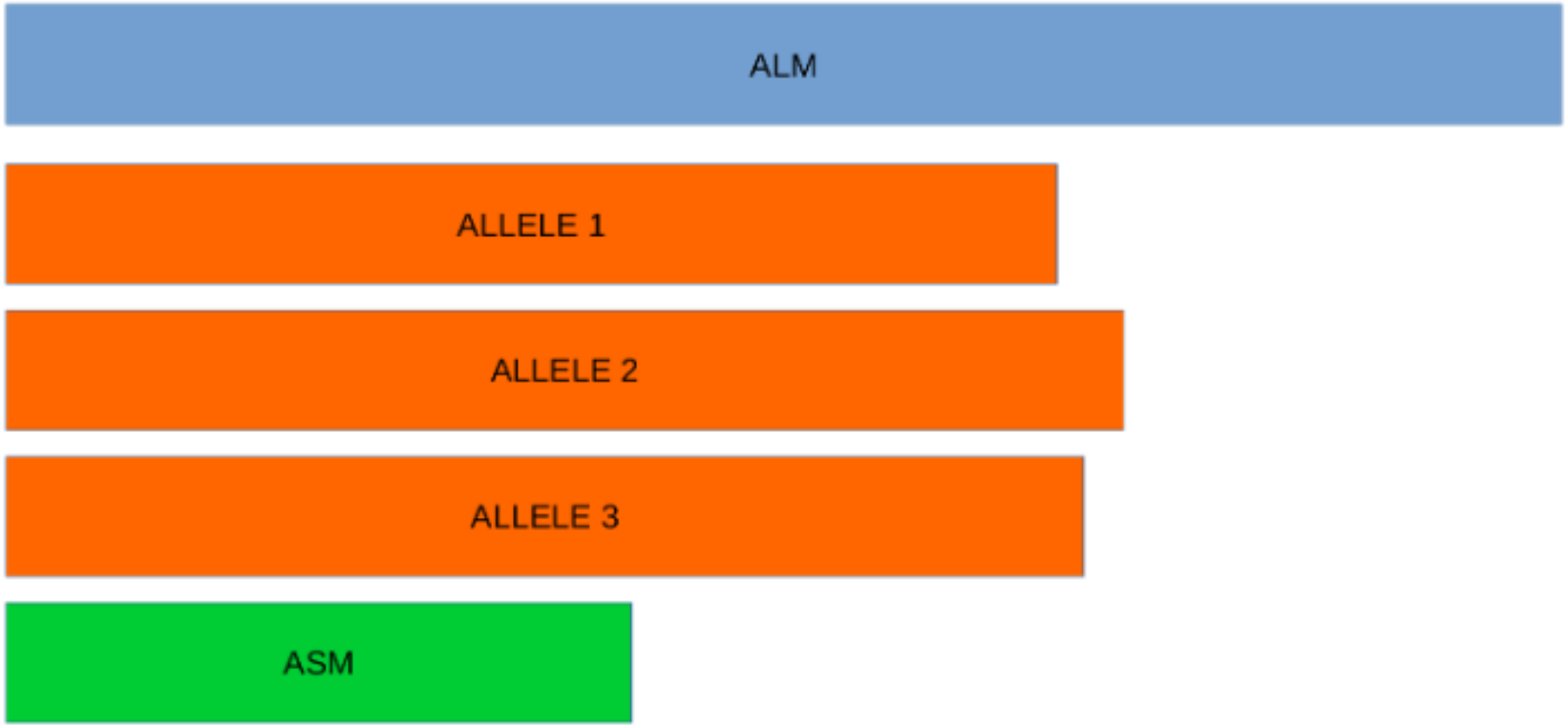
ChewBBACA


NIPH/NIPHEM
LOCUS 1
LOCUS 2
LOCUS 3
MATCH (CDS)
Contig1
MATCH (CDS)
Contig2

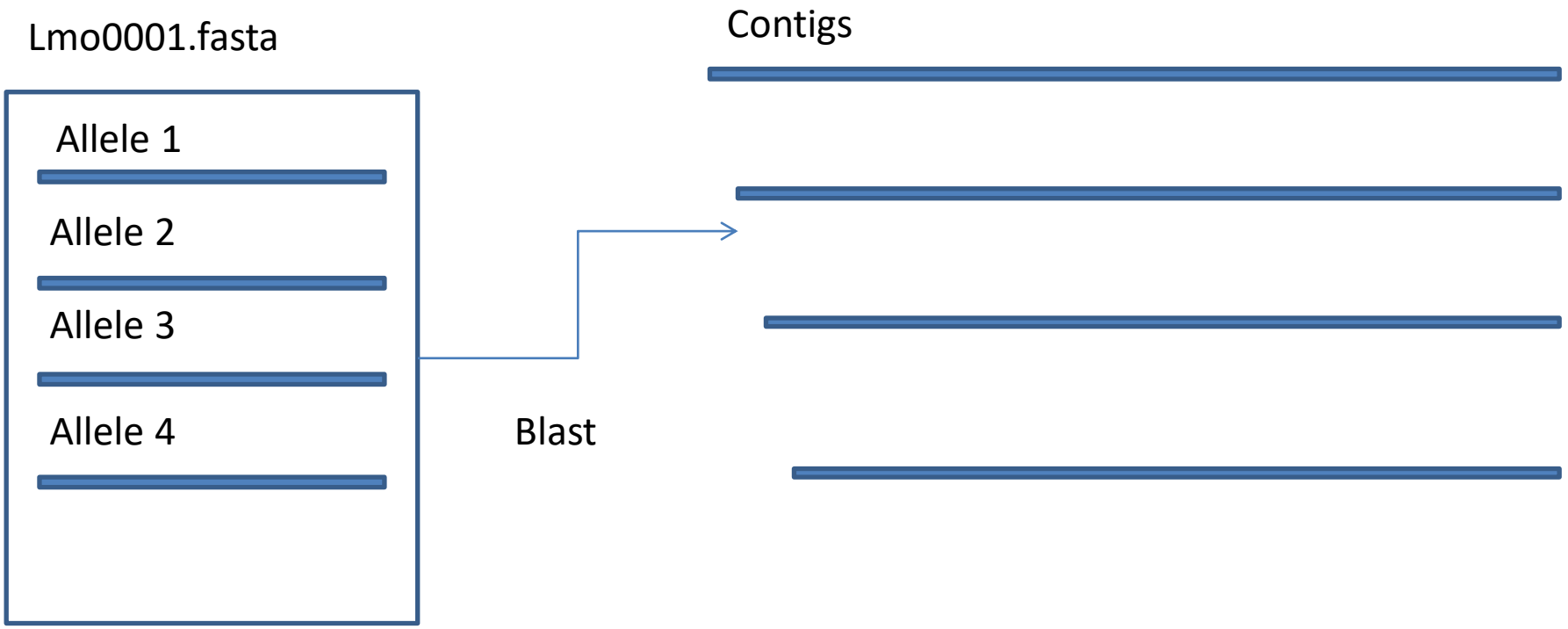
ChewBBACA

- ***ALM*** - alleles 20% larger than length mode of the distribution of the matched loci (CDS length > (locus length mode + locus length mode * 0.2)) (see image below). This determination is based on the currently identified set of alleles for a given locus.
- ***ASM*** - similar to *ALM* but for alleles 20% smaller than length mode distribution of the matched loci (CDS length < (locus length mode - locus length mode * 0.2)).

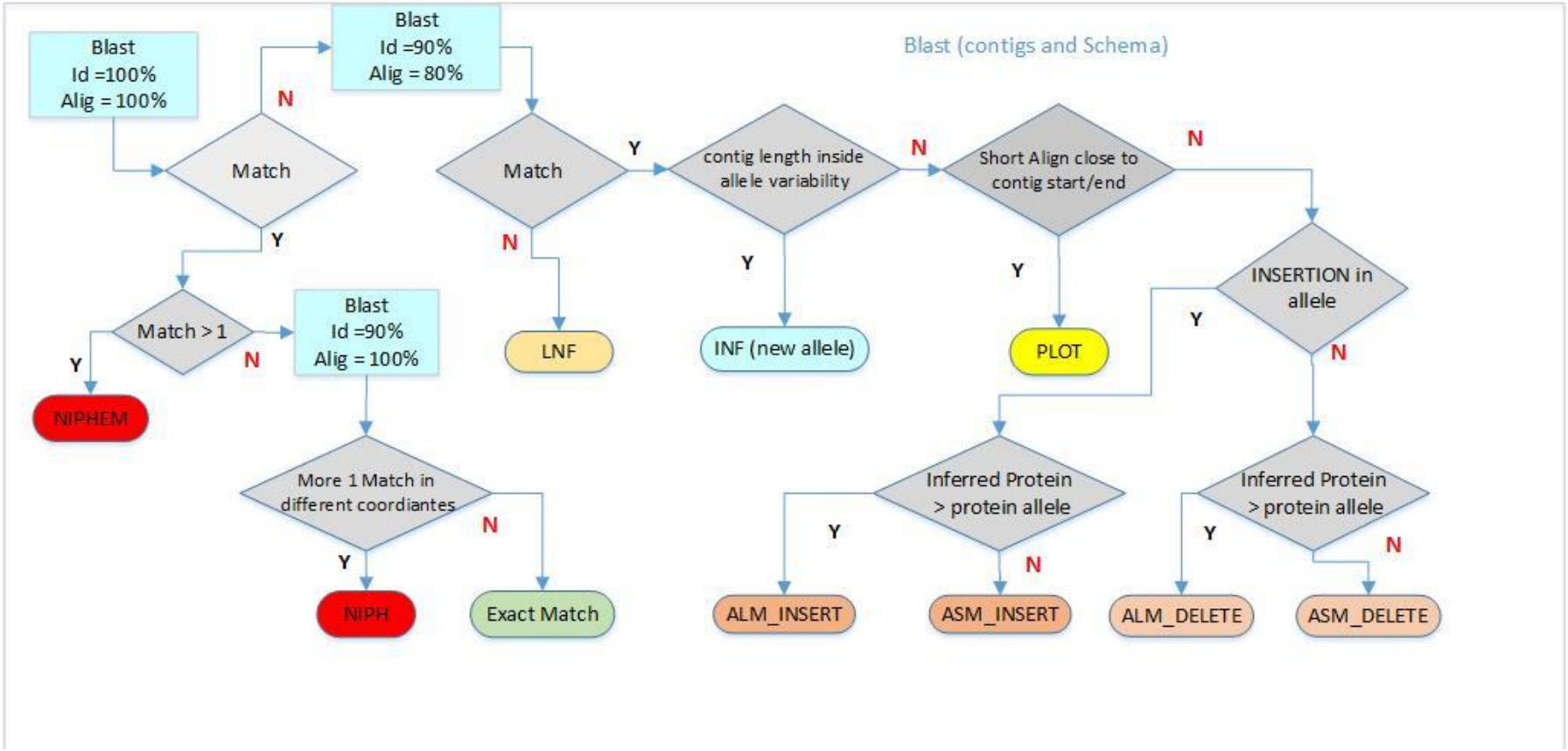
Chewbbaca



Taranis

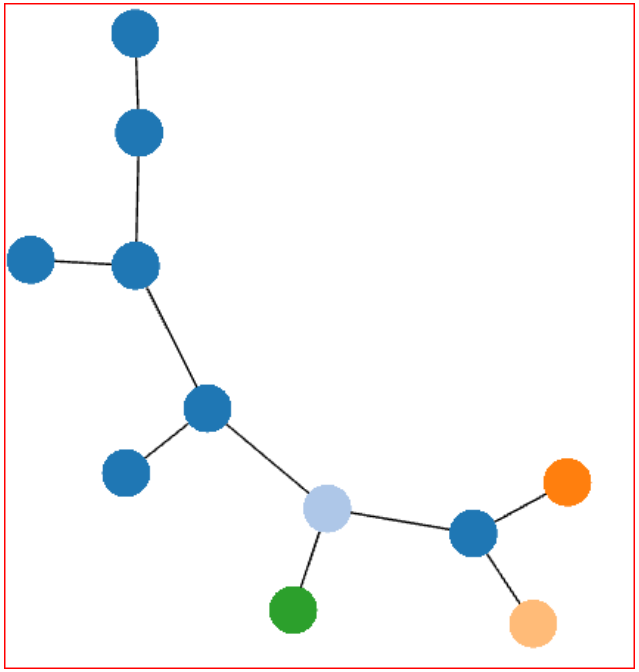

Lmo0001.fasta
Allele 1
Allele 2
Allele 3
Allele 4
Blast
Contigs

Taranis


Blast (contigs and Schema)
Blast Id =100% Alig = 100%
Blast Id =90% Alig = 80%
Match
N
Y
Match
Y
N
contig length inside allele variability
N
Short Align close to contig start/end
N
Y
Y
Y
INSERTION in allele
Y
N
Match > 1
Y
N
Blast Id =90% Alig = 100%
LNF
INF (new allele)
PLOT
NIPHEM
More 1 Match in different coordiantes
N
Y
Inferred Protein > protein allele
Y
N
Inferred Protein > protein allele
Y
N
NIPH
Exact Match
ALM_INSERT
ASM_INSERT
ALM_DELETE
ASM_DELETE

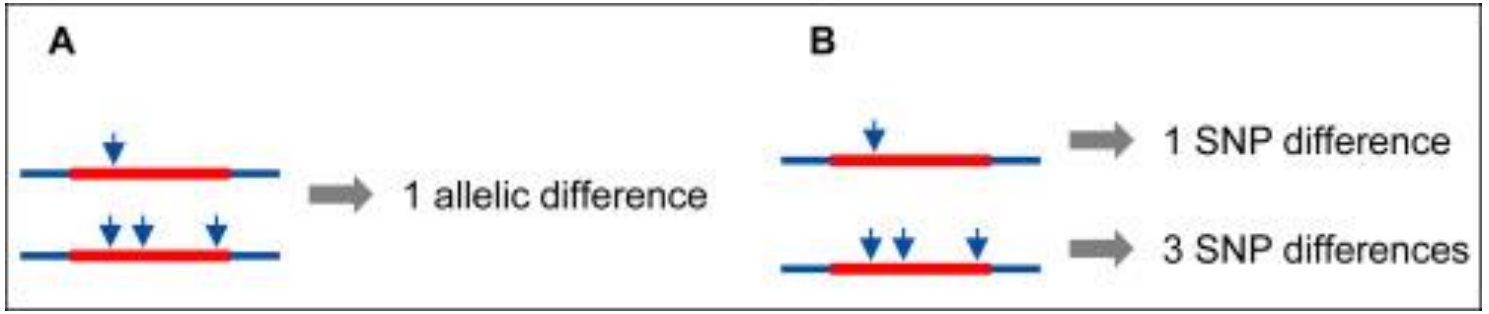
Minimum Spanning tree - Phyloviz

Sample	Gene1	Gene2	Gene3
Sample1	1	45	5
Sample2	1	24	5
Sample3	1	32	6
Sample4	1	12	6



Gene-by-gene vs SNP-based analysis

- In gene-by-gene approaches we account for allelic changes instead of SNP changes.



Lüth et al. Trends in food science and technology. 2018.

Gene-by-gene vs SNP-based analysis

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]

cg, core genome; MLST, multilocus sequence typing; SNP, single nucleotide polymorphism; wg, whole genome.
[a] Data often represent single studies that can be used to begin formulation of species-specific interpretation criteria. Thus, these data should be coupled with newly published similar studies to ensure that resulting values are not atypical and can be generally applied.
[b] Proposed wg/cgMLST schemes are available online (http://www.cgmlst.org/ncs, http://www.applied-maths.com/applications/wgmlst, https://enterobase.warwick.ac.uk/) but as yet have not been epidemiologically validated.

Schürch et al. Clinical Microbiology and infection. 2018

Thanks for your attention!