nature microbiology

SUPPLEMENTARY INFORMATION

ARTICLE NUMBER: 16185 | DOI: 10.1038/NMICROBIOL.2016.185

Whole genome-based population biology and epidemiological surveillance of *Listeria monocytogenes*

Alexandra Moura, Alexis Criscuolo, Hannes Pouseele, Mylène M. Maury, Alexandre Leclercq, Cheryl Tarr, Jonas T. Björkman, Timothy Dallman, Aleisha Reimer, Vincent Enouf, Elise Larsonneur, Heather Carleton, Hélène Bracq-Dieye, Lee S. Katz, Louis Jones, Marie Touchon, Mathieu Tourdjman, Matthew Walker, Steven Stroika, Thomas Cantinelli, Viviane Chenal-Francisque, Zuzana Kucerova, Eduardo P. C. Rocha, Celine Nadon, Kathie Grant, Eva M. Nielsen, Bruno Pot, Peter Gerner-Smidt, Marc Lecuit, Sylvain Brisse

Table of Contents

1. Supplementary Tables

Supplementary Table 1. Characteristics of the 1,696 *Listeria monocytogenes* isolates used in this study. [link]

Supplementary Table 2. Loci (*n*=43) excluded from the initial set of 1,791 core genes. [link]

Supplementary Table 3. Characteristics of the 1,748 loci included in the cgMLST scheme. [link]

Supplementary Table 4. Prevalence of sublineages (SL) identified in this study using cgMLST and correspondence with clonal complexes (CC) and sequence types (ST) defined based on conventional MLST. [link]

Supplementary Table 5. Historical SL1 and SL9 isolates used for temporal analyses. [link]

Supplementary Table 6. International clusters of isolates belonging to the same cgMLST type. [link]

Supplementary Table 7. Detection of recombination regions within major sublineages. [link]

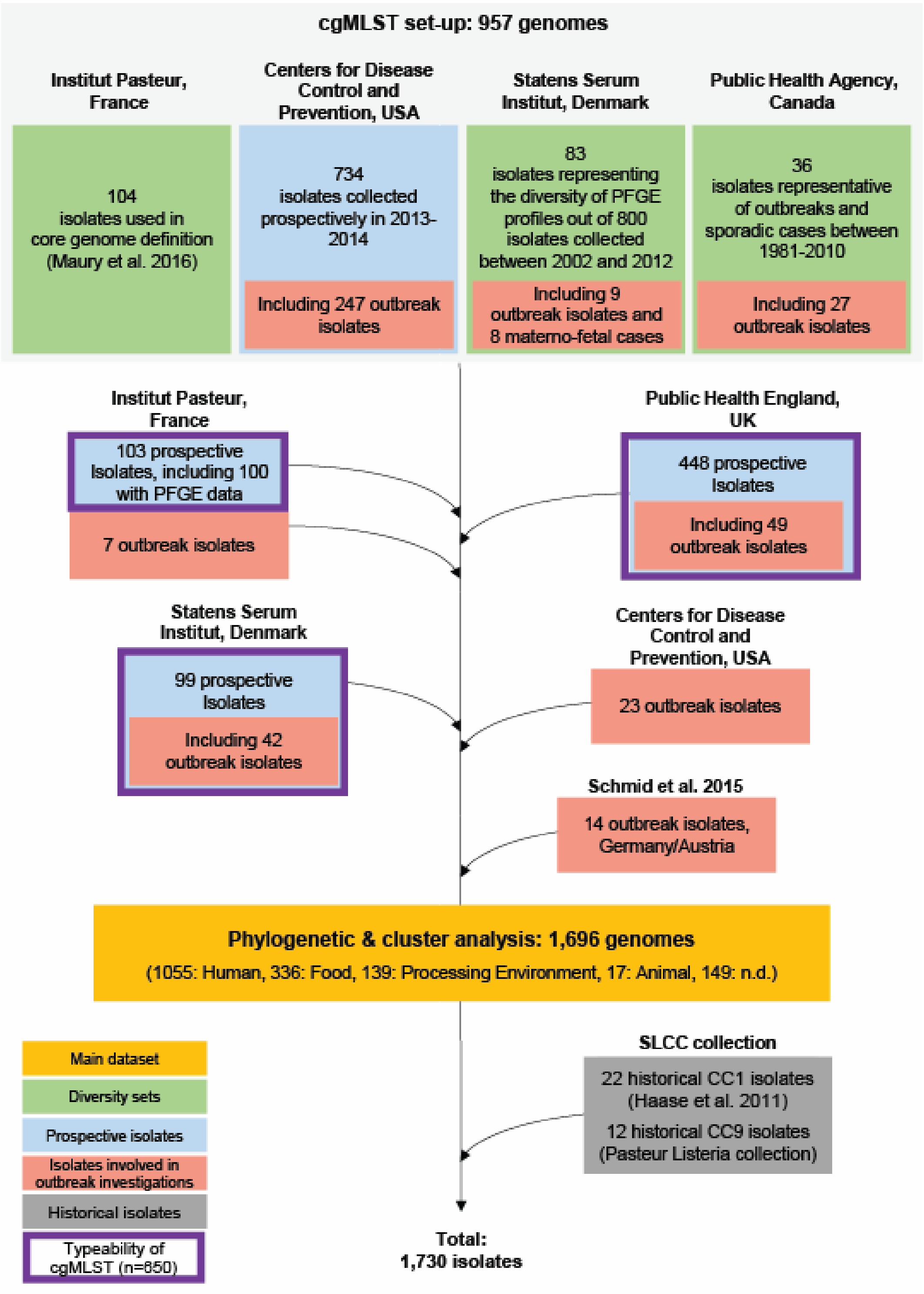
Supplementary Table 8. Frameshifts and mutations identified in this study leading to premature stop codons (PMSC) in *inlA* gene. [link]

2. Supplementary Text

2.1. Validation of a subset of core loci for *Lm* genotyping

The *Lm* core genome was initially defined as the set of 1,791 genes that were always present and complete in 104 genomic sequences of *Lm* strains selected to be representative of all 4 phylogenetic lineages and of the clonal diversity of lineages I and II, as previously described.[1]

To further validate these genes for *Lm* genotyping, an initial set of 996 genomic sequences from four national agencies in charge of listeriosis surveillance (**Supplementary Table 1; Supplementary Figure 1**) was collected. High fragmentation of genomic assemblies was found to have a negative impact on the detection of core genes (**Supplementary Figure 2A**), therefore 39 isolates for which the N50 metric was smaller than 20 kilobases (kb) were removed from further analyses. The remaining 957 high-quality genome sequences (N50 ≥ 20 kb; **Supplementary Table 1; Supplementary Figure 1**) were scanned for gene detection and allele calling using the BLASTN[2] algorithm within BioNumerics v.7.5 (Applied Maths NV, Sint-Martens-Latem, Belgium) and BIGSdb v.1.10 platform[3], with minimum nucleotide identity of 70%, alignment length coverage of 70% and word size of 10, and using EGD-e as reference. To minimize artefactual allele calls, only complete coding sequences were assigned as alleles. Therefore, sequences that lacked start and/or stop codons, as well as genes containing frameshifts, internal stop codons and/or non-GATC characters were discarded. Based on the above, 33 loci for which more than 5% of isolates had missing genes or uncalled alleles were discarded (**Supplementary Table 2**). Further, ambiguous locus detection due to the presence of potential paralogous (highly similar sequences at other genomic locations) was evaluated based on intra-locus *versus* inter-locus variation. Three conflicting pairs of loci (6 paralogous loci) with an average inter-locus similarity >60% were thus eliminated from the final core gene set. Finally, four loci that belong to the conventional 7-gene multilocus sequence typing (MLST) scheme[4] were also eliminated to avoid redundancy, as the MLST loci can be genotyped independently. Applying this filtering procedure led to a final selection of 1,748 core genes used for *Lm* strain genotyping (**Supplementary Table 3**). We defined this subset as the core genome MLST (cgMLST) scheme (**Supplementary Figure 3A**), which can be accessed online through BIGSdb-*Lm* (http://bigsdb.pasteur.fr/listeria/).

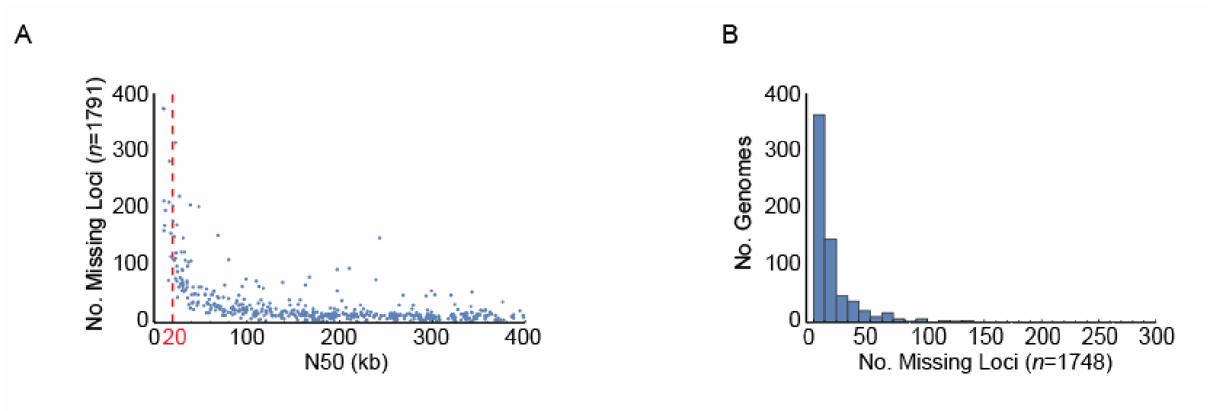

cgMLST set-up: 957 genomes
Institut Pasteur, France
104 isolates used in core genome definition (Maury et al. 2016)
Centers for Disease Control and Prevention, USA
734 isolates collected prospectively in 2013-2014
Including 247 outbreak isolates
Statens Serum Institut, Denmark
83 isolates representing the diversity of PFGE profiles out of 800 isolates collected between 2002 and 2012
Including 9 outbreak isolates and 8 materno-fetal cases
Public Health Agency, Canada
36 isolates representative of outbreaks and sporadic cases between 1981-2010
Including 27 outbreak isolates
Institut Pasteur, France
103 prospective Isolates, including 100 with PFGE data
7 outbreak isolates
Public Health England, UK
448 prospective Isolates
Including 49 outbreak isolates
Statens Serum Institut, Denmark
99 prospective Isolates
Including 42 outbreak isolates
Centers for Disease Control and Prevention, USA
23 outbreak isolates
Schmid et al. 2015
14 outbreak isolates, Germany/Austria
Phylogenetic & cluster analysis: 1,696 genomes
(1055: Human, 336: Food, 139: Processing Environment, 17: Animal, 149: n.d.)
SLCC collection
22 historical CC1 isolates (Haase et al. 2011)
12 historical CC9 isolates (Pasteur Listeria collection)
Main dataset
Diversity sets
Prospective isolates
Isolates involved in outbreak investigations
Historical isolates
Typeability of cgMLST (n=650)
Total: 1,730 isolates

Supplementary Figure 1. **Flowchart of the study.** The different isolates datasets and their relationships and sources are indicated.

cgMLST genes are evenly distributed around the EGD-e reference strain chromosome (**Supplementary Figure 3A**) and represent 62% of its coding sequences.

To estimate cgMLST locus typeability[5], the genomes of an independent set of 650 prospective isolates from the United Kingdom and France were tested (**Supplementary Table 1; Supplementary Figure 1**). Average locus-level typeability was 99.7%, leading to 0 to 135 uncalled alleles (including fully missing loci) per genome (median: 8; average: 15± 20; **Supplementary Figure 2B**).

Allelic diversity among the entire dataset of 1,696 genomes ranged from 2 to 185 alleles per locus (average±standard deviation of 66.6 ± 0.02; **Supplementary Figure 3B**), with an average±standard deviation locus length of 900±550.7 nucleotides (**Supplementary Figure 3C**). The number of distinct allelic profiles (ignoring missing data in pairwise comparisons) was 1,370 out of 1,696 genomes, including epidemiologically related isolates, corresponding to a Simpson's index[6] of discrimination of 99.9% (95% confidence interval = [99.6, 99.7%]). These results show that our cgMLST scheme is very close to maximal resolution of *Lm* isolates.

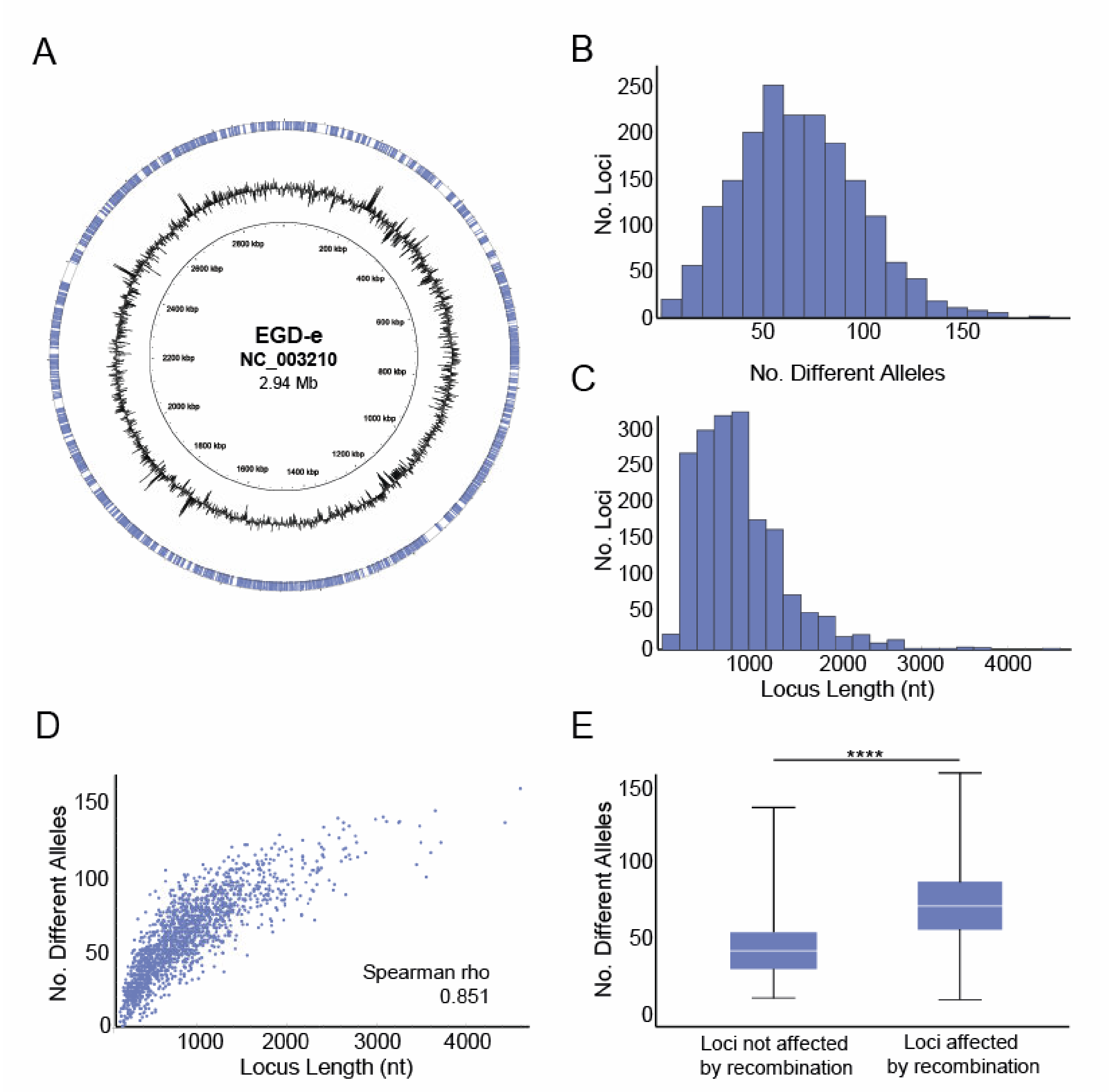


Supplementary Figure 2. Validation of the cgMLST scheme. A) Impact of the N50 assembly size in the number of missing loci (996 genomes). Dashed bar marks N50 of 20 kb, the cut-off established for the genomes used in further analyses. B) Distribution of the number of missing loci per genome (validation set of 650 genomes).

2.2. Diversity, selection and homologous recombination observed at cgMLST loci

The evolutionary factors leading to allelic variation were investigated. Allele nucleotide diversity (π)[7] and mutation rates (θ)[8] were estimated using DnaSP v.5.10.[9] The ratio of synonymous and non-synonymous substitutions (dN/dS) were measured using the Yang and Nielsen method[10] in the yn00 program of PAML v.4.7 package.[11] The presence of homologous recombination was inferred using the PHI statistical test.[12]

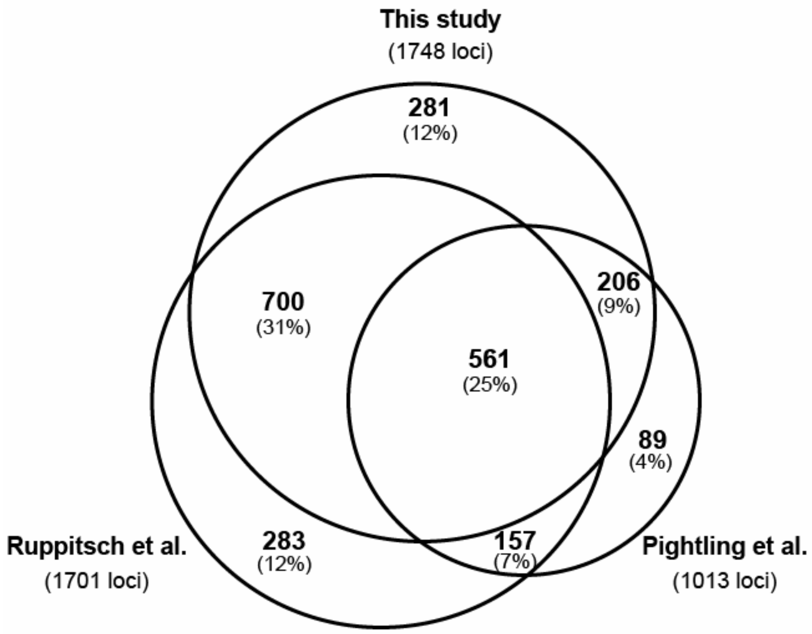
Nearly 60% of loci (1,036/1,748) were significantly affected by homologous recombination events (PHI statistic test, p<0.05, **Supplementary Table 3**), which increased allele diversity (**Supplementary Figure 3E**). As expected, allelic diversity also increased with locus length (Spearman's rho = 0.851, **Supplementary Figure 3D**). Importantly, nearly all core genes (1,731 out of 1,748, 99%) showed dN/dS<1 (**Supplementary Table 3**), suggesting that they were predominantly affected by purifying selection. The remaining 17 loci corresponded to ribosomal (n=13) or hypothetical (n=3) proteins, with limited allelic and nucleotide diversity, thus affecting stochastically the dN/dS ratio, and to *lmo1341,* coding for the ComG competence protein. This 17 loci showed no variation, and were therefore not affected by positive selection, within outbreak sets. These results suggested that most variation at cgMLST loci is neutral and therefore suitable for inference of strain phylogenetic relationships and interpretation of variations within outbreaks.



Supplementary Figure 3. Characteristics of the 1,748 genes (loci) of the cgMLST scheme. A) Location of the core genes (blue lines on external circle) along the EGD-e reference genome. Inner circles represent GC content and the genomic positions, respectively. B) Distribution of the number of different alleles per locus. C) Distribution of the core loci lengths. D) Impact of locus length on allele diversity (*n*=957 genomes). E) Impact of recombination on allele diversity. Boxes show the median, 25th and 75th percentiles and error bars show the highest and lowest values. Stars denote statistical significance (*p*<0.0001, Mann-Whitney U test).

2.3. Comparison with other genome-based MLST schemes

During this work, two other genome-based MLST schemes were proposed, based on a more limited set of genomes.[13,14] An overview of the common loci between the three schemes is presented below (**Supplementary Figure 4**). The present scheme includes a higher number of loci than previously proposed,[13,14] thus providing improved subtyping resolution.



Supplementary Figure 4. Comparison on the set of loci from this study with those from other cgMLST schemes.[13,14]

2.4. Reproducibility of cgMLST allelic calls using distinct bioinformatics strategies and software tools, and its dependency on sequencing depth

Robustness of allele calls was first evaluated by comparing the alleles called from *de novo* assemblies using the BLASTN algorithm, with alleles defined using an assembly-free approach.

The assembly-free allele calling algorithm detects the presence of a particular allele directly from the unassembled reads by determining whether the k-mers (i.e., short sequences of k nucleotides) present in the allelic sequence are also present in the short sequence reads. As every allelic sequence (up to repeated fragments) is unique, every allelic profile of k-mers is unique and therefore the profile can be used as an unambiguous marker for the presence of an allele. It is possible that the k-mers in an allelic profile are found in reads originating from entirely different locations in the genome, but the likelihood thereof is greatly reduced by using larger k-mer lengths (at least 20).

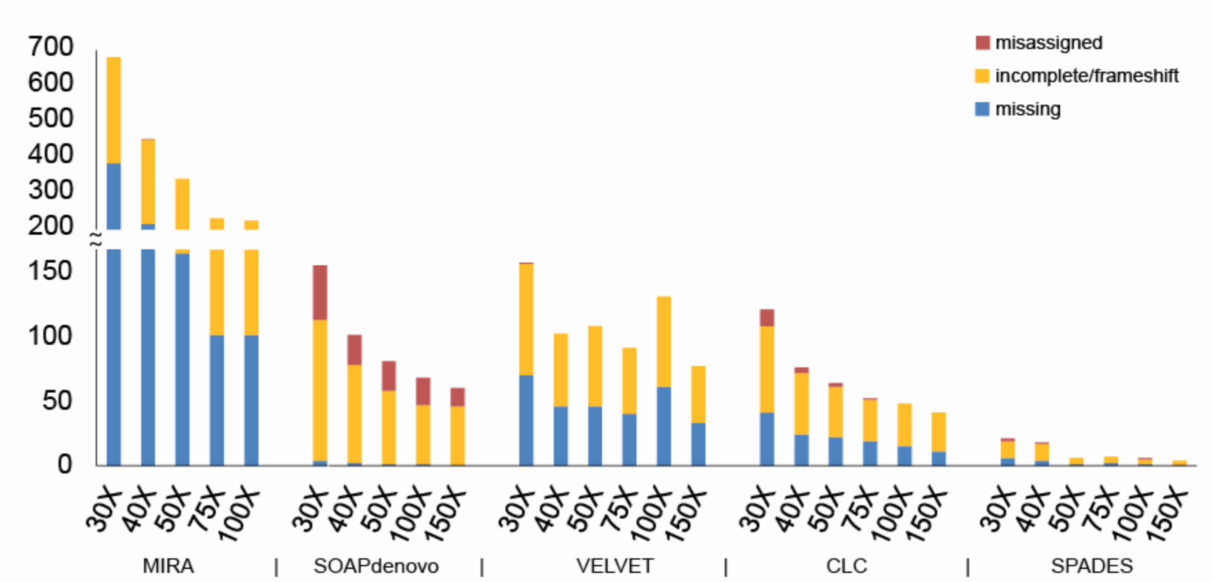
Two mechanisms were put in place to avoid sequencing errors. First, only those parts of a short sequence read that had enough high-quality bases were used (typically, no more than 3 bases could have a Phred score of ≤20). Second, k-mers were considered to be present in the short sequence reads only if they were present at least three times, and at least once in both forward and reverse directions. This method was developed by Applied Maths, is protected by patent EP15153406.2 (pending) and was made available through BioNumerics and the BioNumerics Calculation Engine.

To assess reproducibility and identity of the assembly-based and assembly-free alleles, a subset of 83 genomes from Denmark (**Supplementary Table 1; Supplementary Figure 1**), for which sequence reads were available early during the project, was used. Allele detection from *de novo* assembled genomes was performed using either BIGSdb-*Lm* or BioNumerics, using the BLASTN settings described above. Assembly-free allele calling from raw sequence reads was performed with BioNumerics using k-mers of size 35 after eliminating reads with low sequence quality (at most 3 bases with a Phred quality score below 20 per 35 bases). All assembly-based allele calls defined using BIGSdb and BioNumerics were identical. In addition, all allele calls were confirmed by assembly-free calling. These results correspond to an allele calling error rate smaller than 0.0007 (1 out 145,084 calls).

To assess reproducibility of cgMLST across independent cell cultures, we compared the *Lm* EGD-e reference strain sequence determined by Sanger sequencing[15] (NC_003210.1) to two independent Illumina sequence runs. First, publicly-available reads were retrieved from NCBI's sequence read archive (accession no. ERX705165). Second, the strain EGD-e that was available at the National Reference Laboratory for *Listeria* (Institut Pasteur, France) was independently cultured and sequenced. Reproducibility of allele calls between the three independent cultures and sequencing runs was absolute (error rate < 0.029%, i.e. < 1 error in 3,496 allelic comparisons).

To assess reproducibility across assembly software tools and different sequencing depths, Illumina read sets corresponding to different sequencing depths (10X to 125X) were created *in silico*. Popular assembly tools were then used to assemble the resulting read sets: MIRA v.4.0.rc5[16], CLC Assembly Cell v.4.3.0 (Qiagen, Aarhus, Denmark), SOAPdenovo v.2.04[17], Velvet v.1.2.10[18] and SPAdes v.3.1.0[19], as implemented in iMetAmos[20] using default options, after having trimmed the reads with AlienTrimmer[21] to eliminate adapter sequences and sequencing errors from the read ends. Allele calls were determined for each assembled contigs using BIGSdb-*Lm* and were then compared across sequencing depths and *de novo* assembly softwares in terms of number of loci detected and of allelic mismatches compared to the reference sequence NC_003210.1 (**Supplementary Figure 5**).

The results showed that coverage depths of 40X or above led to a high proportion (>96.86%) of loci being called (**Supplementary Figure 5**), except for MIRA assembler (82.15%). Coverage depths below 40X yielded lower allele call rates due to more fragmented assemblies. Among detected alleles, few allele call differences (see 'misassigned' in **Supplementary Figure 5**) were observed across multiple allele comparisons (<0.02%, corresponding to 6 different alleles out of 26,090 comparisons, at coverage depths ≥40X, using CLC, MIRA, Velvet and SPAdes assemblers). The exception was for SOAPdenovo assembler that showed an error rate of 1.21% (82 wrong calls out of 6,793 comparisons). Overall, these controls demonstrated the very high robustness of cgMLST allele calling based on *de novo* genome assembly, with extremely low error rates with most assemblers.



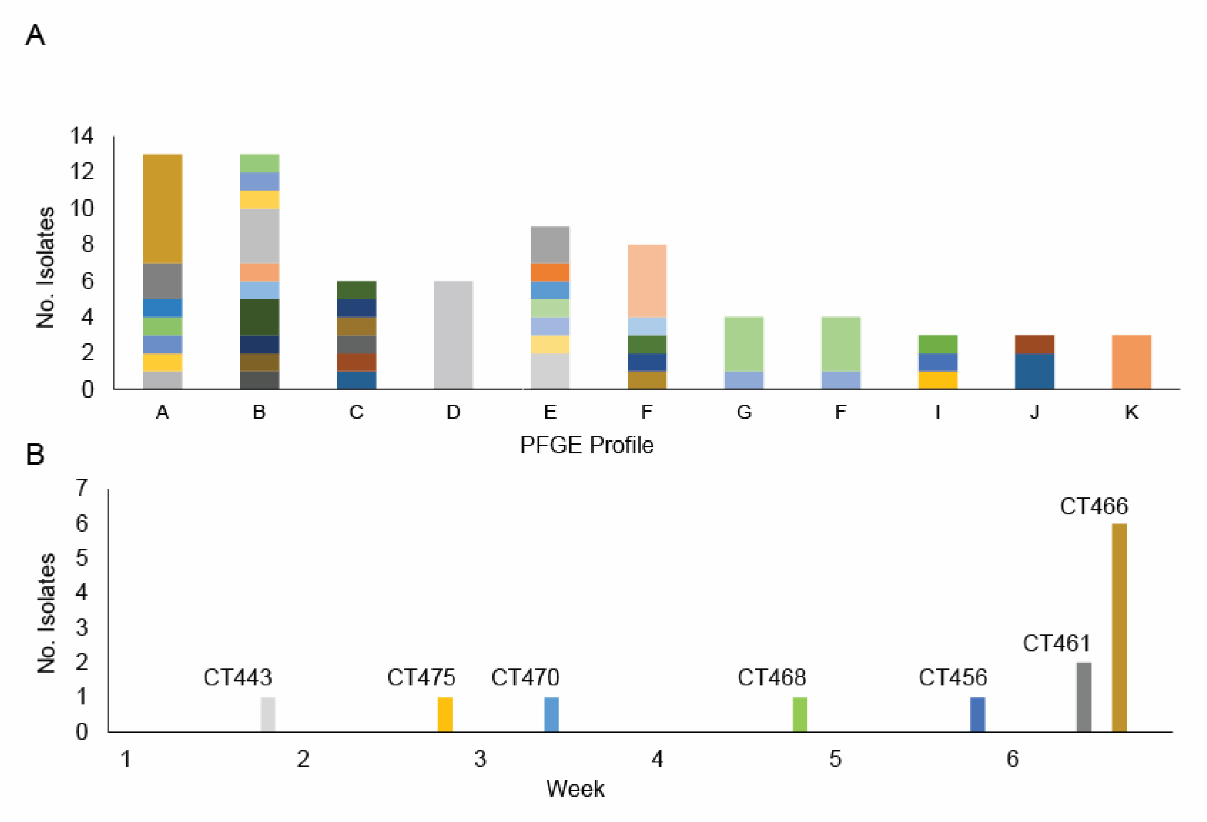
Supplementary Figure 5. Accuracy and sensitivity of assembly-based allele calling using different assemblers (MIRA, SOAPdenovo, Velvet, CLC, SPAdes) and coverage depths (30, 40, 50, 75, 100, and 150X). Each bar summarizes a simulated dataset.

2.5. Comparison of cgMLST and pulsed-field gel electrophoresis (PFGE) genotyping

The comparison of cgMLST and PFGE typing methods was performed using a subset of 100 isolates obtained in the frame of the surveillance system of listeriosis in France by the National Reference Center for *Listeria* of Institut Pasteur, France (**Supplementary Figure 1**). PFGE banding patterns were compared using the complete linkage clustering algorithm based on the number of different bands using BioNumerics. Band comparison parameters were set at 1% for the overall pattern matching optimization setting and at 1% for band position tolerance. *Apa*I types or *Asc*I types were defined for each enzyme separately, as differing from other types by at least two bands (i.e., a difference of only one band was tolerated). Combined *Apa*I-*Asc*I PFGE types were defined as being of a distinct type for at least one of the two enzymes.

The comparison of cgMLST and PFGE discrimination levels was performed using the Simpson's index of diversity[6] and adjusted Wallace index of concordance,[22] calculated using the Comparing Partitions online tool.[23]

Identical PFGE profiles were subdivided into different cgMLST types (**Supplementary Figure 6**), highlighting the improved resolution of cgMLST over PFGE.



Supplementary Figure 6. PFGE *versus* cgMLST discrimination. A) Number of isolates per PFGE profile (A to K) collected in the first trimester of 2015 at the French National Reference Center for *Listeria*. Only PFGE profiles with more than 3 isolates are shown. Within each PFGE profile, different CTs are represented by different colors (independently for each PFGE profile). B) Timeline of the different CTs observed among isolates of PFGE profile A. Note that cgMLST allowed to define narrower clusters (CTs) that were sampled across short time spans. CT461 and CT466 isolates were isolated from two distinct food-production plants, respectively.

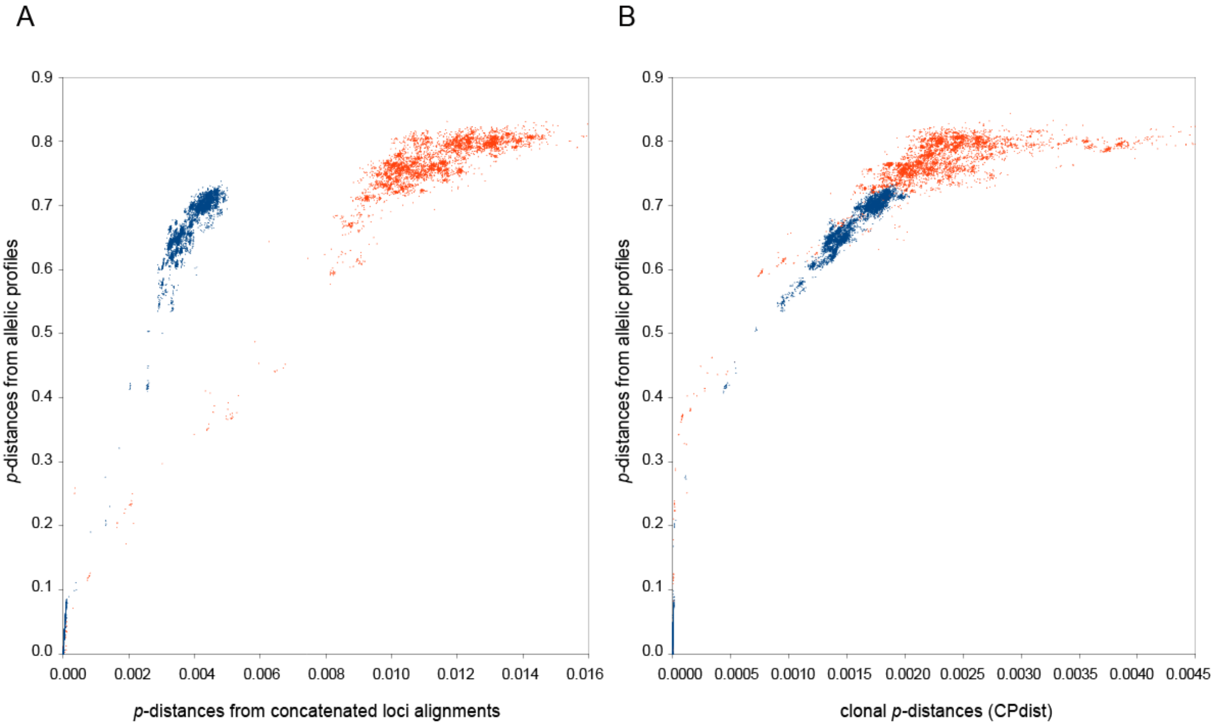
2.6. Phylogenetic analyses

For each core gene, deduced amino acid sequences were first aligned with MAFFT[24] v.7.149 (default options) and back-translated, leading to 1,748 codon-level multiple sequence alignments (1.58 Mb).

In order to obtain a global view of *Lm* diversity, a distance-based phylogenetic tree was inferred from the concatenation of the 1,748 multiple codon sequence alignments. Pairwise dissimilarities were estimated with the p-distance (i.e. the proportion s/l of nucleotide characters at which two aligned sequences are different, where s is the number of differences and l the total number of aligned characters) and used for tree inference with FastME v.2.07[25] (Balanced Minimum Evolution criterion, NNI- and SPR-based minimum evolution tree search).

In order to infer a phylogenetic tree of the *Lm* isolates based on clonal evolutionary events only, pairwise dissimilarities were estimated with the clonal p-distance (cp-distance, i.e. the proportion of nucleotide characters at which two aligned sequences are different owing to only clonal evolutionary events), using CPdist (https://github.com/C3BI-pasteur-fr/CPdist). Given a pair of isolates, when distributing the k = 1,748 p-distance estimates s_m/l_m (m = 1, …, k), one can sometimes observe some unexpectedly too small or large values likely caused by homologous recombination events between the two isolates or with a third one, respectively.[26] To deal with these putative outlier values, the cp-distance S/L was estimated where $S = \sum_{m=1,\dots k} w_m\ s_m$, $L = \sum_{m=1,\dots k} w_m\ l_m$, and w_m a variable used to down-weight the unexpectedly too small or large p-distance estimates s_m/l_m in order to minimize their negative impact on the weighted average S/L. As each s_m could be considered as distributed following a theoretical Negative Binomial (NB) distribution with parameters depending on l_m, the weighting scheme for the cp-distance was defined by $w_m = \min[\mathrm{P}(X < s_m), \mathrm{P}(s_m < X)]$ where X follows the theoretical NB distribution. Of note, estimating the p-distance from concatenated loci is equivalent to estimate S/L with w_m = 1. Tree reconstruction from the estimated cp-distances was performed with FastME v.2.07 (see above). Phylogenetic trees were visualized using the CLC Genomics Workbench v.8.5 (Qiagen, Aarhus, Denmark). Tanglegrams were obtained using Dendroscope v.3.2.10[27]. Clustering compatibility between cgMLST single linkage clusters (see next section) and phylogenetic trees based on concatenated multiple sequence alignments was assessed based on pairwise clade compatibility.[28] Briefly, as a subset S of isolates is compatible with a clade C if and only if $S \cap C \in \{S, C, \emptyset\}$, each cgMLST cluster of isolates was considered as compatible with the phylogenetic tree when compatible with each clade of the tree. With a threshold of 7 and 150 allelic differences, 224 out of 226 (99.1%) and 70 out of 80 (87.5%) cgMLST non-trivial clusters (i.e. of size larger than 1) were compatible with the sequence-based phylogenetic tree, respectively. When considering the entire dataset (i.e., including the trivial clusters of size 1), 995 clusters out of 997 (99.8%) and 129 clusters out of 139 (92.8%) were compatible with the phylogenetic tree at thresholds 7 and 150 allelic differences, respectively.

cgMLST allelic distances (i.e. proportion of mismatched loci among those that are called in both strains) were strongly correlated with sequence-based *p*-distances (**Supplementary Figure 7A**) and with clonal-based distances (**Supplementary Figure 7**). Note that the effect of recombination purging is stronger on lineage II, as expected given its higher recombination rate[29].



Supplementary Figure 7. Correlation between cgMLST (Y-axis) and sequenced-based dissimilarities (X-axis) from A) non-purged and B) recombination-purged concatenated core loci alignments in lineages I (blue) and II (red). Note the continuous increase of cgMLST distance as a function of sequence-based distances, and the more pronounced reduction of the nucleotide-based distances in lineage II when accounting for recombined sequences. Saturation of cgMLST distance is observed for large dissimilarities corresponding to some inter-sublineages comparisons.

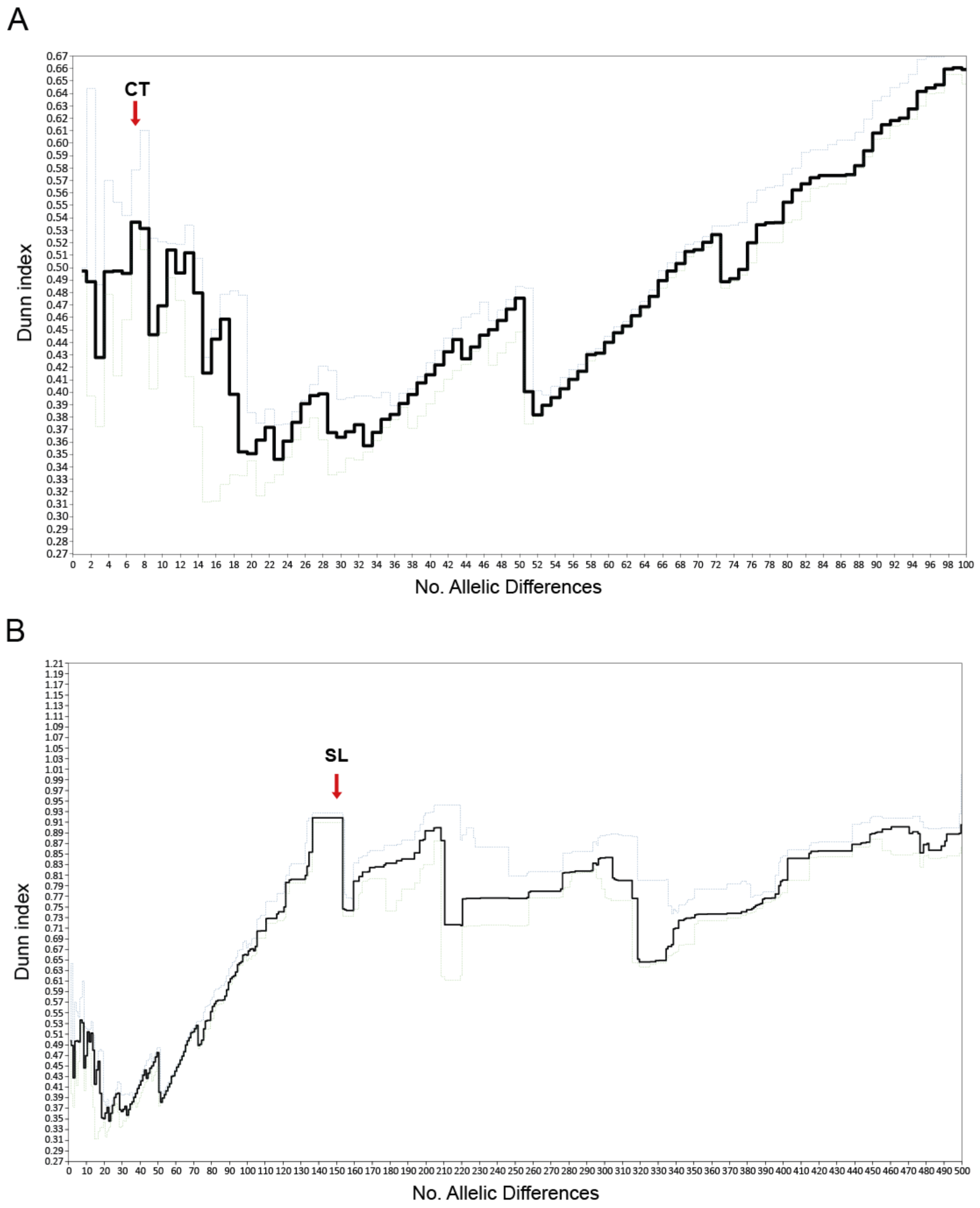
2.7. Choice of cut-off values and nomenclature of sublineages and cgMLST types

In order to optimize the allelic difference thresholds used to define cgMLST types (CTs) and sublineages (SLs), cgMLST allelic distances (i.e. the proportion of mismatched loci among those that are called in both strains) was estimated from each pair of allelic profiles. A single linkage clustering (equivalent to minimum spanning tree or eBURST approaches to define clonal complexes) of the 1,696 isolates was performed from the so-obtained cgMLST allelic distances with varying thresholds $t = d / k$, with k = 1,748 loci and d = 1, 2, ..., 500 allelic differences. In order to assess the compactness and separateness of the clustering obtained for each threshold t, Dunn's index [30] was estimated with the following formula:

$$D = \min_{xy} [\ \Delta(C_x, C_y)\ ] / \max_x [\ \delta(C_x)\ ]$$

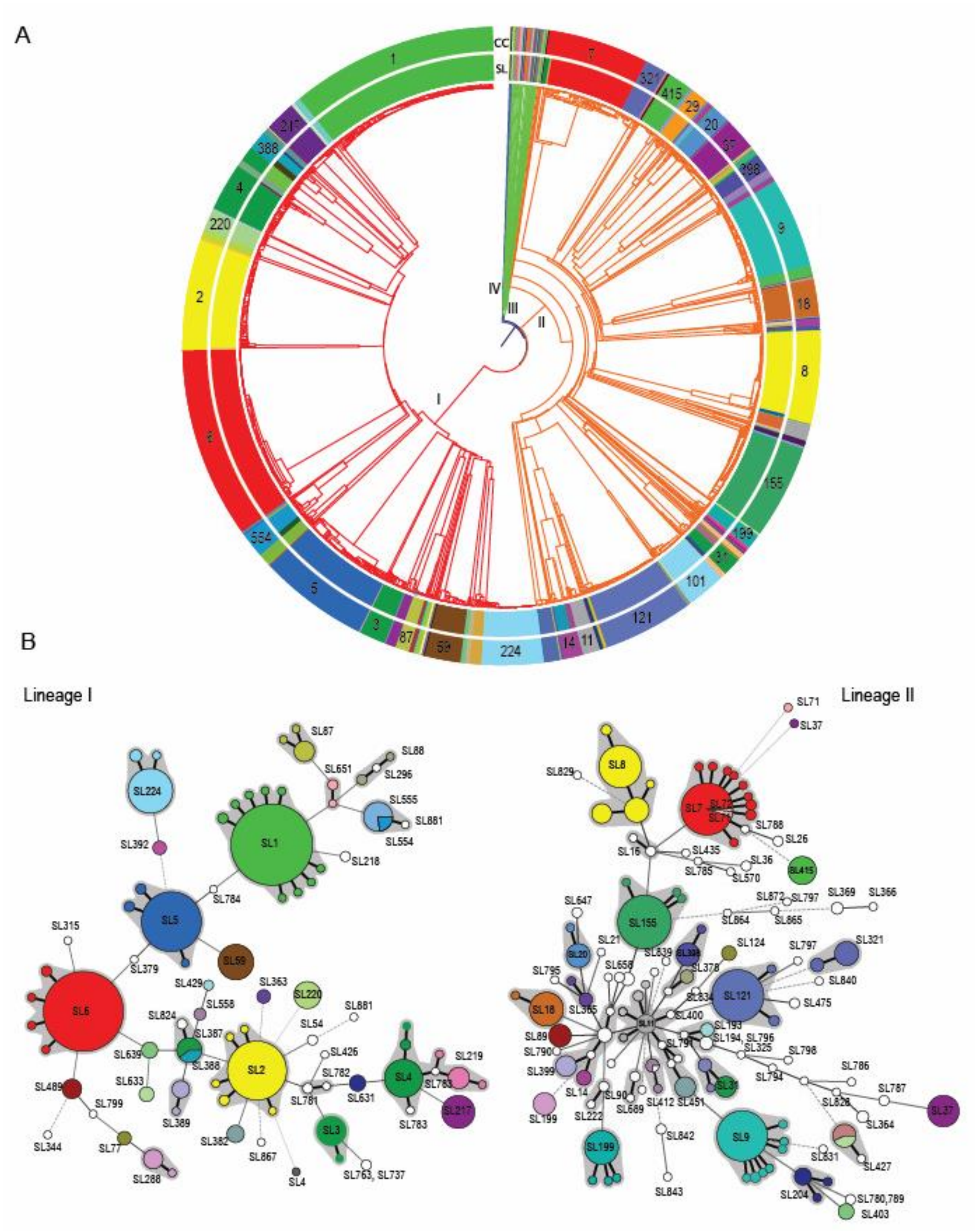
where C_x is the xth cluster (containing at least one isolate), $\Delta(C_x, C_y)$ is the minimum cgMLST allelic distance observed among every pair of isolates $i \in C_x$ and $j \in C_y$, and $\delta(C_x)$ is the diameter of the cluster C_x (i.e. the maximum cgMLST allelic distance observed among every pair of isolates $i,j \in C_x$). According to Dunn,[30] a threshold t that maximizes D corresponds to an optimal clustering (i.e. that best fits the natural tendency of the elements to be clustered).

Finally, in order to assess the stability of the CT and SL threshold values (i.e. d = 7 and 150 allelic differences, respectively) according to the strain sampling, 500 data replicates and corresponding cgMLST allelic distance matrices were generated by performing bootstrapping (i.e., random isolate sampling with replacement). For each threshold t, the 500 Dunn's indices D were computed. As illustrated in **Supplementary Figure 8**, d = 7 and 150 allelic differences both correspond to local optima of Dunn's index (i.e. D = 0.538 and 0.921, respectively). Moreover, Dunn's indices for these two clustering threshold values were significantly better than those of neighboring values (p=2.10^{-16}, Mann Whitney test).



Supplementary Figure 8. Dunn's index values based on single linkage clustering generated from 500 bootstrap replicates of isolates A) within the range of 0 to 100 allelic differences and B) 0 to 500 allelic differences. Arrows represent the optimal clustering values set for cgMLST types (CT) and sublineages (SL). Black: median; green and blue: 1st and 3rd quartiles, respectively.

To facilitate comparisons with previous epidemiological and population biology knowledge based on MLST, the correspondence of sublineages with classical 7-genes MLST nomenclature was established. 156 previously-known STs and 63 novel ones were identified and integrated in the MLST nomenclature database, which we integrated in the BIGSdb-*Lm* database. There was an excellent correspondence between sublineages and MLST-defined clonal complexes,[4] as shown in **Supplementary Figure 9**. Therefore, the MLST nomenclature was mapped onto sublineages where possible (**Supplementary Table 4**).



Supplementary Figure 9. cgMLST *versus* MLST nomenclature. A) Single-linkage based clustering of cgMLST allelic profiles showing a general overview of the correspondence between clonal complexes (CC, based on MLST; external circle) and sublineages (SL, based on cgMLST; inner circle). Branches are colored by phylogenetic lineage (I, red; II, orange; III, green; IV, blue). B) Minimum spanning trees of lineages I and II based on allelic profiles of the 7-locus MLST scheme. STs are represented by circles proportional to the number of isolates and colored by sublineage, as determined by cgMLST. The length of lines connecting STs is proportional to the number of allelic differences; dashed lines represent 4 or more allelic differences between MLST profiles. Grey zones around groups of circles denote MLST clonal complexes.

2.8. Temporal analysis of major sublineages

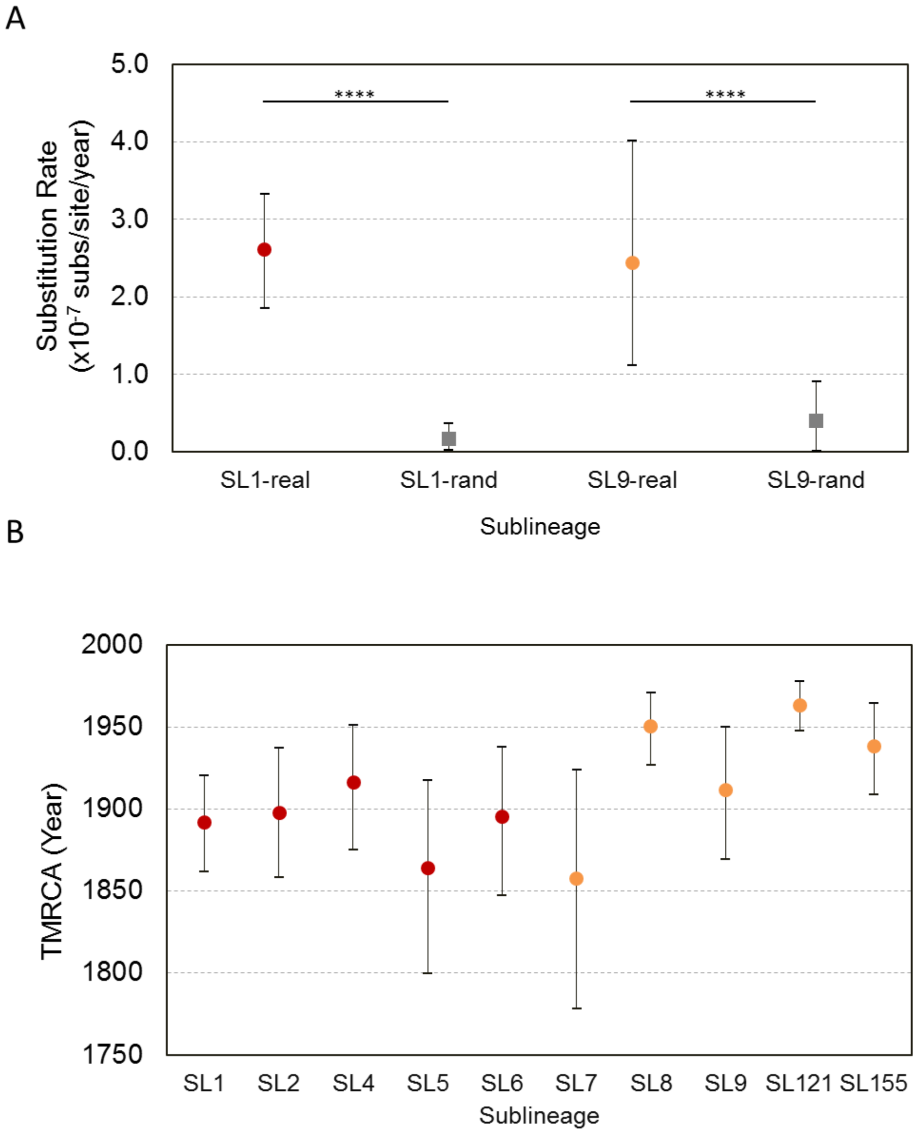
To assess the measurability of a temporal signal in *Lm* genomic sequences and to estimate the evolutionary rate of sequences and cgMLST profiles in lineage I and lineage II, historical isolates from the Seeliger collection (revived by the French Reference Centre for *Listeria* for CC9, and for CC1 by Jana Haase, Mark Achtman and their colleagues[31]) were also sequenced using Illumina technology. These consisted of 22 isolates belonging to MLST clonal complex CC1 (herein called SL1) collected between 1921 and 1974 and 12 isolates from clonal complex CC9 (SL9) collected between 1937 and 1974 (**Supplementary Table 5**). For temporal analyses, these historical isolates were analyzed jointly with the SL1 and SL9 isolates from the main dataset (**Supplementary Table 1**).

Linear regression of the root-to-tip distances of SL1 and SL9 cgMLST profiles against the year of isolation was carried out using Path-O-Gen v1.4 (http://tree.bio.ed.ac.uk/software/pathogen/) using the FastME trees, which do not assume a molecular clock hypothesis, as required for root-to-tip distance analysis. Allelic profiles analysis lead to an estimated evolutionary rate of 1.30x10^{-4} allelic changes per locus per year for SL1 (*n*=195) and of 2.01x10^{-5} allelic changes per locus per year for SL9 (*n*=61).

The rate of evolution of SL1 and SL9 genomes was also independently estimated using BEAST v.2.3.1,[32] from recombination-purged multiple sequence alignments. Gubbins[33] was used to detected recombination within the alignments, using default parameters and a minimum of 3 base substitutions required to identify recombination. Recombination regions (varying from 4 up to 454 bases) were identified in 90 out of 853 isolates (11%, **Supplementary Table 7**). Isolates with recombinant regions were completely discarded from multiple sequence alignments, to avoid amplification of artifacts in the analyses caused by the removal of recombined regions only.[34] BEAST estimations were performed using the nucleotide evolutionary model HKY85+Γ4 and a default gamma prior distribution of 1. To identify the most suitable model, strict, lognormal relaxed and exponential-relaxed molecular clocks and the coalescent constant and exponential population size models were tested. For each parameter combination, three independent runs were performed, each consisting of MCMC chains of 200 million cycles, with the first 100,000 trees discarded as burn-in and further sampling every 10,000 steps. The effective sample size (ESS) values were confirmed to be higher than 200 for all parameters using Tracer v.1.5. Models were compared by marginal likelihood and stronger support was obtained for the exponential-relaxed molecular clock and coalescent exponential population (Bayes factor > 200). To assess the significance of the temporal signatures observed, the substitution rates were also estimated using randomized tip date datasets as controls.[35] Estimated rates based on the real data were significantly different from those estimated based on randomized tip dates (*p*<0.0001, unpaired two tailed t-test; **Supplementary Figure 10A**).

Mean estimated rates ranged from 2.61x10^{-7} substitutions per site per year (equivalent to 1 substitution per 1.58 Mb of cgMLST sequence every 2.4 years) for SL1 and 2.44x10^{-7} substitutions per site per year (equivalent to 1 substitution per 1.58 Mb of cgMLST sequence every 2.6 years), as shown in **Supplementary Figure 10A**.

Estimations of the age of the most recent common ancestor of major sublineages were then performed using BEAST v.2.3.1[32], with the nucleotide evolutionary model HKY85+Γ4 and the coalescent exponential population models along with exponential-relaxed molecular clock of 2.50x10^{-7} substitutions per site per year (which corresponds to the arithmetic mean of SL1 and SL9 estimates) as a prior. Three independent runs were performed, each consisting of MCMC chains of 200 million cycles, with a burn-in of 100,000 and sampling every 10,000 steps. Results of age estimates of the sublineages are presented in **Supplementary Figure 10B.**



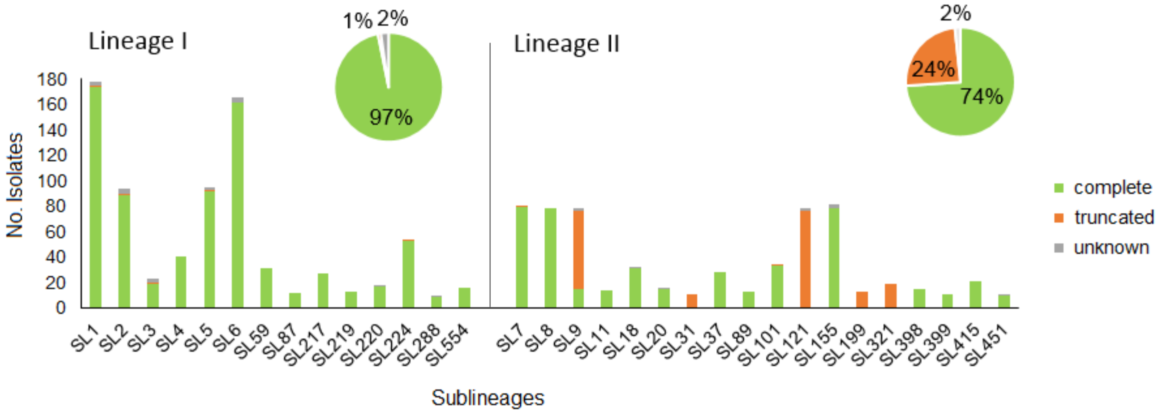
Supplementary Figure 10. Temporal analysis of major sublineages under the best evolutionary model (exponential-relaxed molecular clock with a coalescent exponential population model, Bayes factor >200). Lineage 1, red dots; lineage 2, orange dots. A) Substitution rates (dots: mean; range: 95% highest posterior density intervals) of SL1 and SL9 sublineages estimated from real (-real) and randomized tip dates (-rand) datasets. Real data-based estimations (colored dots) were performed using 3 independent runs. Estimations based on randomized tips datasets (grey squares) were performed using 10 independent randomized datasets. Stars denote statistical significance (*p*<0.0001, unpaired t-student test). B) Estimates of the time of the most recent common ancestor (TMRCA) of major sublineages in lineage I and II, each based on 3 independent runs (dots: mean; range: 95% highest posterior density intervals). The y-axis scale represents the dates (Common Era).

2.9. Determination of PCR-serogroups, virulence and resistance gene profiles from genomic sequences

PCR-serogroups were deduced *in silico* based on the detection of target regions (*lmo0737*, *lmo1118*, ORF2110, ORF2819, *prs*) defined by Doumith et al. 2004[36].

For virulence and resistance gene screening, a set of 76 genes was identified based on the literature [1,37,38] as well as public VFDB [39] and PATRIC [40] databases. Genes were scanned in BIGSdb-*Lm* using BLASTN algorithm, with minimum nucleotide identity of 70%, alignment length coverage of 70% and word size of 10. The virulome/resistome included genes involved in teichoic acid biosynthesis (*gltAB, tagB, gtcA*), genes located in the pathogenicity islands LIPI-1 (*prfA, plcA, hly, mpl, actA, plcB*), LIPI-3 (*llsAGHXBYDP*) and LIPI-4 (LM9005581_70009 to LM9005581_70014), genes coding for internalins (*inlABCEFGHJK*), and other genes involved in adherence (*ami, dltA, fbpA, lap, lapB*), invasion (*aut, aut_IVb, cwhA, lpeA, vip*), intracellular survival (*hpt, lplA1, oppA, prsA2, purQ, svpA*), regulation of transcription and translation (*agrAC, cheAY, fur, lisKR, rsbV, sigB, stp, virRS*), surface protein anchoring (*lgt, lspA, srtAB*), peptidoglycan modification (*oatA, pdgA),* immune modulation *(lntA),* bile resistance *(bsh, brtA, mdrM, mdrT),* resistance to detergents (*bcrABC, ermE, qac*) and biofilm formation and virulence *(comK).* The association between gene presence and phylogenetic lineages was accessed using Fisher's exact test.[41]

Results on the virulence/resistance profiles are presented in Figure 6 and discussed in the main text. A detailed breakdown of the distribution of *inlA* variants per sublineage is shown in **Supplementary Figure 11.** The details on the 22 different *inlA* mutations and frameshifts leading to premature internal stop codons identified in 200 isolates (out of 1,696; 12%) are listed in **Supplementary Table 8**.



Supplementary Figure 11. Distribution of *inlA* variants among frequent sublineages (≥10 isolates). Pie charts represent the overall proportion of *inlA* variants in lineage I (*n*=884 isolates) and II (*n*=783).

3. Supplementary References

1. Maury, M. *et al.* Uncovering *Listeria monocytogenes* hypervirulence by harnessing its biodiversity. *Nat. Genet.* **48,** 308–313 (2016).
2. Altschul, S., Gish, W. & Miller, W. Basic Local Alignment Search Tool. *J. Mol. Biol.* 403–410 (1990). doi:10.1016/S0022-2836(05)80360-2
3. Jolley, K. A. & Maiden, M. C. J. BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* **11,** 595 (2010).
4. Ragon, M. *et al.* A new perspective on *Listeria monocytogenes* evolution. *PLoS Pathog.* **4,** e1000146 (2008).
5. van Belkum, A. *et al.* Guidelines for the validation and application of typing methods for use in bacterial epidemiology. *Clin. Microbiol. Infect.* **13,** 1–46 (2007).
6. Simpson, E. H. Measurement of diversity. *Nature* **163,** 688–688 (1949).
7. Nei, M. & Li, W. H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. U. S. A.* **76,** 5269–5273 (1979).
8. Watterson, G. A. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7,** 256–276 (1975).
9. Librado, P. & Rozas, J. DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25,** 1451–1452 (2009).
10. Yang, Z. & Nielsen, R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* **17,** 32–43 (2000).
11. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24,** 1586–1591 (2007).
12. Bruen, T. C., Philippe, H. & Bryant, D. A simple and robust statistical test for detecting the presence of recombination. *Genetics* **172,** 2665–81 (2006).
13. Ruppitsch, W. *et al.* Defining and Evaluating a core genome MLST scheme for whole genome sequence-based typing of *Listeria monocytogenes*. *J. Clin. Microbiol.* (2015). doi:10.1128/JCM.01193-15
14. Pightling, A. W., Petronella, N. & Pagotto, F. The *Listeria monocytogenes* core-genome sequence typer (LmCGST): a bioinformatic pipeline for molecular characterization with next-generation sequence data. *BMC Microbiol.* **15,** 224 (2015).
15. Glaser, P. *et al.* Comparative genomics of *Listeria* species. *Science (80-.).* **294,** 849–852 (2001).
16. Chevreux, B., Wetter, T. & Suhai, S. Genome sequence assembly using trace signals and additional sequence information. *Comput. Sci. Biol. Proc. Ger. Conf. Bioinforma.* 45–56. (1999). doi:10.1.1.23/7465
17. Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1,** 18 (2012).
18. Zerbino, D. R. & Birney, E. Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* **18,** 821–829 (2008).
19. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19,** 455–477 (2012).
20. Koren, S., Treangen, T. J., Hill, C. M., Pop, M. & Phillippy, A. M. Automated ensemble assembly and validation of microbial genomes. *BMC Bioinformatics* **15,** 126 (2014).
21. Criscuolo, A. & Brisse, S. AlienTrimmer: A tool to quickly and accurately trim off multiple short contaminant sequences from high-throughput sequencing reads. *Genomics* **102,** 500–506 (2013).
22. Severiano, A., Pinto, F. R., Ramirez, M. & Carriço, J. a. Adjusted Wallace coefficient as a measure of congruence between typing methods. *J. Clin. Microbiol.* **49,** 3997–4000 (2011).
23. Carriço, J. a. *et al.* Illustration of a common framework for relating multiple typing methods by application to macrolide-resistant *Streptococcus pyogenes*. *J. Clin. Microbiol.* **44,** 2524–2532 (2006).
24. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30,** 772–80 (2013).
25. Desper, R. & Gascuel, O. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *J. Comput. Biol.* **9,** 687–705 (2002).
26. Didelot, X., Achtman, M., Parkhill, J., Thomson, N. R. & Falush, D. A bimodal pattern of relatedness between the *Salmonella* Paratyphi A and Typhi genomes: Convergence or divergence by homologous recombination? *Genome Res.* **17,** 61–68 (2007).
27. Huson, D. H. & Scornavacca, C. Dendroscope 3: An interactive tool for rooted phylogenetic trees

and networks. *Syst. Biol.* **61,** 1061–1067 (2012).

28. Berger-Wolf, T. Online consensus and agreement of phylogenetic trees. *Lect. Notes Comput. Sci.* **3240,** 350–361 (2004).
29. den Bakker, H. C., Didelot, X., Fortes, E. D., Nightingale, K. K. & Wiedmann, M. Lineage specific recombination rates and microevolution in *Listeria monocytogenes*. *BMC Evol. Biol.* **8,** 277 (2008).
30. Dunn, J. C. Well-separated clusters and optimal fuzzy partitions. *J. Cybern.* **4,** 95–104 (1974).
31. Haase, J. K., Murphy, R. A., Choudhury, K. R. & Achtman, M. Revival of Seeliger's historical 'Special *Listeria* Culture Collection'. *Environ. Microbiol.* **13,** 3163–71 (2011).
32. Bouckaert, R. *et al.* BEAST 2: A software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **10,** 1–6 (2014).
33. Croucher, N. J. *et al.* Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.* **43,** e15 (2015).
34. Lapierre, M., Blin, C., Lambert, A., Achaz, G. & Rocha, E. P. C. The impact of selection, gene conversion, and biased sampling on the assessment of microbial demography. *Mol. Biol. Evol.* **33,** 1711–1725 (2016).
35. Firth, C. *et al.* Using time-structured data to estimate evolutionary rates of double-stranded DNA viruses. *Mol. Biol. Evol.* **27,** 2038–2051 (2010).
36. Doumith, M., Buchrieser, C., Glaser, P., Jacquet, C. & Martin, P. Differentiation of the major *Listeria monocytogenes* serovars by multiplex PCR. *J. Clin. Microbiol.* **42,** 3819–3822 (2004).
37. Kuenne, C. *et al.* Reassessment of the *Listeria monocytogenes* pan-genome reveals dynamic integration hotspots and mobile genetic elements as major components of the accessory genome. *BMC Genomics* **14,** 47 (2013).
38. Gahan, C. G. M. & Hill, C. *Listeria monocytogenes*: survival and adaptation in the gastrointestinal tract. *Front. Cell. Infect. Microbiol.* **4,** 9 (2014).
39. Chen, L. *et al.* VFDB: A reference database for bacterial virulence factors. *Nucleic Acids Res.* **33,** 325–328 (2005).
40. Wattam, A. R. *et al.* PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res.* **42,** 581–591 (2014).
41. Agresti, A. A survey of exact inference for contingency tables. *Stat. Sci.* 131–153 (1992).