Session 4.1 – Tipificación basada en perfil alélico o gene-by-gene

Isabel Cuesta

BU-ISCIII

Unidades Comunes Científico Técnicas – SGSAFI-ISCIII

04-15 Noviembre 2019, 2ª Edición
Programa Formación Continua, ISCIII

Index

- Typing resolution
- Concepts: homology, core, accessory and pan-genome
- Schema definition
- e.g. *Listeria monocytogenes*

Typing methods: DNA-based methods

PFGE *Gold standard*

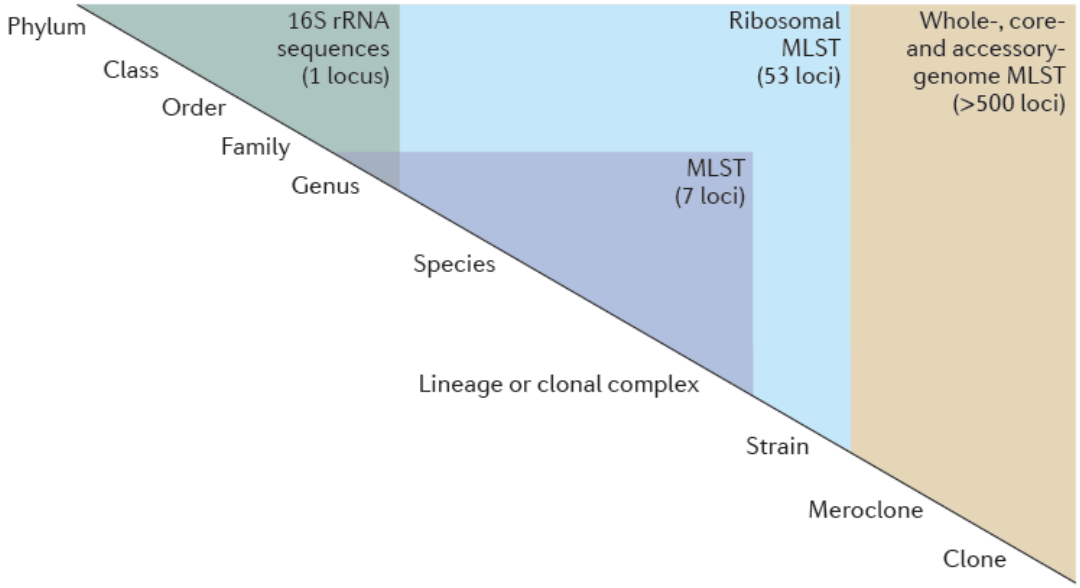
Is a rather time-consuming and labour-intensive technique.

Discriminatory power of PFGE profiles is limited as only nucleotide changes in the restriction enzyme recognition sites are detected.

Relatedness of strains may be over- or underestimated

Epidemiologically unrelated isolates may be assigned to one ‘pseudo’-cluster whereas even highly related strains fall into distinct clusters.

Sequence data for taxonomy and typing



Different levels of sequence information can be associated with different taxonomic levels.

The need for higher-resolution characterization of isolates has led to the development of a wide range of strain-typing methods

Variability between bacterial genomes of the same species

GENOME EVOLUTION

Vertical transfer: is the passing of genetic material by descent

Horizontal transfer: is the movement of genetic material among bacteria that do not necessarily share a mother cell.

- transformation: the uptake of DNA by a cell
- conjugation: transfer facilitated by conjugative elements
- phage-mediated transduction

- Point mutations: SNPs, single nucleotide insertion / deletion.
- Large insertions / deletions
- Genome rearrangements
- Transfer of exogenous DNA
- Plasmids or phages

Concepts

Homology: share a common ancestor, either by descent or recombination. No such thing as “significantly homologous”. A sequence either is or is not homologous.
Infer homology from knowledge of evolutionary relationships and from degrees of similarity between sequences, features or other data.

Orthologues: sequences have common ancestor and have split due to speciation event.

Paralogues: genes arise by gene duplication

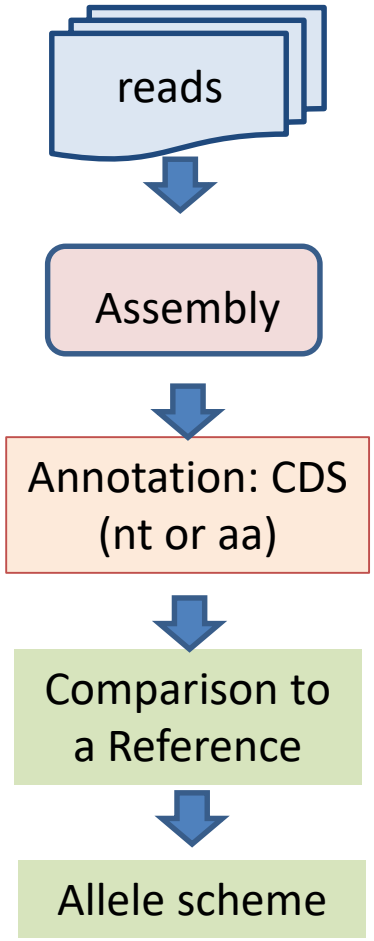
Concepts

Core genome: the number of shared features in a pool of genomes. Shared genes among multiple strains are mostly related to house-keeping genes or central metabolic processes, most of the structural information and main genotypic features. **Orthologues** in all genomes of bacteria belonging to the same taxa

Accessory genome or adaptative genome: includes genes conferring adaptive advantages to the strain in order to survive in a specific environment. In most cases, these factors are linked to antibiotic resistance, virulence, capsular serotype, adaptation, and might reflect the organisms predominant lifestyle.

Pangenome: The term "pan-genome" refers to pan (from Greek παν, whole) and genome (genome) referring to the inclusion of the core and the dispensable genome.

General analytical process for cgMLST / wgMLST



Gene-by-gene: Defining a schema

Sequence-clustering algorithms:

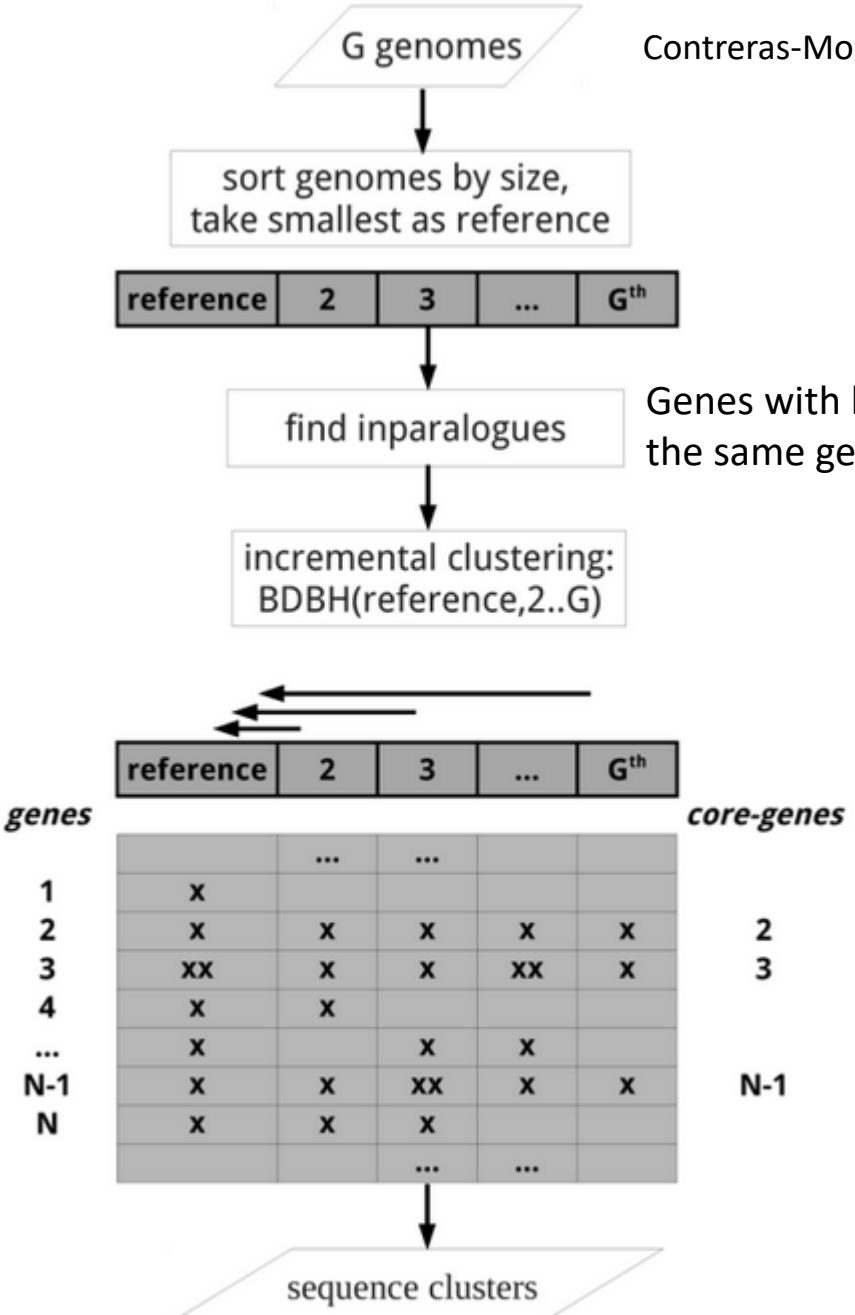
BDBH:
OMCL
COGtriangles

name	option	
BDBH	default	Starting from a reference genome, keep adding genomes stepwise while storing the sequence clusters that result of merging the latest bidirectional best hits, as illustrated in Figure 3.
COGS	-G	Merges triangles of inter-genomic symmetrical best matches, as described in PubMed=20439257. Note that a single sequence might occasionally be included in several COGS clusters with option -x.
OMCL	-M	OrthoMCL v1.4, uses the Markov Cluster Algorithm to group sequences, with inflation (-F) controlling cluster granularity, as described in PubMed=12952885.

BLAST
- % of coverage in the pairwise aligments query/subject
- % of sequence identity in query/subject pairs
- Genome uses as reference genome

BDBH

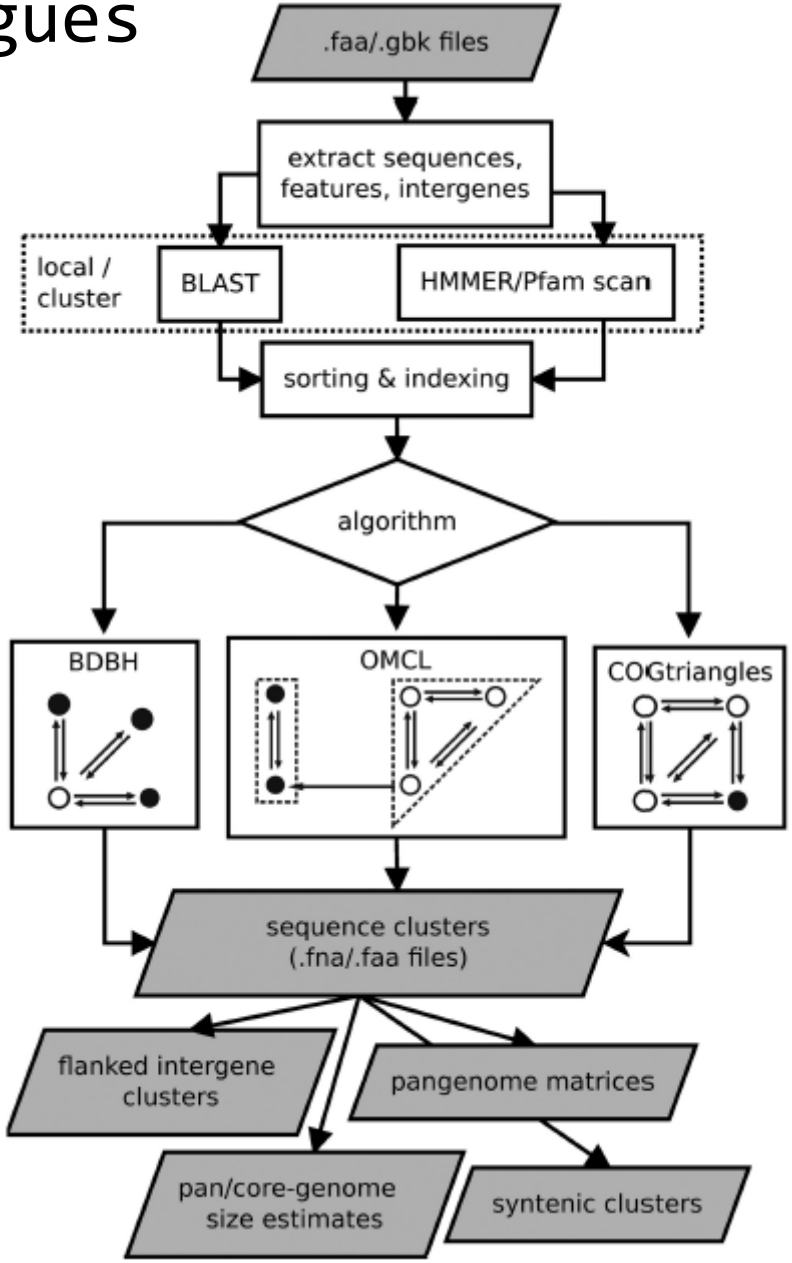
Contreras-Moreira, Appl Envirn Microbiol 2013



Genes with best hits in the same genome

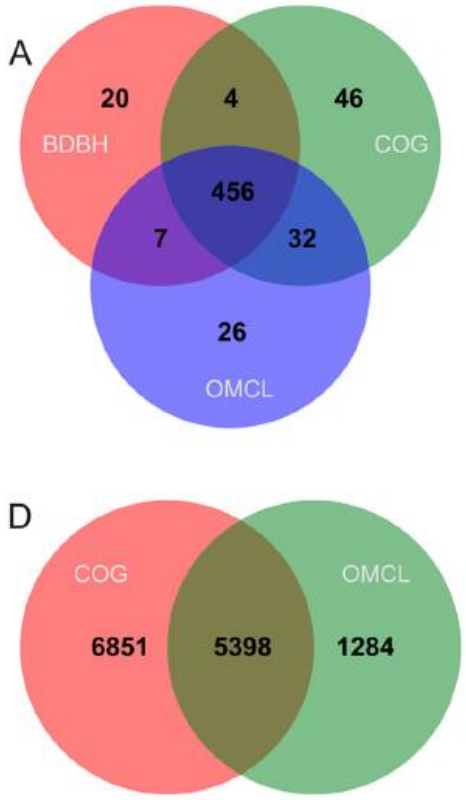
Contreras-Moreira, Appl Envirn Microbiol 2013

Get_Homologues

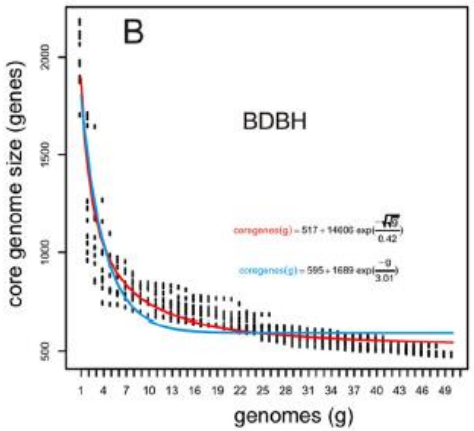


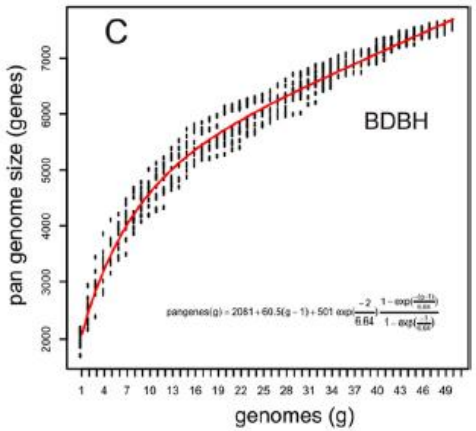
Get_Homologues

Contreras-Moreira, Appl Envirn Microbiol 2013



50 Streptococcus proteomes from 14 species
BLAST: mínimum pairwise aligment covarage of 75%





Reasons why schemas are different

Van Tonder et al., PlosCompBiol 2014

Any collection of isolates is a subset of the entire population for the species of interest, and if the subset of isolates has limited genetic diversity then the number of ‘‘core’’ genes shared by all isolates in that sample will be higher than in a dataset which is genetically more diverse.
More generally, the **size of the core genome is dependent on the size of the data set**, with the core genome decreasing in size as more genomes are added to the analysis

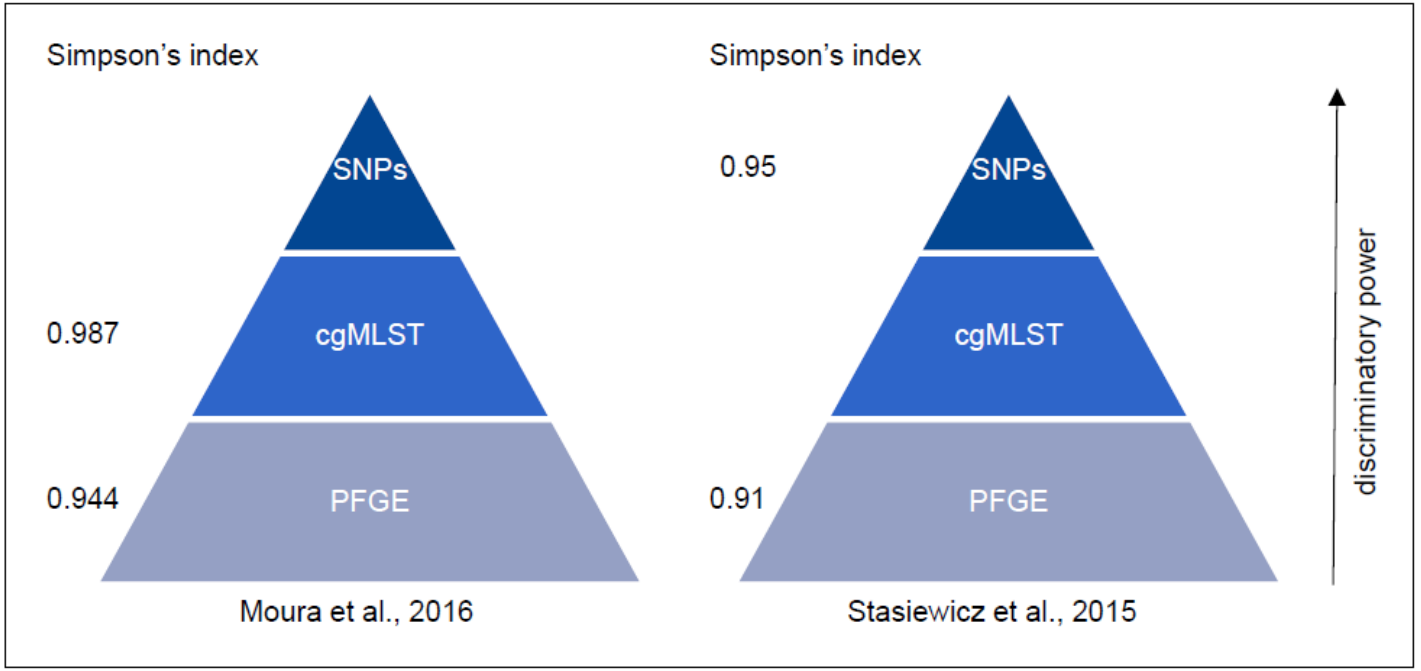
Technical reason:

Incomplete or ‘‘draft’’ genomes. This is acceptable for most studies, but analyses of these genomes may exclude a gene from a list of core genes simply because it contains a sequence gap or is otherwise incomplete at that locus in the assembly of one or a few genomes

BLAST parameters

Discriminatory power of typing methods

Listeria monocytogenes



100 genomes

188 genomes, 3 US states, 2 years

The Simpson's index is used to quantify the probability that two unrelated strains are assigned to different typing groups

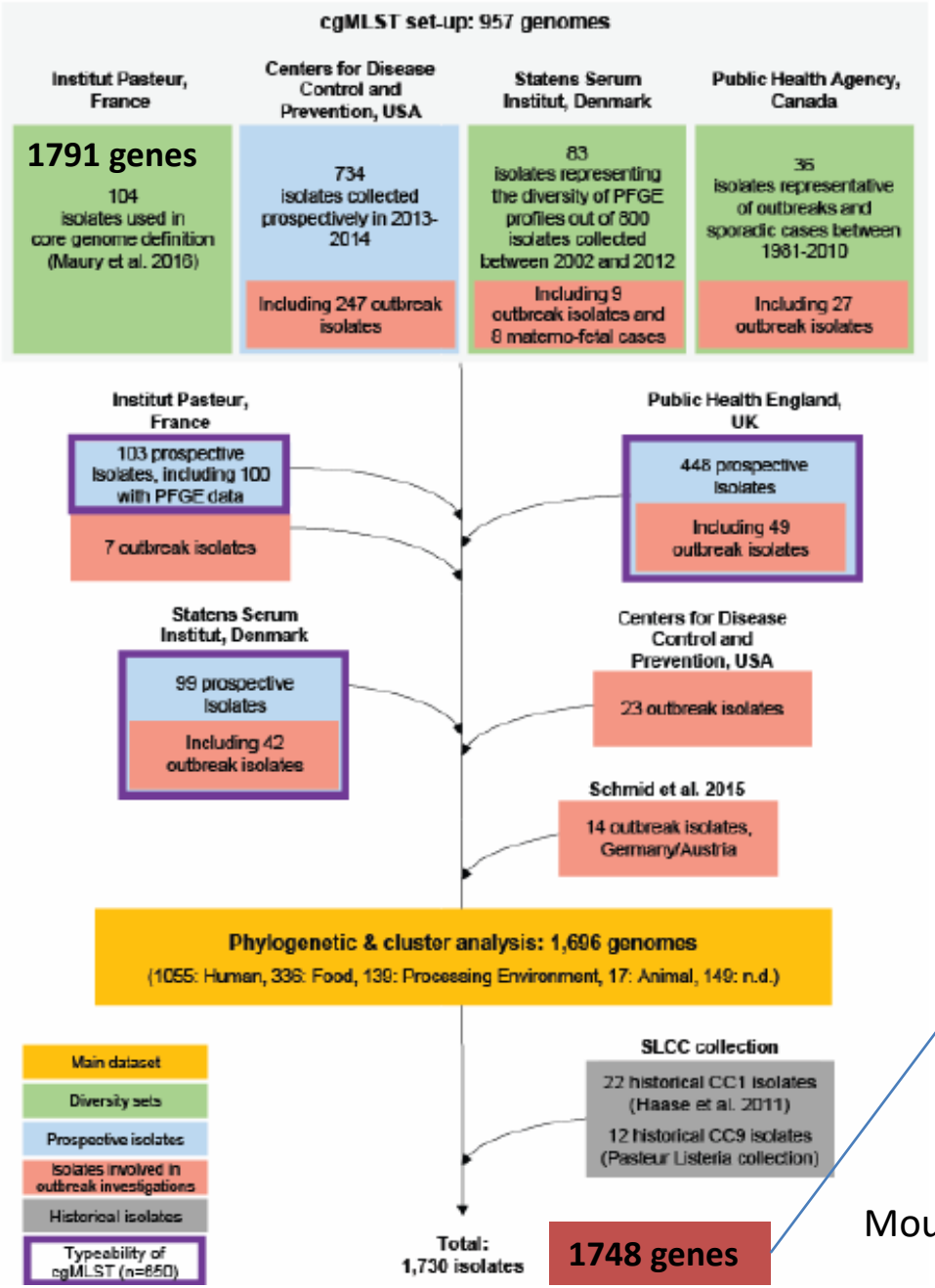
Core-genome schemas for *Listeria monocytogenes*

Ruppitsch et al., 2015
SeqSphere+

Pightling et al, 2015
Bioinformatics pipeline that takes raw sequence reads as input and calculating a core genome profile by comparing it to an expandable database to compile a phylogeny

Moura et al., 2016
1748 loci

All 4 lineages



BLASTN
Minimum nucleotide identity 70%
Alignment length coverage 70%

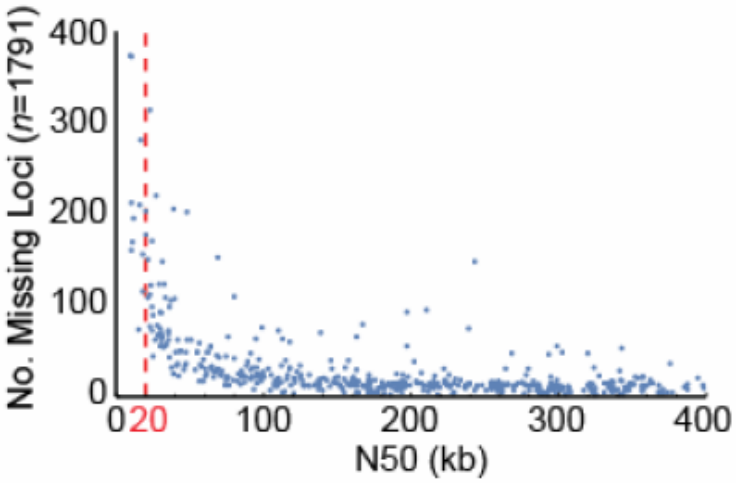
62% of its CDS

Moura et al., Nature Microbiology 2016

Moura core-genome schema for *Listeria monocytogenes*

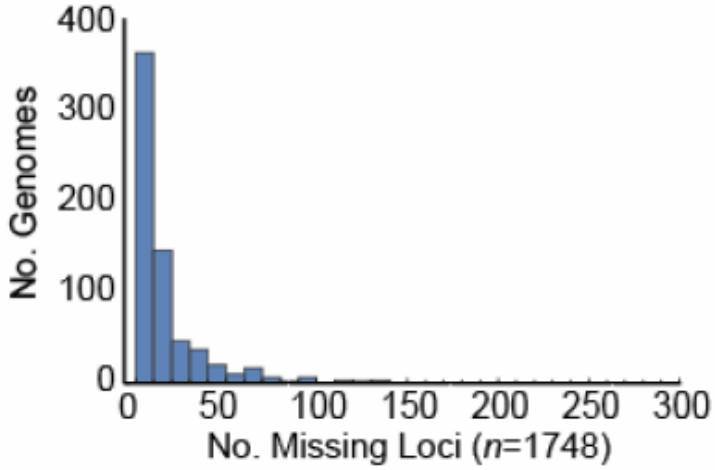
Validation of the cgMLST scheme with a set of 650 genomes

A

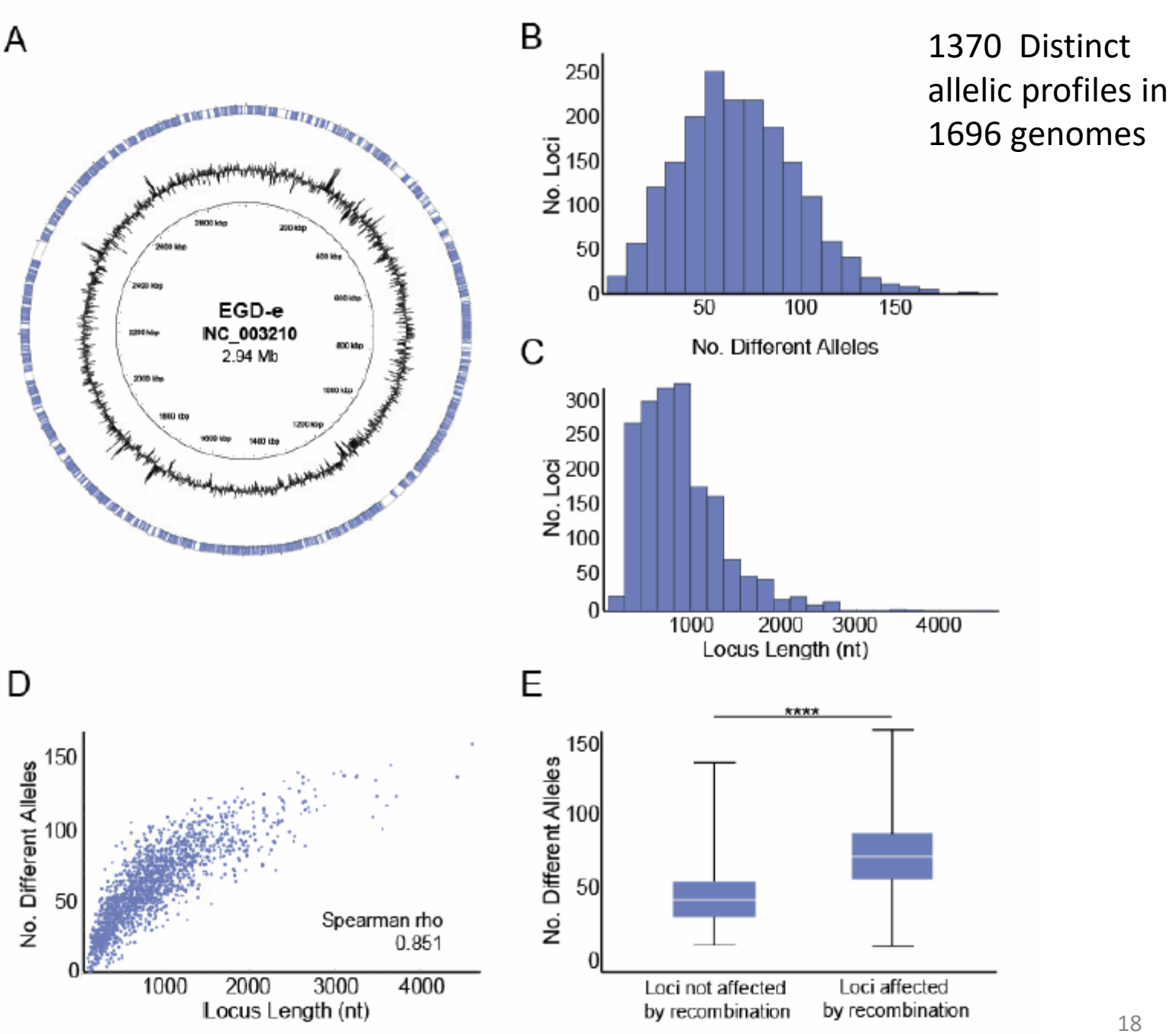


Impact of the N50 assembly size in the number of missing loci. Cut-off N50 of 20kb

B

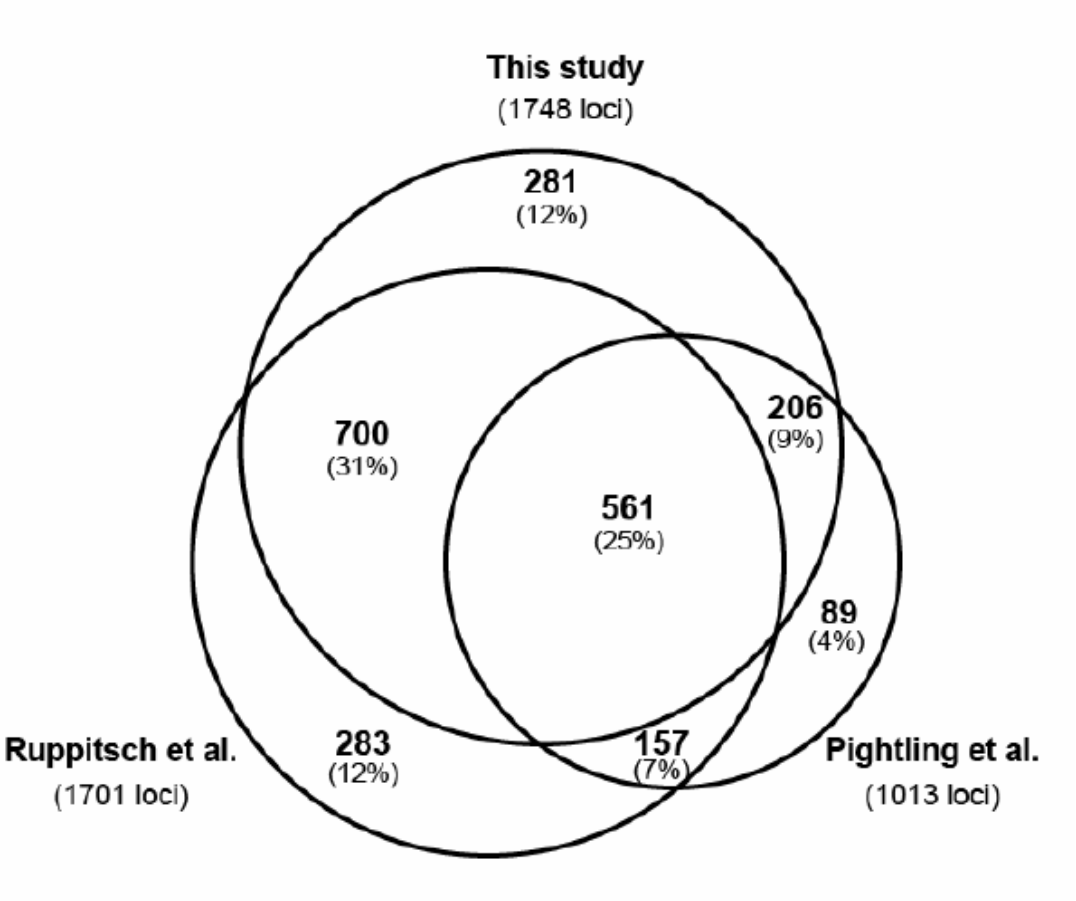


Distribution of the number of missing loci per genome

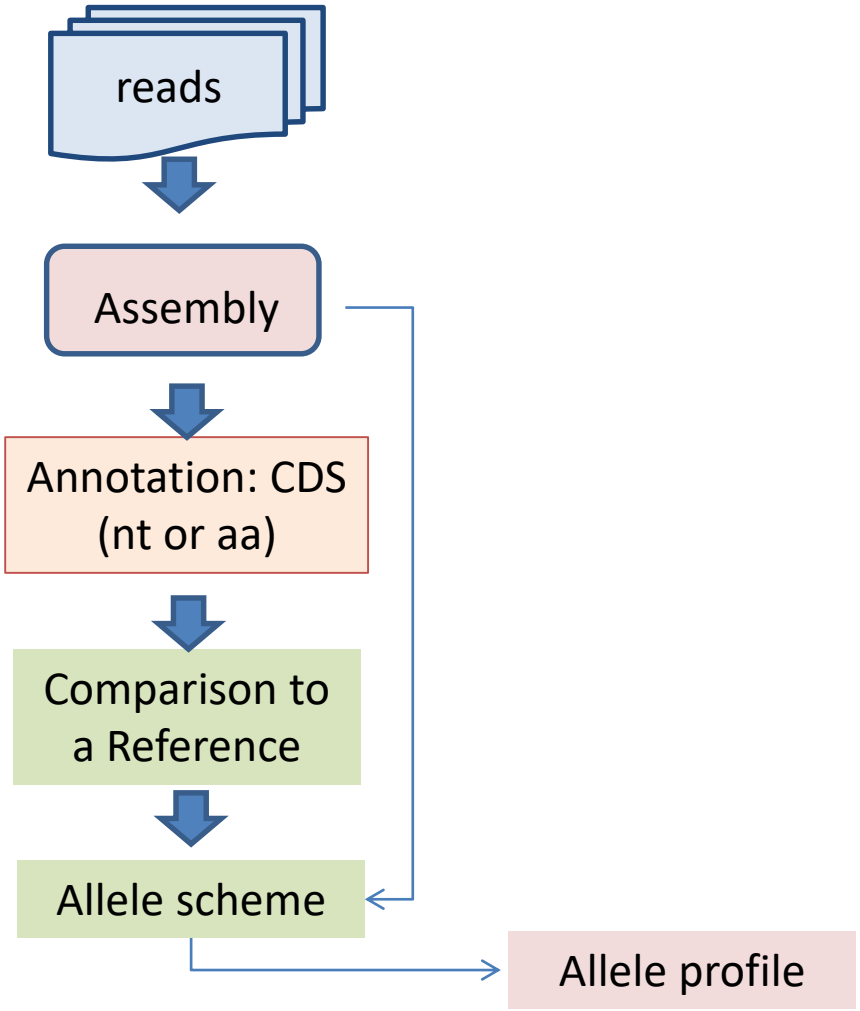


1370 Distinct allelic profiles in 1696 genomes

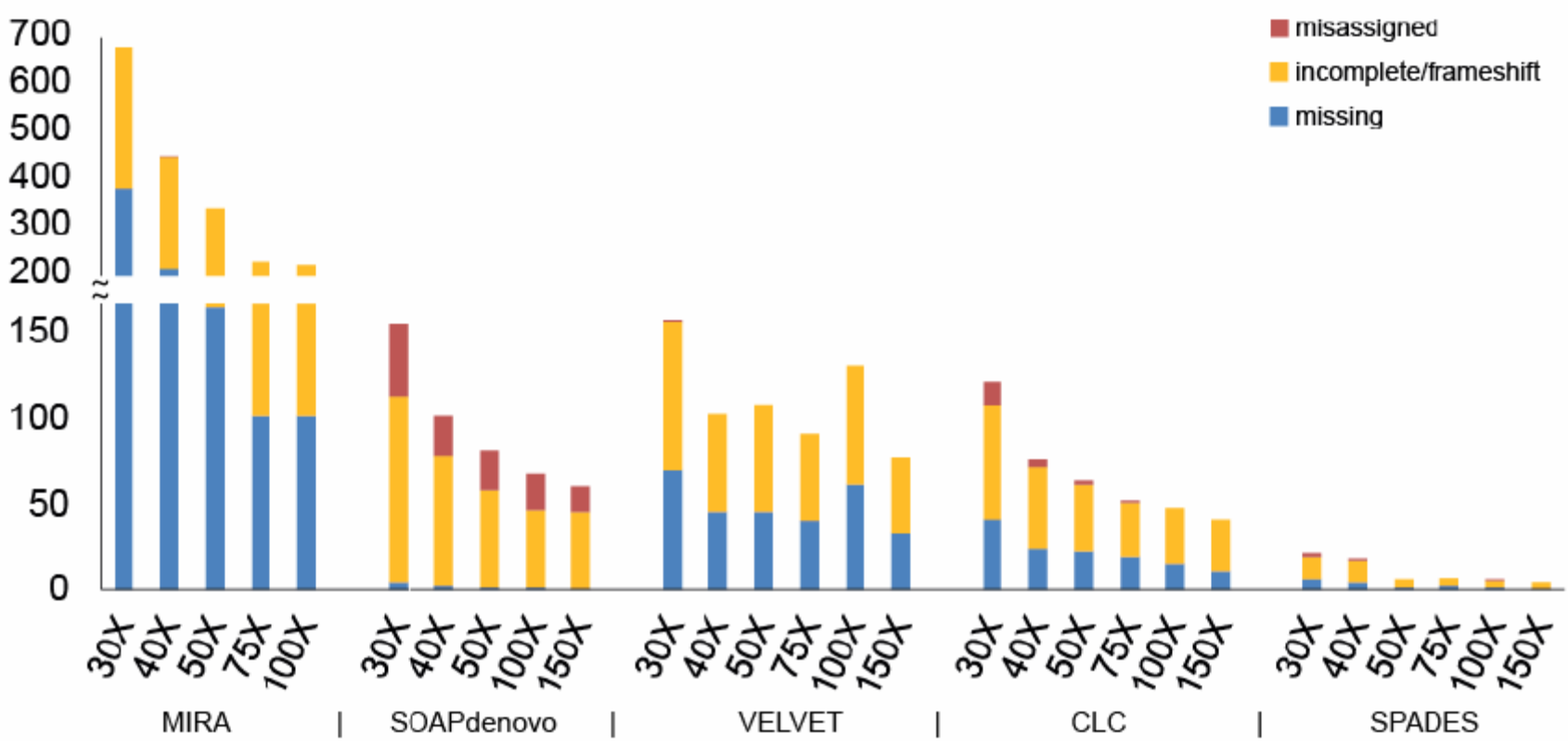
Comparison with other genome-based MLST schemes



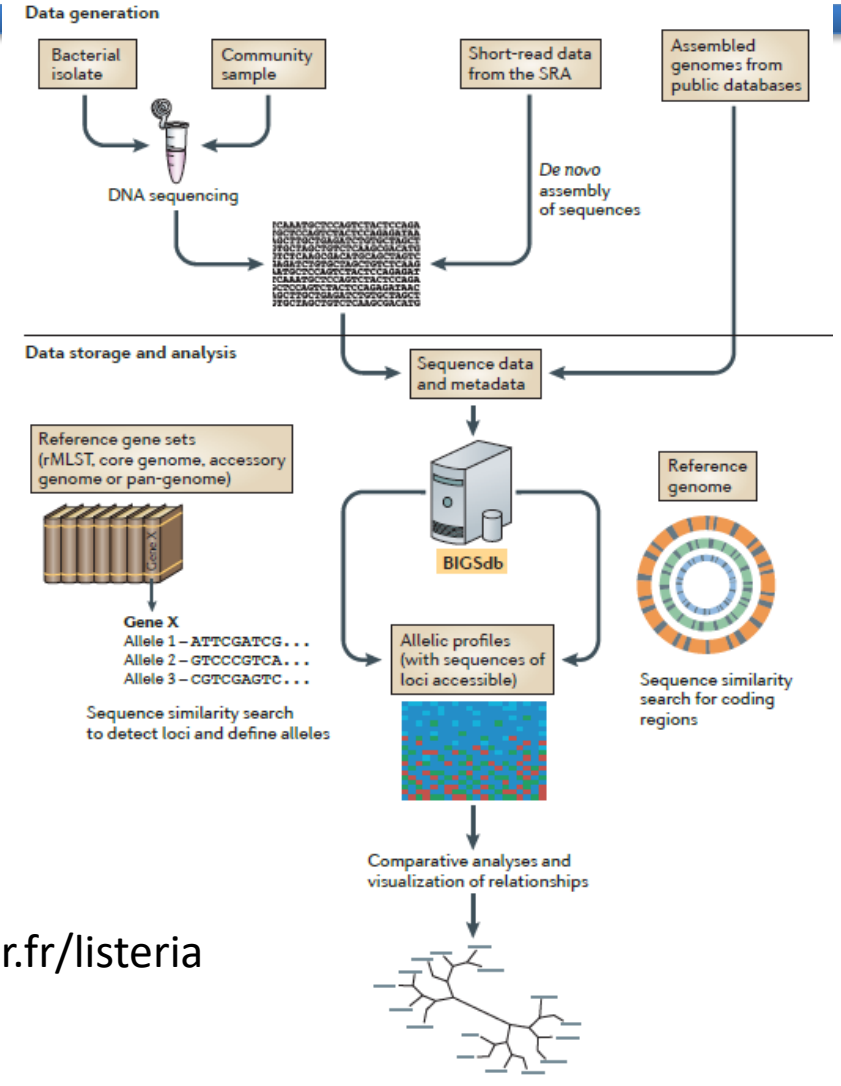
General analytical process for cgMLST / wgMLST



Reproducibility of cgMLST allelic calls, sequencing depth & assembly strategie



BIGSdb-lm



http://bigsdb.pasteur.fr/listeria

Retrospective validation of whole genome sequencing-enhanced surveillance of listeriosis in Europe, 2010 to 2015

Van Walle et al., eurosurveillance 2018

2,726 Lm isolates from human cases from 27 EU/EEA countries, 2010-2015
1,069 isolates -> public health laboratories
1,657 isolates -> commercial sequencing provider
MiSeq (2x150, 2x250, 2x300) NextSeq (2x150), HiSeq (2x100), Ion Torrent PGM

Trimming

- Removal of any adaptor sequences
- Removal of leading bases and trailing bases with Q<25
- Window of 20 bases has average Q<25
- Removal reads with length <36b

Assembly: Spades 3.7.1 or Velvet 1.1.04. Minimum contig length 300nt. Assembly with the highest N50 was retained.

Allelic profile: two subseets of isolate pairs, AD ≤ 7 (closely related isolates likely to share a common epidemiological link), AD ≤ 150 (sublineages where isolates are still likely to have common phenotypic properties that may be relevant, e.g. source atribution)
Moura (Bionumerics: reads (kmers) or BLASTN), 1748 loci
Ruppitsch (SeqSphere +3.4.1), 1701 loci

Retrospective validation of whole genome sequencing-enhanced surveillance of listeriosis in Europe, 2010 to 2015

Van Walle et al., eurosurveillance 2018

Conclusions:

- The average coverage up to around 55x before trimming and 45x after trimming (Illumina)
- Assembly-based allele calling outperforms reads-based allele calling -> more loci were detected (increase typeability) and the average distances between isolates were slightly smaller.
- Velvet including k-mer optimisation performed slightly worse than SPAdes, but both produced near-equivalent results.
- cgMLST analysis to share assembled genomes rather than sequence read data (fastq , for SNPs analysis or to verify the analyses in some cases, e.g. multi-country outbreaks)
- Important: individual differences in Ads between Moura and Ruppitsch CG schemes can be relatively large, since onluy 1,261 loci are common to both schemes.
- The AD≤7 cutoff is useful for cluster detection bor both schemas, in general, although there are exceptions when there are epidemiologically linked isolates with more allele differences than the cutoff (more than one strain or specific sublineages may have higher average mutation rates -> sublineages specific cutoff ??)
- WGS-enhanced surveillance of listeriosis: many clusters found involved more than one country. Earlier detection of clusters.
- The molecular typing results must also be combined with epidemiological and food exposure investigations.

Fifth external quality assessment scheme for *Listeria monocytogenes* typing

Table 1. Number and percentage of laboratories submitting results for each method

	Serotyping				Cluster analysis			
	Conventional only	Molecular only	Both	Total	PFGE-only	WGS-only	Both	Total
Number of participants	1	12	5	18	3	8	4	15
Percentage of participants	6%	67%	28%	90%*	20%	53%	27%	75%*

Thirteen of the 20 participants (65%) completed both parts (serotyping and cluster analysis) of the EQA.
** Percentage of total number of participating laboratories (20)*

Fifth external quality assessment scheme for *Listeria monocytogenes* typing

Annex 7. Reported sequencing details

Sequencing performed	Protocol (library prep)	Commercial kit	Sequencing platform
In own laboratory	Commercial kits	Nextera XT DNA library Preparation Kit*	HiSeq2500
In own laboratory	Commercial kits	NEBNext® Fast DNA Fragmentation & Library Prep Set for Ion Torrent, New England 'Biolabs**	Ion Torrent PGM
Externally	Commercial kits	Illumina	HiSeq 2500
In own laboratory	Commercial kits	Ion Xpress™ Plus Fragment Library Kit for AB Library Builder™ System	IonTorrent S5XL
In own laboratory	Commercial kits	Nextera XT	MiSeq
In own laboratory	Commercial kits	NEXTERA	MiSeq
In own laboratory	Commercial kits	SureSelect QXT Library Prep Kit (Agilent)	MiSeq
In own laboratory	Commercial kits	Nextera XT DNA Library Preparation Kit	MiSeq
In own laboratory	Commercial kits	Nextera XT	MiSeq
In own laboratory	Commercial kits	Nextera XT***	Miniseq
In own laboratory	Commercial kits	Nextera XT Libray Prep kit (96 samples)***	NextSeq
In own laboratory	Commercial kits	Illumina Nextera XT library Prep Kit	MiSeq

** 5ng input DNA (as opposed to 1ng)*
Altered PCR protocol to favour longer fragment sizes
Adjustment of extension temperature (and final extension) from 72° to 65°C
'Manual' normalisation using library concentration and fragment size as opposed to bead-based normalisation.
*** Shearing carried out for 15 minutes at 25°C instead of 20 minutes because 400bp sequencing protocol was used*
**** Half volume for all reagents.*

Fifth external quality assessment scheme for *Listeria monocytogenes* typing

Table 7. Results of raw reads submitted by participants evaluated by EQA provider QC pipeline summarised by laboratory

		Laboratory ID											
Parameters	Ranges*	19	35	56	70	105	108	129	135	141	142	144	146
No. of genera detected	{1}	1	1	1	1	1	1	1	1	1	1	1	1
Detected species	{Lm}	Lm	Lm	Lm	Lm	Lm-N	Lm	Lm	Lm	Lm	Lm	Lm	Lm
Unclassified reads (%)		1.5-2.5	0.6-2.5	0.6-2.2	1.5-2.9	0.7-50.8	1.1-1.8	0.6-1.7	0.5-1.0	0.9-1.8	0.8-1.4	0.2-1.3	0.4-2.0
Length at 25 x min. coverage (Mbp)	{>2.8 ∧ <3.1}	2.9-3.0	2.9-3.0	1.8-2.7	2.9-3.0	0.1-3.0	2.9-3.0	2.9-2.9	2.9-3.0	1.0-3.0	2.9-3.0	2.9-3.0	2.9-3.0
Length [0-25] x min. coverage (Mbp)	{<0.25}	0	0	0	0	0-0.9	0	0-0.1	0	0.0-1.8	0	0	0
No. of contigs at 25 x min. coverage	{>0}	14-21	12-25	876-1056	17-45	14-193	57-146	15-47	17-24	19-85	13-17	11-17	17-25
No. of contigs [0-25] x min. coverage#	{<1000}	0	0	0	0-4	0-517	0-5	0-24	0	0-165	0-2	0	0-1
Average coverage	{>50}	160-224	40-175	61--104	51-100	8-94	30-70	50-244	153-221	24-126	40-58	75-128	140-200
No. of reads (x 1000)		1741-2457	250-1120	707-1278	528-1035	345-622	285-689	530-2704	1898-2835	158-883	261-385	525-881	2148-3169
No. of trimmed reads (x1000)		1721-2428	248-1110	691-1235	524-1028	342-609	521-617	523-2677	1878-2800	150-865	295-380	534-870	2148-3169
Maximum read length		151	301	285-365	151	301	241-319	151	126	301	251	251	101
Mean read length		140-142	215-251	217-229	143-146	204-241	186-200	139-145	123-124	218-235	245-234	210-227	97-100
Read insert size		267.9-305	333-394	NA	288-391	199-363	NA	244-450	326-351	279-358	361-399	280-327	204-360
Insert size StdDev		100-106	158-199	NA	100-149	67-158	NA	108-196	175-188	102-130	157-174	93-125	85-169
N50 (kbp)		238-551	274.4-558	1.4-3.4	162.3-318	1.3-407	34.0-87	125-551	295-482	22-263	262-556	353-558	286-510
N75 (kbp)		143.3-257.3	139-263	0.9-1.9	78-238	0.8-262	23-45	61-258	142-258	11-236	198-262	183-263	144-262

** Indicative QC range*
Lm: L. monocytogenes
N: Neisseria
 # Number of contigs with coverage < 25 (Figure 10B)

Fifth external quality assessment scheme for *Listeria monocytogenes* typing

Table 4. Results of SNP-based cluster analysis

Lab ID	SNP-based						
	Approach	Reference	Read mapper	Variant caller	Assembler	Distance within cluster	Distance outside cluster
Provider	Reference-based	ST6 (REF4)	BWA	GATK		0-3	38-71
19*	Reference-based	ST6 ID 2362	BWA	GATK		0-4	43-81
56	Assembly-based			ksnp3	SPAdes	0-57#	561-591 (6109)
105	Reference-based	ST6 J1817	Bowtie2	VARSCAN 2		0-2#	22-42 (1049)
108	Reference-based	In-house strain resp ST	CLC assembly cell v4.4.2	CLC assembly cell v4.4.2		0-2	37-72
142*	Reference-based	*Listeria* EGDe (cc9)	CLC Bio	CLC Bio		0-1219	1223-2814 (8138)
146	Reference-based	ST6 ref. CP006046 ST1 ref. F2365 ST213/ST382 no ref.	BWA	In-house		0-358	

** Additional analysis*
 # Only three isolates included due to data quality not meeting laboratory's own QC thresholds
¤ Reported distance to ST6 (non-ST6) isolates (Annex 9).

Fifth external quality assessment scheme for *Listeria monocytogenes* typing

Table 5. Results of allele-based cluster analysis

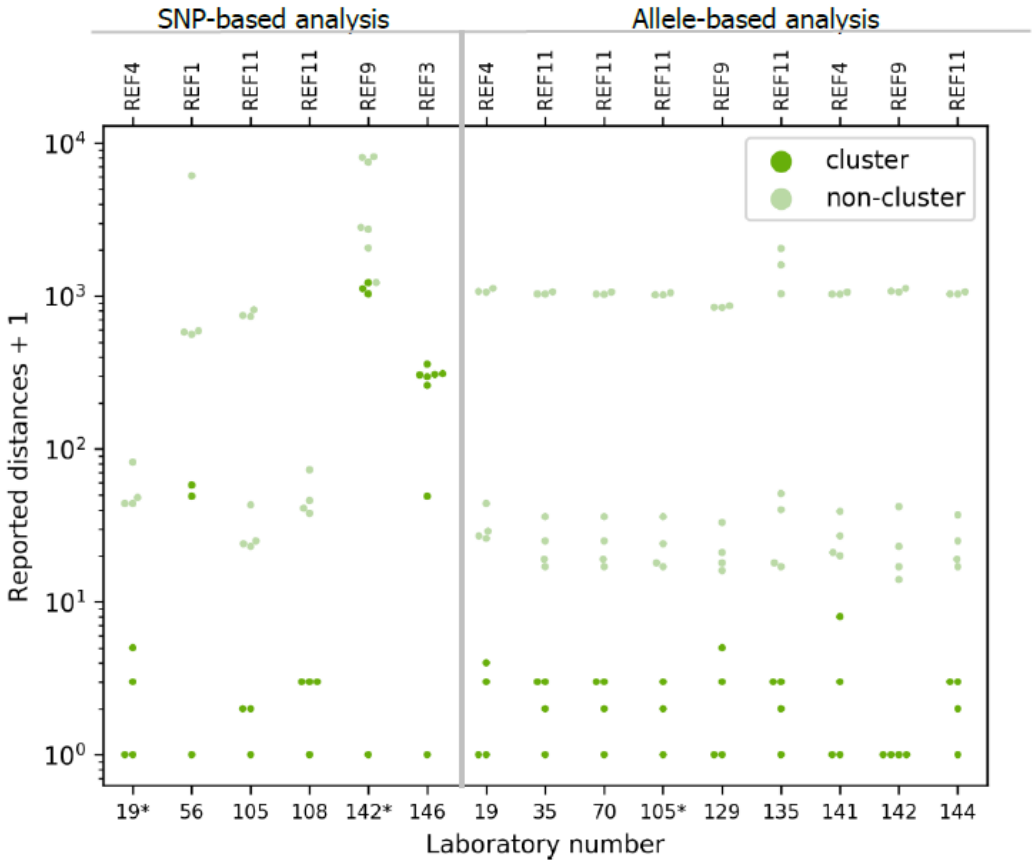
Lab ID	Allele based analysis					
	Approach	Allelic calling method	Assembler	Scheme	Difference within cluster	Difference outside cluster
EQA provider	BioNumerics	Assembly- and mapping-based	SPAdes	Applied Math (cgMLST/Pasteur)	0-3	24-1112
19	BioNumerics	Assembly- and mapping-based	SPAdes	Applied Math (cgMLST/Pasteur)	0-3	25-1120
35	SeqPhere	Assembly-based only	Velvet	Ruppitsch (cgMLST)	0-2	16-1065
70	SeqPhere	Assembly-based only	Velvet	Ruppitsch (cgMLST)	0-2	16-1062
105*	SeqPhere	Assembly-based only	SPAdes v 3.80	Ruppitsch (cgMLST)	0-1#	23-812
129	SeqPhere	Assembly-based only	Velvet	In-house (cgMLST)	0-4	15-862
135	SeqPhere	Assembly-based only	CLC Genomics Workbench 10	Ruppitsch (cgMLST)	0-2	16-2042
141	SeqPhere	Assembly-based only	SPAdes 3.9.0	Ruppitsch (cgMLST)	0-7	19-1060
142	Inhouse	Assembly-based only	SPAdes	Pasteur (cgMLST)	0	13-1120
144	SeqPhere	Assembly-based only	Velvet	Ruppitsch (cgMLST)	0-2	16-1065

** Additional analysis*

 # Only three isolates included due to data quality not meeting laboratory's own QC thresholds (Annex 9).

Fifth external quality assessment scheme for *Listeria monocytogenes* typing

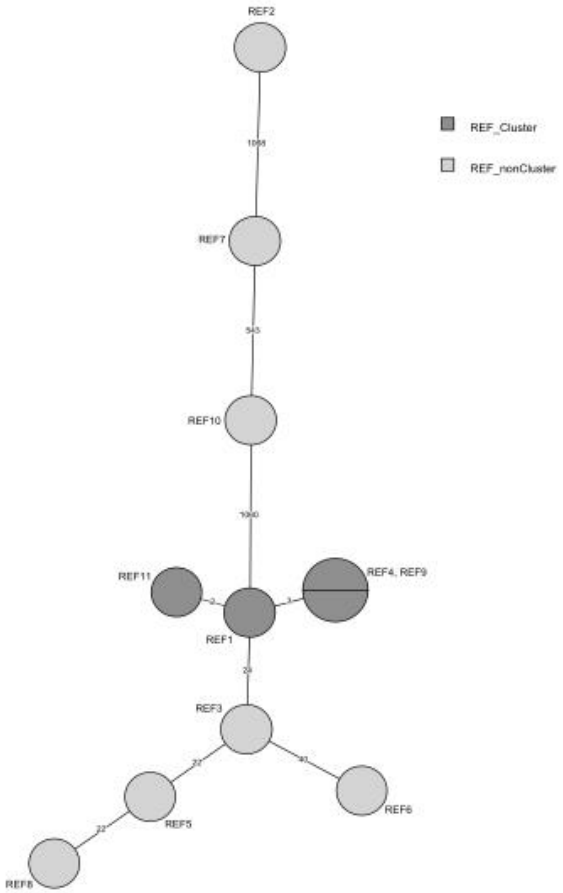
Figure 7. Reported SNP distances or allelic differences for each test isolate to selected cluster representative isolate



** Additional analysis*
SNP: Single nucleotide polymorphism
Selected cluster representative marked as REF in dark green: Reported cluster of closely related isolates
Light green: Not reported as part of cluster.

Fifth external quality assessment scheme for *Listeria monocytogenes* typing

Annex 4. EQA provider cluster analysis based on WGS-derived data



Minimum spanning tree of core genome multi locus sequence typing (cgMLST, [6]) profiles of L. monocytogenes *EQA-5 isolates. Logarithmic scaling in BioNumerics.*

Dark grey: Cluster isolates
Light grey: Outside cluster isolates

Fifth external quality assessment scheme for *Listeria monocytogenes* typing

Annex 5. Reported cluster of closely related isolates based on PFGE-derived data

Lab ID	Reported cluster	Corresponding REF isolates	Correct
Provider	REF1, REF4, REF9, REF11 (4 and 9 technical duplicates)		
19	3362# 2539 2691 2719	REF4, REF1, REF9, REF11	Yes
100	2080 2295 2405 2499	REF4, REF9, REF11, REF1	Yes
105	2073 2709 2805 2978	REF4, REF9, REF1, REF11	Yes
138	2141 2349 2778 2947	REF9, REF1, REF4, REF11	Yes
141	2022 2050 2092 2872	REF1, REF4, REF11, REF9	Yes
142	2385 2529 2794 2837	REF9, REF4, REF11, REF1	Yes
145	2027 2235 2287 2444 2514 2592 2680 2699 2904 2961 2967	REF5, REF9, REF1, REF3, REF11, REF4 REF7, REF2, REF8, REF6, REF 10	No

 # Writing error 2362

Fifth external quality assessment scheme for *Listeria monocytogenes* typing

Annex 8. Reported cluster of closely related isolates based on WGS-derived data

Lab ID	Reported cluster	Corresponding to REF isolates	Correct
Provider	REF1, REF4, REF9, REF11 (4 and 9 technical duplicates)		
19	#3562 3539 2691 2719	REF4, REF1 REF9, REF11	Yes
35	2251 2737 2783 2993	REF11, REF9, REF1, REF4	Yes
56	2341 2165 2612	REF9, REF1, REF11	Yes
70	2104 2216 2567 2767	REF4, REF1, REF11, REF9	Yes
105	2073 2805 2978	REF4, REF1, REF11	Yes
108	2098 2788 2582 2422	REF1, REF11, REF9, REF4	Yes
129	2079 2640 2912 2950	REF1, REF9, REF11, REF4	Yes
135	2161 2423 2673 2897	REF1, REF4, REF11, REF9	Yes
141	2022 2050 2092 2872	REF1, REF4, REF11, REF9	Yes
142	2385 2529 2794 2837	REF9, REF4, REF11, REF1	Yes
144	2143 2626 2727 2822	REF4, REF11, REF1, REF9	Yes
146	2068 2197 2377 2488 ##2353 2575 2655 2726	REF5, REF8, REF3, REF1, REF6, REF4, REF9, REF11	No

#Writing error 2362
##Writing error 2553

Fifth external quality assessment scheme for *Listeria monocytogenes* typing

Annex 9. Reported SNP distance and allelic differences

SNP distances

Isolate no.	ST	Provider	Laboratory ID					
			19*	56	105	108	142*	146
REF1‡	6	3	4	0ᵉ	1	2	1030	306
REF2	1	9999	9999	9999	812	9999	7502	9999
REF3	6	41	47	579	23	45	2814	0ᵉ
REF4‡#	6	0ᵉ	0ᵉ	9999	1	2	1219	309
REF5	6	40	43	561	24	37	2056	259
REF6	6	72	81	591	42	72	2732	358
REF7	213	9999	9999	9999	734	9999	8050	9999
REF8	6	39	43	9999	22	40	1223	48
REF9‡#	6	0	0	57	9999	2	0ᵉ	296
REF10	382	9999	9999	6109	745	9999	8138	9999
REF11‡	6	1	2	48	0ᵉ	0ᵉ	1114	304

Fifth external quality assessment scheme for *Listeria monocytogenes* typing

Allelic distances

Isolates no.	ST	Provider	Laboratory ID								
			19	35	70	105*	129	135	141	142	144
REF1‡	6	3	3	1	1	1	4	1	7	0	1
REF2	1	1118	1120	1065	1062	812	862	2042	1060	1120	1065
REF3	6	25	25	16	16	23	15	16	19	16	16
REF4‡#	6	0¤	0¤	2	2	1	0	2	0¤	0	2
REF5	6	26	26	18	18	24	17	17	20	13	18
REF6	6	44	43	35	35	42	32	50	38	41	36
REF7	213	1073	1070	1028	1026	734	842	1031	1024	1074	1028
REF8	6	28	28	24	24	22	20	39	26	22	24
REF9‡#	6	0	0	2	2	9999	0¤	2	0	0¤	2
REF10	382	1060	1060	1027	1021	745	839	1592	1025	1063	1027
REF11‡	6	3	2	0¤	0¤	0¤	2	0¤	2	0	0¤

** Additional analysis*
‡ Closely related isolates
 # Technical duplicate isolate
¤Isolate used as cluster representative by participant
9999: Isolates not included in analysis by participant
ST: Sequence type

Criteria for wg/cgMLST and SNP typing schemes

Schürch et al., CMI, 2018

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]

Thanks for your attention!