

Session 4.2 – Gene-by-Gene analysis

Sara Monzón Fernández

BU-ISCIII

Unidades Comunes Científico Técnicas – SGSAFI-ISCIII

24-28 Octubre 2022, 4ª 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

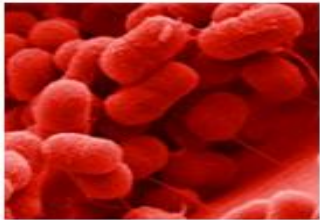


cg/wgMLST schemes available

- 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.
- Macroccoccus canis*
- Macroccoccus 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 damsela*
- 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

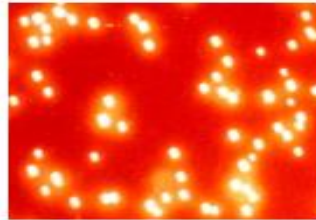
Acinetobacter baumannii



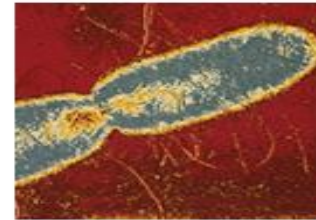
Bifidobacterium



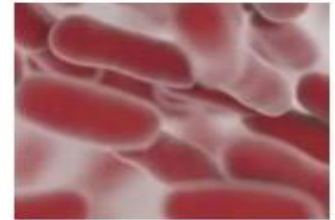
Bordetella pertussis



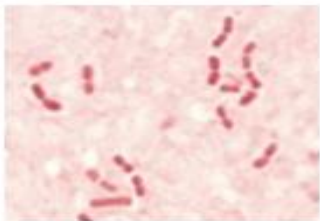
Escherichia coli



Elizabethkingia



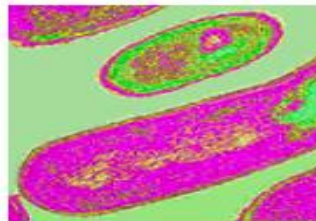
Kingella kingae



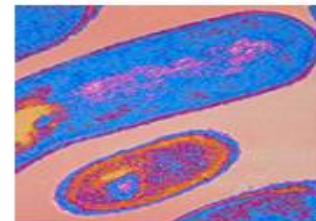
Klebsiella pneumoniae



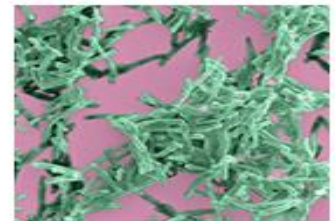
Lactobacillus casei



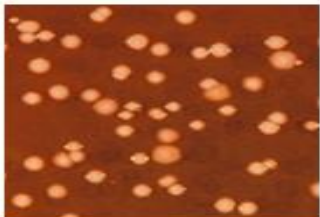
Listeria monocytogenes



Mycobacterium abscessus



Pantoea agglomerans



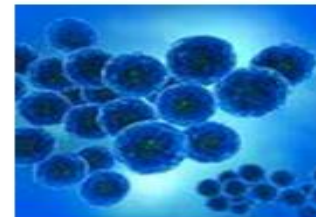
Plesiomonas shigelloides



Propionibacterium freudenreichii



Staphylococcus lugdunensis



Streptococcus thermophilus



cg/wgMLST schemes available

Salmonella
Strains:184198

Assembled

- Legacy:7229
- From NGS:176969
- In Progress:857

Schemes

- rMLST:175680
- Achtman 7 Gene MLST:182851
- cgMLST V2:175008
- wgMLST:171995
- CRISPR:51158

Database Home →

Escherichia/Shigella
Strains:92086

Assembled

- Legacy:9811
- From NGS:82275
- In Progress:30

Schemes

- wgMLST:80922
- Achtman 7 Gene MLST:91747
- rMLST:81928
- cgMLST V1:81665

Database Home →

Clostridioides
Strains:7215

Assembled

- From NGS:7215
- In Progress:1

Schemes

- Griffiths 7 Gene:7208
- cgMLST V1:7205
- rMLST:7206
- wgMLST:7202

Database Home →

Vibrio
Strains:6840

Assembled

- From NGS:6840
- In Progress:3

Schemes

- rMLST:6822

Database Home →

Yersinia
Strains:3576

Assembled

- Legacy:1165
- From NGS:2411
- In Progress:0

Schemes

- Achtman 7 Gene:3207
- McNally 7 Gene:2773
- cgMLST V1:2411
- rMLST:2408
- wgMLST:2410

Database Home →

Moraxella
Strains:557

Assembled

- Legacy:420
- From NGS:137
- In Progress:0

Schemes

- Achtman 7 Gene:559
- rMLST:137

Database Home →

Helicobacter
Strains:535

Assembled

- From NGS:535
- In Progress:0

Schemes

- rMLST:531

Database Home →

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
```

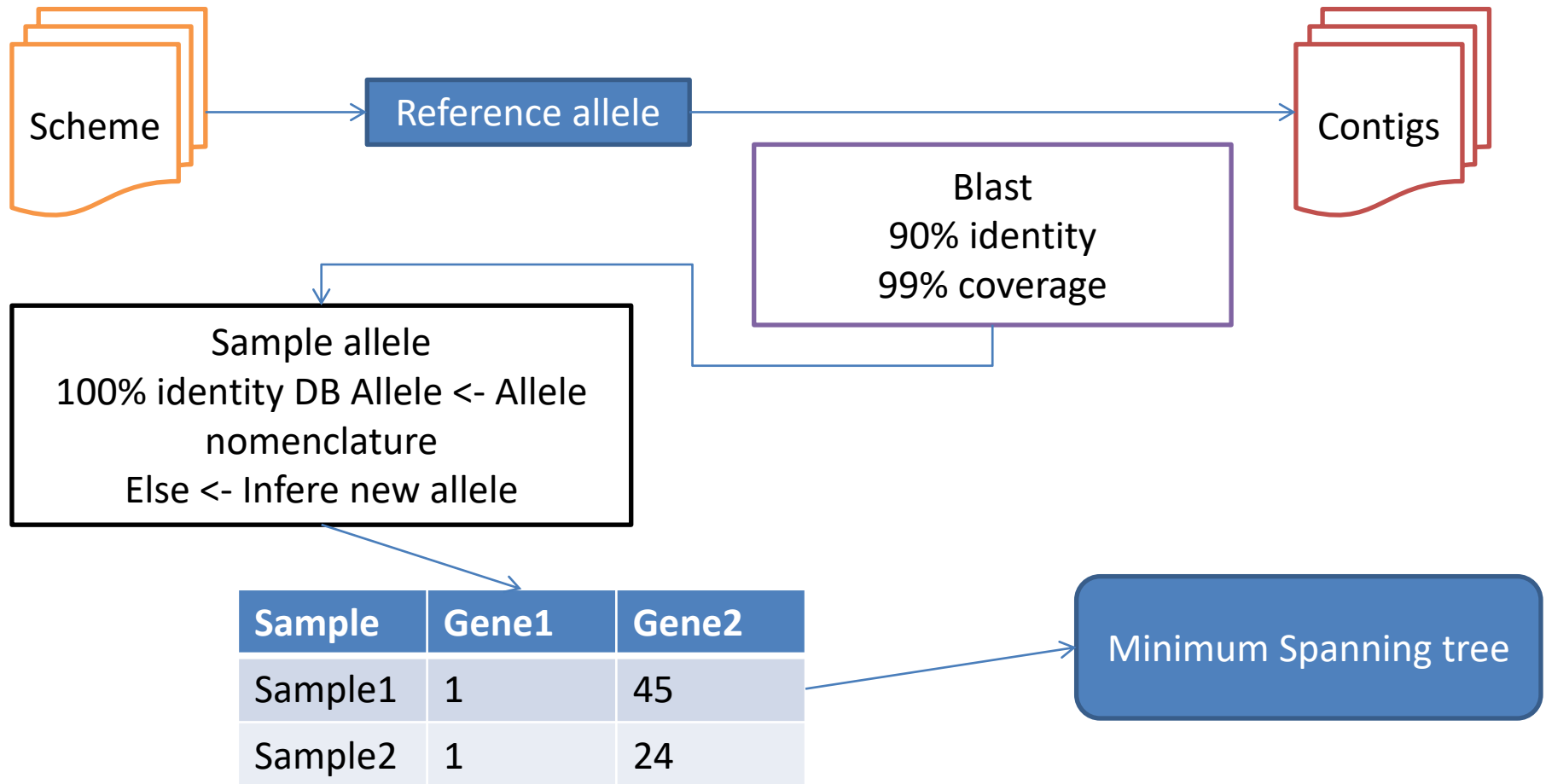
Allele calling

- 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)

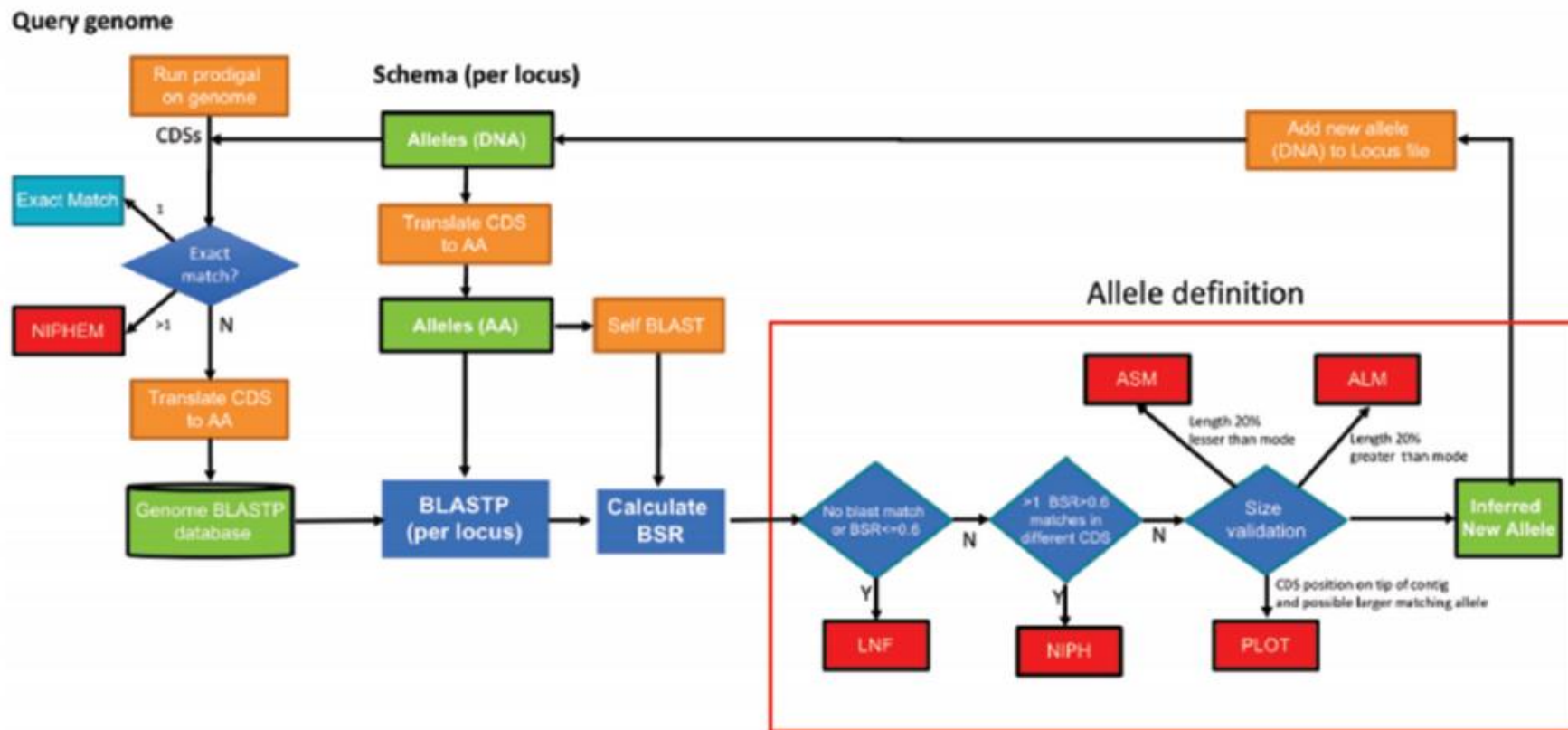
Software

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

SeqSphere



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

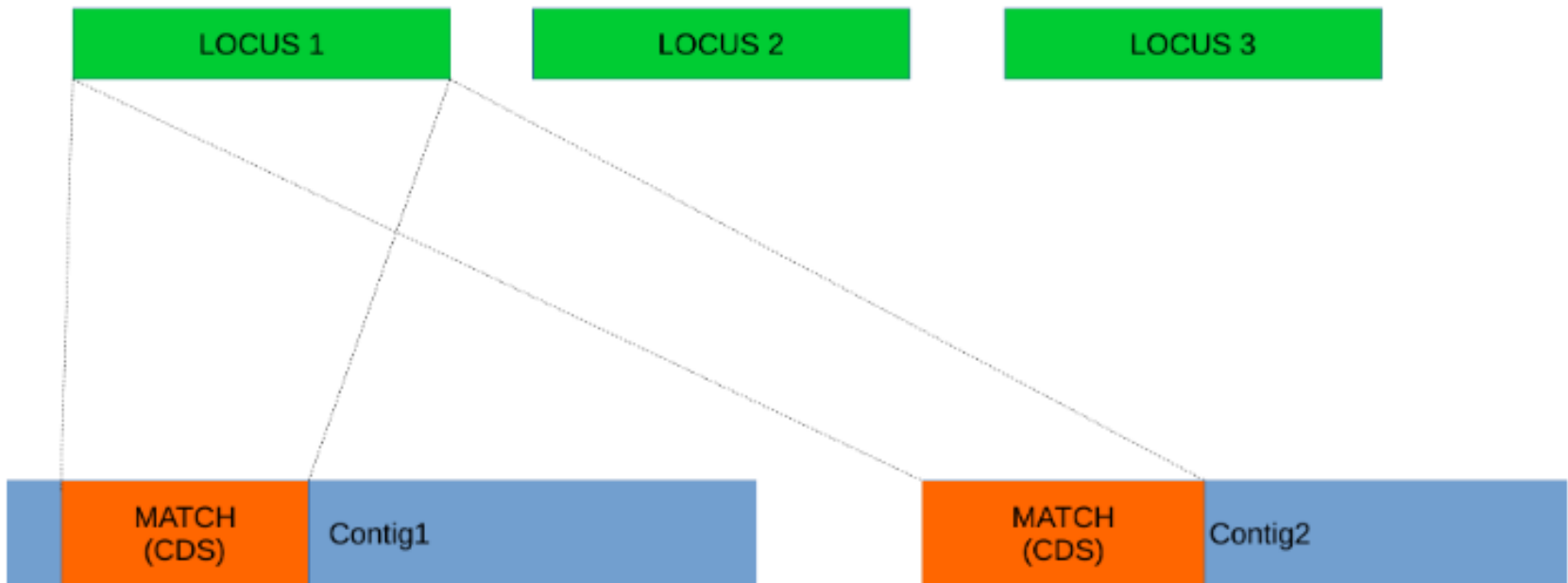


Chewbbaca

- ***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 may 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

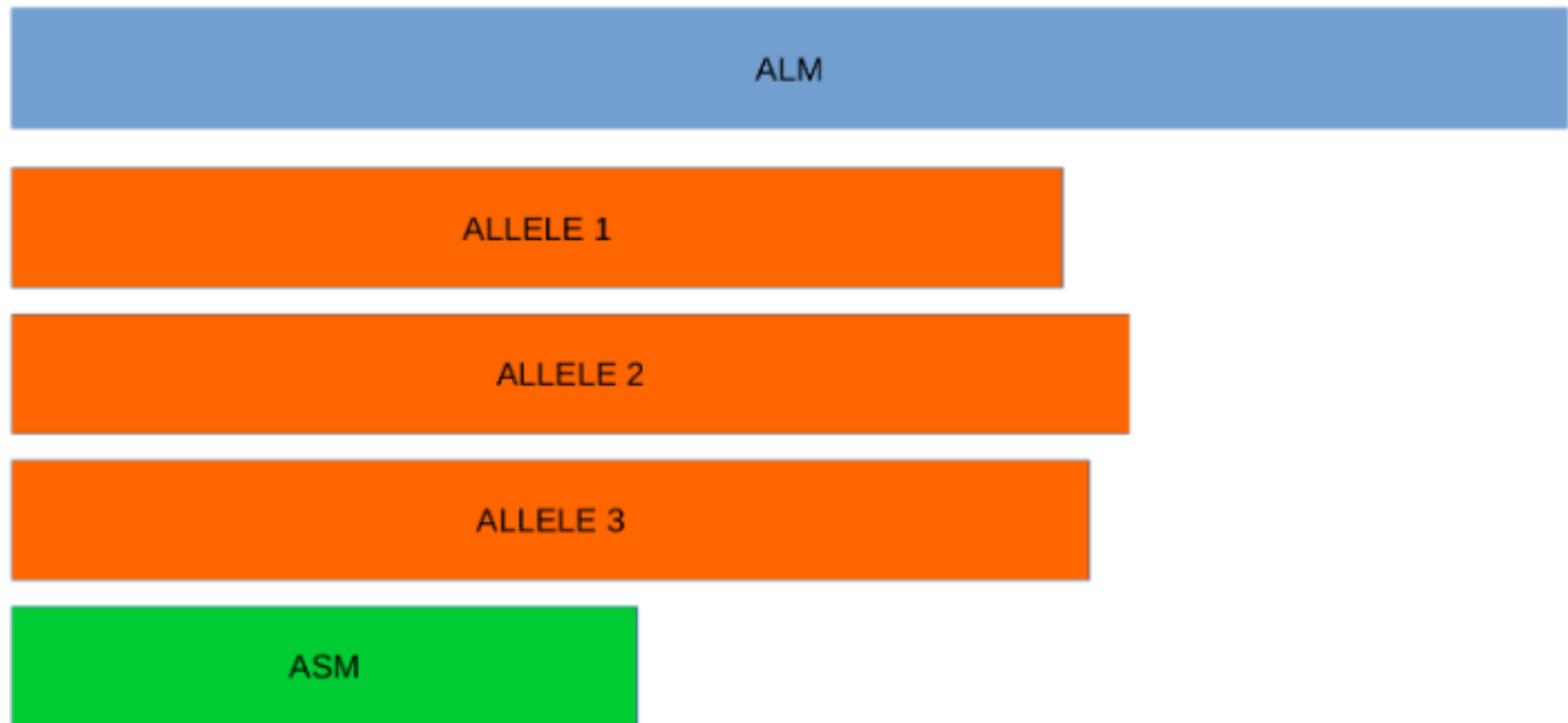
NIPH/NIPHEM



ChewBBACA

- **ALM** - alleles 20% larger than length mode of the distribution of the matched loci ($\text{CDS length} > (\text{locus length mode} + \text{locus length mode} * 0.2)$) (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 ($\text{CDS length} < (\text{locus length mode} - \text{locus length mode} * 0.2)$).

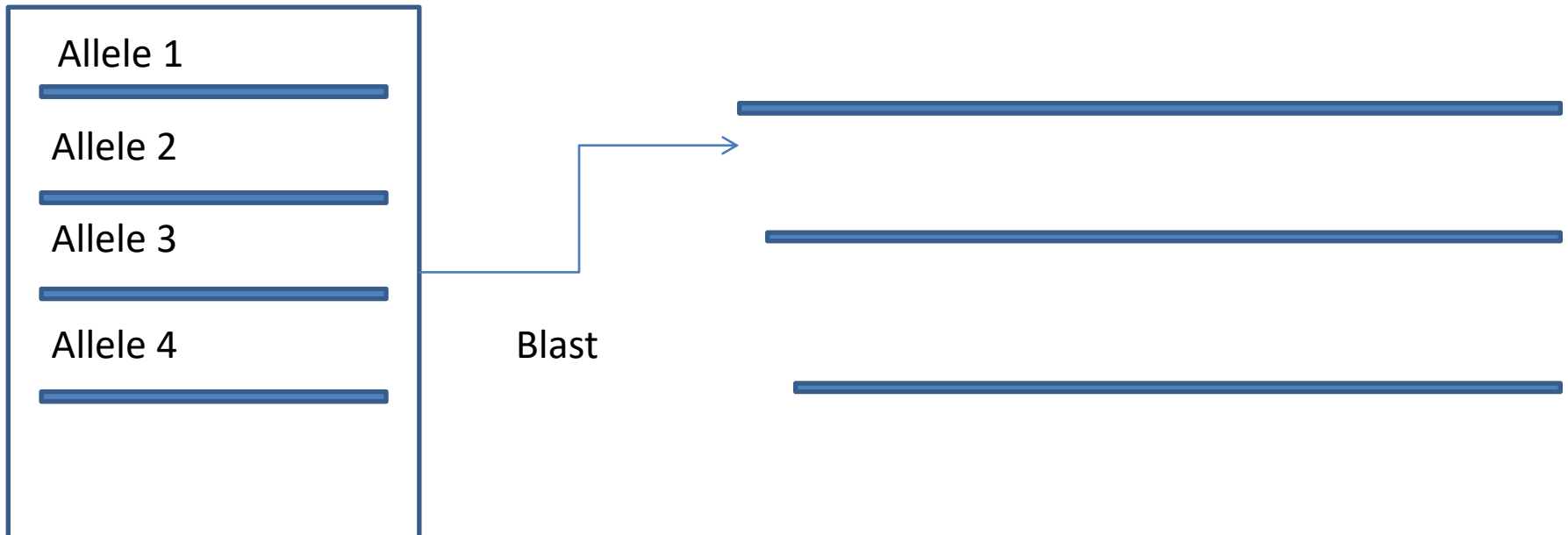
Chewbbaca



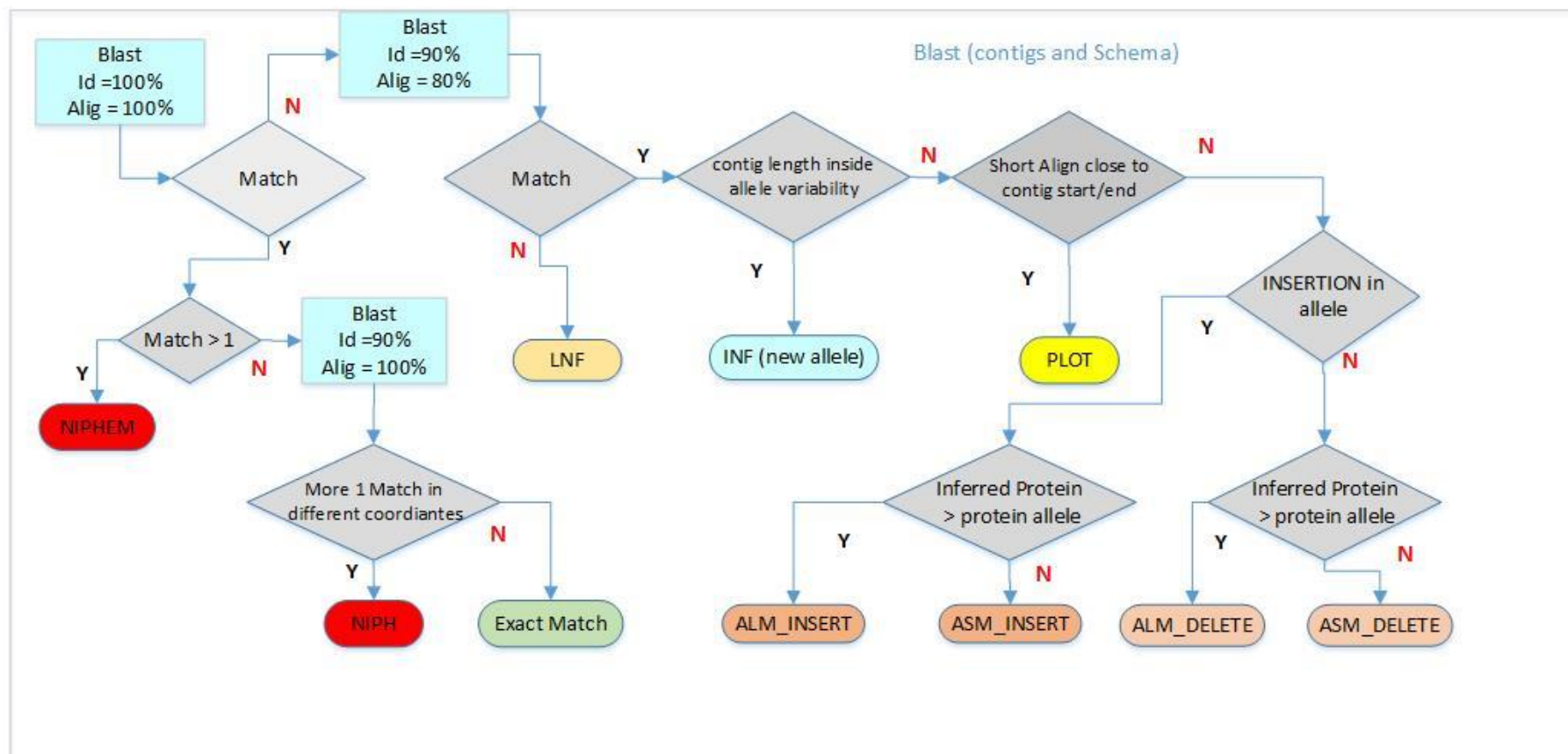
Taranis

Lmo0001.fasta

Contigs

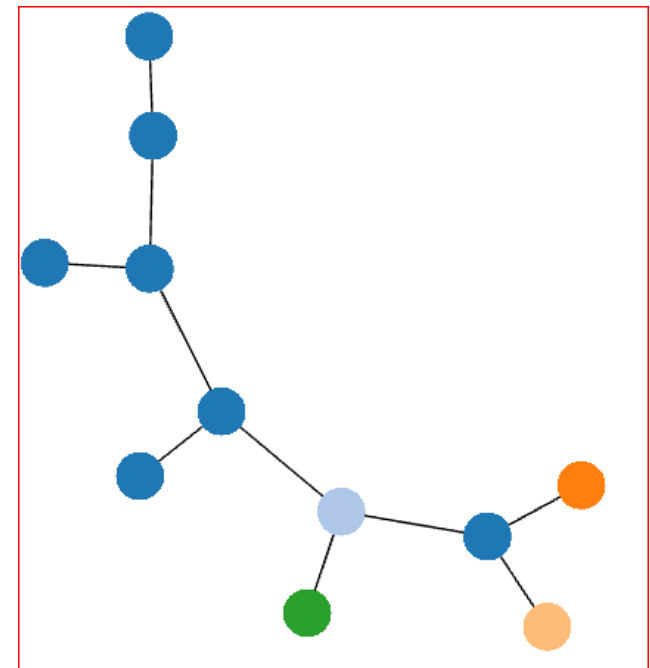


Taranis



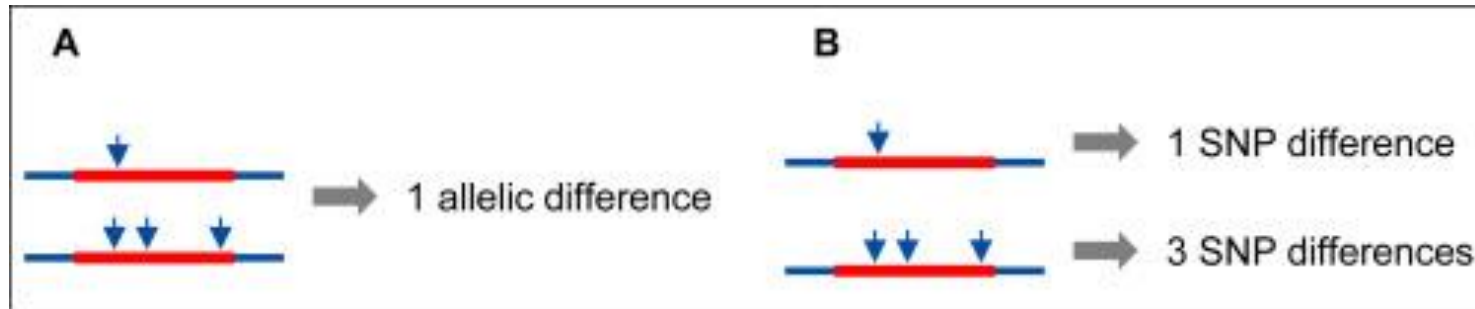
Minimum Spanning tree - Phyloviz

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



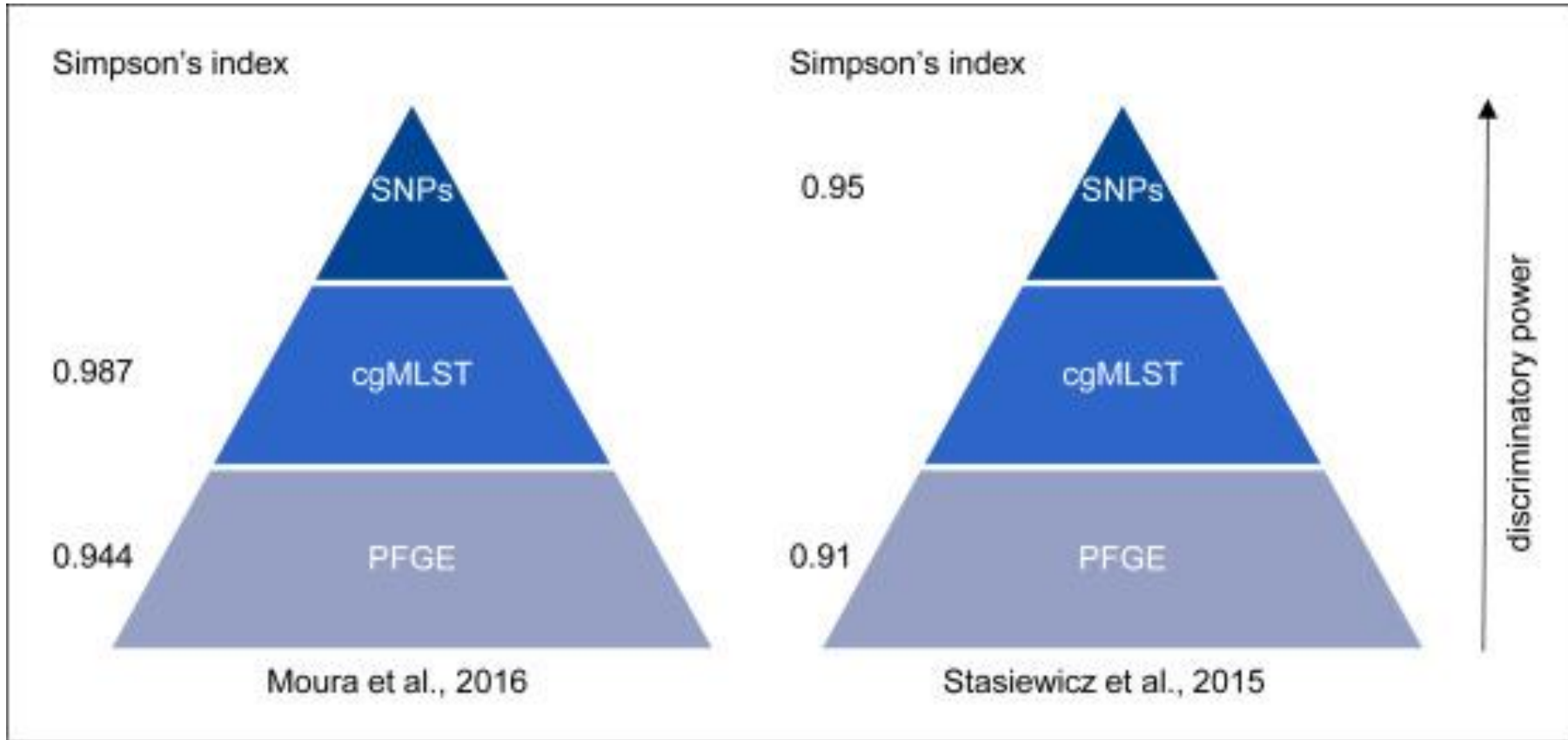
Gene-by-gene vs SNP-based analysis

- 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



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	
<i>Acinetobacter baumannii</i>	≤8	≤3 [25,26]
<i>Brucella</i> spp.	Epidemiologic validation in progress ^b	http://www.applied-maths.com/applications/wgmlst
<i>Campylobacter coli</i> , <i>C. jejuni</i>	≤14	≤15 [27,28]
<i>Cronobacter</i> spp.	Epidemiologic validation in progress ^b	http://www.applied-maths.com/applications/wgmlst
<i>Clostridium difficile</i>	Epidemiologic validation in progress ^b	≤4 [29], http://www.cgmlst.org/ncs , http://www.applied-maths.com/applications/wgmlst
<i>Enterococcus faecium</i>	≤20	≤16 [30]
<i>Enterococcus raffinosus</i>	Epidemiologic validation in progress ^b	http://www.applied-maths.com/applications/wgmlst
<i>Escherichia coli</i>	≤10	≤10 [31,32], https://enterobase.warwick.ac.uk/
<i>Francisella tularensis</i>	≤1	≤2 [33,34]
<i>Klebsiella oxytoca</i>	Epidemiologic validation in progress ^b	http://www.applied-maths.com/applications/wgmlst
<i>Klebsiella pneumonia</i>	≤10	≤18 [35,36]
<i>Legionella pneumophila</i>	≤4	≤15 [37]
<i>Listeria monocytogenes</i>	≤10	≤3 [38,39]
<i>Mycobacterium abscessus</i>		≤30 [40]
<i>Mycobacterium tuberculosis</i>	≤12	≤12 [41]
<i>Neisseria gonorrhoeae</i>	Epidemiologic validation in progress ^b	≤14 [42], http://www.applied-maths.com/applications/wgmlst
<i>Neisseria meningitidis</i>	Epidemiologic validation in progress ^b	http://www.cgmlst.org/ncs
<i>Pseudomonas aeruginosa</i>	≤14	≤37 [31,43]
<i>Salmonella dublin</i>	Epidemiologic validation in progress ^b	≤13 [44], https://enterobase.warwick.ac.uk/
<i>Salmonella enterica</i>	Epidemiologic validation in progress ^b	≤4 [45], http://www.cgmlst.org/ncs , http://www.applied-maths.com/applications/wgmlst , https://enterobase.warwick.ac.uk/
<i>Salmonella typhimurium</i>	Epidemiologic validation in progress ^b	≤2 [46], https://enterobase.warwick.ac.uk/
<i>Staphylococcus aureus</i>	≤24	≤15 [47,48]
<i>Streptococcus suis</i>		≤21 [49]
<i>Vibrio parahaemolyticus</i>	≤10	[50]
<i>Yersinia</i> 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!
