



**HAL**  
open science

## 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, et al.

► **To cite this version:**

Alexandra Moura, Alexis Criscuolo, Hannes Pouseele, Mylène M. Maury, Alexandre Leclercq, et al.. Whole genome-based population biology and epidemiological surveillance of *Listeria monocytogenes*. Nature Microbiology, 2016, 2, pp.16185. 10.1038/nmicrobiol.2016.185 . pasteur-01415883

**HAL Id: pasteur-01415883**

**<https://pasteur.hal.science/pasteur-01415883>**

Submitted on 13 Dec 2016

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - ShareAlike 4.0 International License

1 **Whole genome-based population biology**  
2 **and epidemiological surveillance**  
3 **of *Listeria monocytogenes***  
4  
5

6 Alexandra Moura<sup>1,2,3,4,5</sup>, Alexis Criscuolo<sup>6</sup>, Hannes Pouseele<sup>7</sup>, Mylène M. Maury<sup>1,2,3,4,5,8</sup>, Alexandre  
7 Leclercq<sup>1,2</sup>, Cheryl Tarr<sup>9</sup>, Jonas T. Björkman<sup>10</sup>, Timothy Dallman<sup>11</sup>, Aleisha Reimer<sup>12</sup>, Vincent  
8 Enouf<sup>13</sup>, Elise Larssonneur<sup>4,6,14</sup>, Heather Carleton<sup>9</sup>, Hélène Bracq-Dieye<sup>1,2</sup>, Lee S. Katz<sup>9</sup>, Louis Jones<sup>6</sup>,  
9 Marie Touchon<sup>4,5</sup>, Mathieu Tourdjman<sup>15</sup>, Matthew Walker<sup>12</sup>, Steven Stroika<sup>9</sup>, Thomas Cantinelli<sup>1</sup>,  
10 Viviane Chenal-Francisque<sup>1</sup>, Zuzana Kucerova<sup>9</sup>, Eduardo P. C. Rocha<sup>4,5</sup>, Celine Nadon<sup>12</sup>, Kathie  
11 Grant<sup>11</sup>, Eva M. Nielsen<sup>10</sup>, Bruno Pot<sup>7</sup>, Peter Gerner-Smidt<sup>9</sup>, Marc Lecuit<sup>1,2,3,16,\*</sup>, Sylvain Brisse<sup>4,5,\*</sup>

12  
13 <sup>1</sup> Institut Pasteur, National Reference Centre and World Health Organization Collaborating Center  
14 for *Listeria*, Paris, France;

15 <sup>2</sup> Institut Pasteur, Biology of Infection Unit, Paris, France;

16 <sup>3</sup> Inserm U1117, Paris France;

17 <sup>4</sup> Institut Pasteur, Microbial Evolutionary Genomics Unit, Paris, France;

18 <sup>5</sup> CNRS, UMR 3525, Paris, France;

19 <sup>6</sup> Institut Pasteur – Hub Bioinformatique et Biostatistique – C3BI, USR 3756 IP CNRS – Paris, France;

20 <sup>7</sup> Applied-Maths, Sint-Martens-Latem, Belgium;

21 <sup>8</sup> Paris Diderot University, Sorbonne Paris Cité, Cellule Pasteur, Paris, France;

22 <sup>9</sup> Centers for Disease Control and Prevention, Atlanta, Georgia, United States;

23 <sup>10</sup> Statens Serum Institut, Copenhagen, Denmark;

24 <sup>11</sup> Public Health England, London, United Kingdom;

25 <sup>12</sup> Public Health Agency of Canada, Winnipeg, Canada;

26 <sup>13</sup> Institut Pasteur, Pasteur International Bioresources network (PIBnet), Mutualized Microbiology  
27 Platform (P2M), Paris, France;

28 <sup>14</sup> CNRS, UMS 3601 IFB-Core, Gif-sur-Yvette, France;

29 <sup>15</sup> Public Health France, Saint-Maurice, France;

30 <sup>16</sup> Paris Descartes University, Sorbonne Paris Cité, Institut Imagine, Necker-Enfants Malades  
31 University Hospital, Division of Infectious Diseases and Tropical Medicine, APHP, Paris, France;

32 \* Correspondence to [sylvain.brisse@pasteur.fr](mailto:sylvain.brisse@pasteur.fr) and [marc.lecuit@pasteur.fr](mailto:marc.lecuit@pasteur.fr)  
33

34

35 **Abstract**

36 *Listeria monocytogenes* (*Lm*) is a major human foodborne pathogen. Numerous *Lm*  
37 outbreaks have been reported worldwide, associated with high case fatality rate,  
38 reinforcing the need for strongly coordinated surveillance and outbreak control. We  
39 developed a universally applicable genome-wide strain genotyping approach and  
40 investigated the population diversity of *Lm* using 1,696 isolates from diverse  
41 sources and geographical locations. We define, with unprecedented precision, the  
42 population structure of *Lm*, demonstrate the occurrence of international circulation  
43 of strains, and reveal the extent of heterogeneity in virulence and stress resistance  
44 genomic features among clinical and food isolates. Using historical isolates, we show  
45 that the evolutionary rate of *Lm* from lineage I and lineage II is low ( $\sim 2.5 \times 10^{-7}$   
46 substitutions per site per year, as inferred from the core genome) and that major  
47 sublineages (corresponding to so-called 'epidemic clones') are estimated to be at  
48 least 50 to 150 years old. This work demonstrates the urgent need of monitoring *Lm*  
49 strains at the global level and provides the unified approach needed for global  
50 harmonization of *Lm* genome-based typing and population biology.

51

## 52 **Introduction**

53 Pathogens know no border and can cause multi-country outbreaks and  
54 pandemics,<sup>1,2</sup> emphasizing the importance of international coordination for  
55 infectious diseases surveillance.<sup>3</sup> Microbiological surveillance programs rely on the  
56 continuous monitoring of circulating genotypes in space and time, enabling the  
57 rapid detection of common-source clusters and the implementation of control  
58 measures.<sup>4</sup> Despite outstanding exceptions,<sup>5-7</sup> most pathogens are so far  
59 monitored only at the national level. The lack of international coordination implies  
60 that outbreaks affecting multiple countries are either not detected or not  
61 controlled optimally.<sup>3,4</sup> International and cross-sector surveillance of pathogens  
62 requires strain subtyping methods that combine high resolution, reproducibility  
63 and exchangeability, so that epidemiologically relevant groups of matching isolates  
64 can be rapidly recognized across space and time.<sup>4</sup> Besides, harmonized and  
65 universally shared strain nomenclatures, which must be rooted in the  
66 microorganism population biology, are a prerequisite for rapid detection and  
67 efficient communication on emerging strain types.

68 The foodborne pathogen *Listeria monocytogenes* (*Lm*) causes listeriosis, a human  
69 systemic infection characterized by septicemia, central nervous system and  
70 maternal-fetal invasion, with high hospitalization and fatality rates.<sup>8</sup> Less severe  
71 manifestations include gastroenteritis<sup>9</sup> and may often remain undiagnosed. In the  
72 PulseNet program,<sup>6</sup> the microbiological typing golden standard, pulsed-field gel  
73 electrophoresis (PFGE), has been standardized internationally, but naming of  
74 profiles is not coordinated between the different international PulseNet networks.  
75 Furthermore, PFGE does not reflect evolutionary relationships and certain profiles

76 are highly prevalent leading to insufficient discriminative power. In contrast,  
77 multilocus sequence typing (MLST) based on seven genes provides highly  
78 standardized genotypes and nomenclature,<sup>10-12</sup> but lacks the discriminatory power  
79 required for epidemiological surveillance of most bacterial pathogens. Advances in  
80 high-throughput sequencing technologies have established whole genome  
81 sequencing (WGS) as a powerful epidemiological typing tool<sup>1,13,14</sup> that has been  
82 applied to investigate outbreaks and *Lm* contamination of food production  
83 plants.<sup>15-20</sup> However, these studies were restricted to local or national levels and a  
84 relatively small number of isolates.

85 Wide-range transmission of *Lm* strains can occur through international food  
86 trade<sup>21</sup> and the major MLST-defined clonal complexes (CCs) of *Lm* are distributed  
87 globally.<sup>11</sup> However, the rate of evolution of *Lm* genomes and the speed at which  
88 strains can spread over large distances are currently unknown. Further, a global  
89 view of the relationships between genotype and virulence potential of *Lm* strains  
90 remains to be established.

91 To enable population biology studies of global *Lm* collections and for prospective  
92 international epidemiological surveillance, a harmonized protocol to translate  
93 genomic sequence into its corresponding nomenclatural genotype needs to be  
94 established. Although single-nucleotide polymorphism (SNP)-based approaches  
95 can provide maximal discrimination,<sup>20</sup> they are difficult to standardize and can be  
96 difficult to interpret.<sup>18,22</sup> In contrast, genome-wide MLST approaches rely on well-  
97 defined standard sets of hundreds of genes that can be validated *a priori* for strain  
98 genotyping.<sup>14,19,23,24</sup>

99 Here we developed a core genome MLST (cgMLST) method for *Lm* and applied it to  
100 a large number of strains from a wide spectrum of geographic, temporal and  
101 epidemiological origins. This enabled us to decipher the population structure and  
102 evolutionary rate of *Lm*, to demonstrate international transmission of major  
103 sublineages, and to develop a unified genome-based nomenclature of *Lm* strains  
104 accessible through an open bioinformatics platform, allowing international  
105 collaboration on research and public health surveillance based on high-throughput  
106 genome sequencing.

107

## 108 **Results**

### 109 ***Universal Lm cgMLST***

110 A core genome MLST (cgMLST) scheme of 1,748 loci was defined based on a high  
111 level of conservation of this set of genes among 957 genomes of diverse origins  
112 (**Supplementary Information 2.1**). Using an independent set of 650  
113 prospectively collected isolates to estimated typeability<sup>4</sup>, each of these genes could  
114 be detected in 644 genomes (99.1%) on average, resulting in half of the genomes  
115 having 8 or fewer uncalled alleles (average±standard deviation of uncalled alleles  
116 15±20; **Supplementary Information 2.1**). These results demonstrate the  
117 universal applicability of this cgMLST scheme for *Lm* strain genotyping.

118 Reproducibility of allele calls based on genomic sequences obtained from  
119 independent cultures and sequencing protocols of the reference EGD-e strain was  
120 absolute (error rate <0.029%, *i.e.* <1 error in 3,496 allelic comparisons). cgMLST  
121 genotyping was also reproducible irrespective of assembly pipeline for coverage  
122 depths ≥40 (with per-site Phred quality score ≥20, *i.e.* corresponding to ≥99% base  
123 accuracy) and *de novo* assembly allele calls were identical to assembly-free  
124 methods (**Supplementary Information 2.4**). Altogether, the cgMLST scheme  
125 developed herein constitutes an extremely robust genotyping method, even when  
126 applied on a very wide variety of *Lm* strains sequenced from diverse sources and  
127 geographical locations.

128

### 129 ***Definition of cgMLST types***

130 To provide a definition of cgMLST types (CTs) that would be maximally useful for  
131 surveillance purposes, we compared the genetic heterogeneity between

132 epidemiologically related isolates on the one hand, and between isolates with no  
133 documented epidemiological link on the other hand. Pairwise allelic mismatches  
134 revealed two distinct distributions (**Fig. 1A**). First, most isolates sampled during  
135 investigations of single outbreaks had seven or fewer allelic mismatches (**Fig. 1A**).  
136 Among these, pairs of isolates from vertical maternal-neonatal transmission cases  
137 had no allelic differences (not shown). Second, taking into account the entire  
138 dataset (**Fig. 1A**), a sharp discontinuity was observed, with few pairs of isolates  
139 having between 7 and 10 allelic mismatches, showing that isolates with no  
140 documented epidemiological link differed most generally by more than 10  
141 mismatches. Clustering efficiency was optimal when using a cut-off value of 7.3  
142 allelic mismatches (*i.e.* 0.414% of mismatched loci; **Supplementary Information**  
143 **2.7**). Therefore, we propose to define CTs as groups of cgMLST profiles that differ  
144 by up to 7 allelic mismatches out of 1,748 loci (*i.e.*, in case of uncalled alleles,  
145 0.400% of mismatched loci among those that are called in both profiles), from at  
146 least one other member of the group.

147

#### 148 ***Comparison of cgMLST and PFGE genotyping***

149 PFGE is the current reference method for *Lm* epidemiological surveillance and  
150 outbreak investigation.<sup>6</sup> Among the 100 *Lm* isolates used for cgMLST and PFGE  
151 comparison, only 36 distinct *AscI-ApaI* combined PFGE profiles (Simpson's  
152 diversity index = 0.944, 95% confidence interval CI = [0.926, 0.963]) were  
153 identified, whereas cgMLST distinguished 68 CTs (Simpson's index = 0.987; 95% CI  
154 = [0.981, 0.994]). This indicates that cgMLST greatly improves discrimination  
155 among *Lm* isolates as compared with PFGE ( $p < 0.001$ ; **Supplementary**

156 **Information 2.5).** Consistent with this, PFGE did not subtype any CT, whereas  
157 multiple PFGE types could be subdivided using cgMLST (adjusted Wallace index of  
158 concordance = 0.215; 95% CI = [0.156, 0.304]). Retrospective analysis indicated no  
159 epidemiological link among isolates that were grouped by PFGE but not by cgMLST  
160 (NRC and InVS, France). These results are consistent with previous work that  
161 reported improved discrimination of genome sequence typing over PFGE,<sup>18,19</sup> and  
162 our collective unpublished experience covering more than one year with WGS for  
163 real-time surveillance of listeriosis in Denmark, France, the United Kingdom and  
164 the United States. Implementation of cgMLST in *Lm* surveillance therefore shows  
165 great promise to improve the definition of clusters of cases, thus facilitating  
166 investigations of contamination sources.

167

### 168 ***Phylogenetic structure and nomenclature of *Lm* sublineages***

169 A unified nomenclature of *Lm* subtypes is critically needed for real-time exchange  
170 of information on the emergence and geographic dispersal of strains. To provide  
171 an optimized subtype definition, we analyzed the phylogenetic structure of *Lm*.  
172 The four major phylogenetic lineages of *Lm* were clearly separated (**Fig. 2A**).  
173 cgMLST-based clustering of isolates into lineages and their sublineages was highly  
174 concordant with the sequence-based phylogenetic tree (**Fig. 2B**). Whereas the  
175 strains of lineages III and IV (which are rarely isolated in the context of  
176 surveillance) were scattered into multiple rare sublineages, lineages I and II were  
177 strongly structured into major sublineages, each comprising multiple closely-  
178 related isolates (**Fig. 3A**). Two atypically divergent sublineages within lineage II

179 were identified (sublineages SL842 and SL843, **Fig. 2A**), showing that lineage II is  
180 more diverse than previously reported.

181 The observed trimodal distribution of allelic mismatches among all pairs of  
182 isolates (**Fig. 1A**) was consistent with phylogenetic structure: isolates belonging to  
183 distinct major phylogenetic lineages differed by 1,500 loci or more out of 1,748  
184 loci, isolates from different sublineages within a given lineage typically showed  
185 between 1,000 and 1,400 allelic differences, and most isolates within the same  
186 sublineage were up to 150 allelic mismatches distant. Moreover, clustering  
187 efficiency was optimal between 140 and 150 allelic mismatches (**Supplementary**  
188 **Information 2.7**). Therefore, a threshold of 150 allelic mismatches (8.58%  
189 dissimilarity) was chosen to define sublineages. This cut-off value led to the  
190 identification of 163 sublineages. Remarkably, the flat rarefaction curve obtained  
191 for sublineages within lineages I and II suggests that this study has captured most  
192 of the phylogenetic sublineages of these two epidemiologically major lineages (**Fig.**  
193 **1B**). In contrast, the almost linear rarefaction curve of CT richness indicates that  
194 the 1,013 CTs sampled represent only a small fraction of those expected to be  
195 uncovered upon further sampling (**Fig. 1B**), underlining the fine subtyping power  
196 of cgMLST and its ability to subdivide *Lm* biodiversity into a multitude of  
197 epidemiologically relevant genotypic groups.

198 We next analyzed the correspondence of sublineages with classical 7-genes MLST  
199 nomenclature.<sup>10-12</sup> Whereas 156 sequence types (STs) have been previously  
200 defined in the Institut Pasteur MLST database (now in BIGSdb-*Lm*,  
201 <http://bigsdb.pasteur.fr/listeria>), 63 new ones were identified, revealing a  
202 significant amount of novel diversity of *Lm* strains. MLST-defined CCs were

203 mapped onto the cgMLST-based phylogenetic structure (**Supplementary**  
204 **Information 2.7**), largely revealing a one-to-one correspondence with cgMLST  
205 sublineages. Therefore, the MLST nomenclature was mapped onto sublineages  
206 where possible (**Supplementary Table 4**). As expected, frequent sublineages  
207 corresponded to previously recognized major MLST clones.<sup>10-12,25,30</sup> As a result, the  
208 sublineage cgMLST-based nomenclature can be easily matched with the widely  
209 used MLST nomenclature, which remains a valuable tool for first line identification  
210 of sublineages.<sup>31</sup>

211

### 212 ***Evidence for international spread of *Lm* strains***

213 To investigate international transmission of *Lm* strains, we first mapped the  
214 geographic origin of isolates onto the phylogeny (**Fig. 3A**). All sublineages  
215 represented by more than 50 isolates were recovered from at least four distinct  
216 countries (**Fig. 3B**). Using a stochastic mapping approach to reconstruct ancestral  
217 states, we estimated the average number of cross-country transmission events as  
218 ranging from 13 to 48 in the 10 most frequent sublineages (**Fig. 3C and 3D**). These  
219 results show that subsequent to the evolutionary origin of major sublineages,  
220 geographical shifts have occurred repeatedly. When normalizing the number of  
221 geographical transitions by taking into account the number of isolates and  
222 evolutionary time (jointly approximated by total tree length), large differences in  
223 cross-country transition rates were apparent (**Fig. 3D**). Interestingly, the most  
224 food-associated sublineages SL9 and SL121<sup>25</sup> had among the highest geographical  
225 transition rates.

226 To investigate international transmission at a more recent epidemiological time-  
227 scale, we searched for internationally distributed CTs. Interestingly 9 CTs,  
228 comprising a total of 34 isolates, included isolates from at least two countries (**Fig.**  
229 **4, Supplementary Table 6**). These results demonstrate the international  
230 distribution of genotypic groups of *Lm* isolates that exhibit levels of genetic  
231 divergence typical of those observed within documented outbreaks and  
232 transmission events.

233

#### 234 ***Temporal accumulation of variation within *Lm* lineages and outbreaks***

235 Phylogenetic analysis of the most prevalent sublineage (SL1) (**Supplementary**  
236 **Table 5**) showed that the root-to-tip distances were significantly associated  
237 ( $p < 0.0001$ , F-test) with the isolation year of isolates (**Fig. 5B**). The inferred slope of  
238 the linear regression indicated an accumulation of 0.23 allelic mismatches per  
239 cgMLST profile (i.e., 1.58 Mb) per year. BEAST analysis of the concatenated  
240 multiple sequence alignments confirmed the existence of a temporal signal  
241 (**Supplementary Information 2.8**) and estimated an evolutionary rate of  $2.6 \times 10^{-7}$   
242 subst/site/year (0.41 subst/1.58Mb/year), i. e. 1 substitution on the core genome  
243 every 2.5 years (95% HPD=[1.9-3.4]). We also estimated independently the rate of  
244 SL9, as a representative of major lineage II. Remarkably, the SL9 rate was  $2.4 \times 10^{-7}$   
245 subst/site/year (0.38 subst/1.58Mb/year), indicating a highly similar rate in SL1  
246 and SL9. These results demonstrate measurable evolution of *Lm* genomes over a  
247 few decades and provide an estimate of the short-term rate of accumulation of  
248 genetic variation in representative sublineages of the two major lineages of *Lm*.  
249 Based on the hypothesis that the substitution rate is conserved in *Lm*, we

250 estimated that the root of the other major sublineages was 50 to 150 years old  
251 (**Supplementary Information 2.8**). Note that these estimates must be taken with  
252 care: the rate may vary in some sublineages, and it is likely that our sampling has  
253 missed some divergent branches, implying that our estimates are minimal ages.  
254 Nevertheless, our current age estimates suggest an expansion of major sublineages  
255 in modern times. Whether the dissemination of *Lm* was driven by an increase in  
256 the intensity of exchange of people, animals and food in recent times is an  
257 intriguing possibility.

258 *Lm* can survive for long periods of time in various sources, where genetic  
259 diversification from a single population founder can occur.<sup>15,20,32</sup> Consistently, we  
260 observed that allelic divergence within outbreak sets and international clusters  
261 was positively associated with the time span between the first and last isolate  
262 collected ( $p < 0.05$ , F-test; **Fig. 5C**), with an accumulation of 0.28 allelic mismatch  
263 per year, highly consistent with the phylogenetic tree-based evolutionary rate  
264 estimate. These results illustrate the importance and possibility of taking the  
265 temporal dimension into account when interpreting genomic data in the context of  
266 persistent contaminations.<sup>20</sup>

267 The phylogeny of SL1 (**Fig. 5A**) showed that outbreaks strains were dispersed in  
268 multiple branches, suggesting that all SL1 isolates have the potential to cause  
269 outbreaks. Moreover, it demonstrated that the multiple outbreaks caused by this  
270 sublineage, previously called 'epidemic clone ECI',<sup>30</sup> are actually independent  
271 epidemiological events. The most recent common ancestor of SL1 was estimated to  
272 have existed around 1876 (95% HPD=[1861-1891]), reinforcing the idea that  
273 extant SL1 isolates do not derive from a single recent epidemic.

274

275 ***Biological features of Lm sublineages and CTs***

276 Important genomic differences among sublineages are shown in **Fig. 6**. PCR-  
277 serogroup distribution across the phylogenetic tree was consistent with previous  
278 knowledge, with major PCR-serogroups being strong markers of the main divisions  
279 of *Lm* diversity.<sup>10,33,34</sup> In contrast, PCR-serogroup variant IVb-v1<sup>35</sup> was found in  
280 various branches. Likewise, and as expected, serogroup L<sup>35</sup> was present in lineages  
281 III and IV, but also in lineage I. These results underline that caution is needed when  
282 interpreting molecular serotyping data for *Lm* epidemiological purposes.

283 The screening for virulence and stress resistance genes showed important  
284 differences among *Lm* lineages and sublineages (**Fig. 6**). As expected, the major  
285 pathogenicity island LIPI-1 was highly conserved. A complete LIPI-3<sup>36</sup> was almost  
286 exclusively detected within lineage I. The recently described LIPI-4<sup>25</sup> was nearly  
287 universally present in SL4 and closely related sublineages (**Fig. 6**), and was also  
288 found in few other isolates, including in lineages III and IV. *inlA* alleles encoding  
289 truncated InlA variants, which are associated with hypovirulence,<sup>37</sup> were observed  
290 in most isolates of lineage II sublineages SL9, SL31, SL121, SL199 and SL321 (**Fig.**  
291 **6 and Supplementary Information 2.9**) and were significantly associated with  
292 food and food-production isolates ( $p < 0.0001$ ). The presence of a non-disrupted  
293 form of the *comK* gene, involved in intracellular survival switch and biofilm  
294 formation<sup>38,39</sup>, was dispersed across multiple sublineages and far more frequent in  
295 lineage I than in lineage II (79% vs 38%, respectively,  $p < 0.0001$ , Fisher's exact  
296 test). Finally, genes that confer resistance to benzalkonium chloride, a major  
297 disinfectant applied on food-industry surfaces,<sup>40</sup> were significantly associated

298 ( $p < 0.0001$ , Fisher's exact test) with lineage II and particularly frequent in SL121,  
299 consistent with persistence of this clonal complex in food processing plants.<sup>41</sup>  
300 Taken together, these results demonstrate the strong heterogeneity among *Lm*  
301 sublineages with regards to genomic features involved in either pathogenesis or  
302 food contamination.  
303

## 304 **Discussion**

305 Listeriosis surveillance is currently organized almost exclusively at national levels,  
306 thereby limiting our capacity to trace sources of infections involving international  
307 transmission through food trade or human travel. An efficient global laboratory  
308 surveillance system would consist of three parts: standardized methods and  
309 databases, open sharing of data between public health laboratories, and rapid  
310 communication about outbreaks. Here, we have addressed these issues by  
311 developing a genome-wide genotyping system validated on a large international  
312 collection of *Lm* strains. Furthermore, we have set up an openly accessible  
313 database and analysis tool (BIGSdb-*Lm* at <http://bigsdbs.pasteur.fr/listeria>), which  
314 provides a unified nomenclature that will ease global communication on *Lm*  
315 genotypes. Real-time incorporation of genotypic variation of future *Lm* isolates  
316 uncovered through prospective genomic surveillance will enable global  
317 coordination of epidemiological surveillance.

318

319 Although alternative sets of cgMLST loci (**Supplementary Information 2.3**) have  
320 been recently proposed for *Lm* typing,<sup>19,24</sup> the scheme developed here in the  
321 context of a global collaboration contains more genes, was validated using isolates  
322 from diverse origins, and was extremely reproducible when comparing the results  
323 from independent allele calling approaches. We also show that cgMLST has a far  
324 greater discriminative power than PFGE when applied to the prospective  
325 surveillance of isolates. It is worth noting that although we already identified 1,013  
326 CTs, they represent only a small fraction of existing CTs that will be revealed by  
327 future genomic surveillance (**Fig. 1B**), indicating that referenced CTs should

328 rapidly surpass the number of PFGE types distinguished during 20 years of  
329 PulseNet surveillance (4,119 unique *ApaI/AscI* combinations among 21,158  
330 isolates with PFGE as of December 21, 2015). The largely improved refinement of  
331 *Lm* genotyping using cgMLST is expected to (i) reduce in size clusters accurately  
332 detected by PFGE, (ii) erase clusters falsely inferred from PFGE, and (iii) allow  
333 detecting earlier, clusters that would likely remain ignored when belonging to  
334 hyper-prevalent PFGE profiles. Together, these highly significant improvements of  
335 *Lm* typing will strongly reduce and even eliminate unnecessary epidemiological  
336 investigations, which is a major drawback of the lack of discrimination of the  
337 current standard PFGE, and will help to identify the food source of clusters of  
338 human cases by refining the definition of cases in case-control studies.

339

340 The analysis of a large and geographically diverse collection of *Lm* genomes also  
341 allowed us to determine the population structure of this species with  
342 unprecedented precision. The sharp discontinuities observed within the  
343 phylogenetic diversity of *Lm* allowed to identify and define sublineages  
344 unambiguously, which will constitute the basis of a universal genome-based  
345 nomenclature. This nomenclature has the advantage of being congruent with the  
346 previously widely adopted 7-genes MLST nomenclature and the corresponding  
347 major clinically- and food-associated CCs.<sup>10,25</sup> In addition, by including a large  
348 number of sets of epidemiologically related isolates, we could also define cgMLST  
349 types relevant for epidemiological purposes using a statistically optimized cut-off.  
350 As cgMLST dissimilarity is highly congruent with phylogenetic relationships, *Lm*  
351 strains can be assigned with high confidence to sublineages and types based on

352 their cgMLST profile. Because this does not require a multiple sequence alignment  
353 step, this approach is considerably faster than sequence-based identification, and  
354 easier to interpret by microbiologists, epidemiologists and public health  
355 professionals. Thus, CT classification is poised to become a universal tool for  
356 cluster detection and international communication during regional or global *Lm*  
357 outbreaks.

358

359 By applying genomic sequencing to a large collection of *Lm* isolates from diverse  
360 geographic origins, we were able to clearly demonstrate repeated international  
361 transmission of multiple sublineages of *Lm*. Further, we identified international  
362 groups of genetically highly related isolates, suggestive of recent cross-country and  
363 intercontinental transmissions. These results provide a unique population-level  
364 based snapshot of *Lm* international transmission and suggest that cross-country  
365 outbreaks that were recognized up to now<sup>17,42</sup> are only the tip of the iceberg of *Lm*  
366 long distance dissemination.<sup>43</sup> Given the retrospective nature of our analyses, we  
367 were not able to identify the epidemiological links among isolates of these  
368 international clusters, but these observations suggest that their detection in real  
369 time would allow tracing back to common sources, and firmly establish the  
370 importance of monitoring in real time the diffusion of *Lm* genotypes at the  
371 international level. The cgMLST collaborative approach here developed makes this  
372 goal achievable and paves the way for future research aimed at better  
373 understanding the routes and contributing factors of *Lm* dissemination.

374

375 We calibrated the short-term evolutionary rate of *Lm* genomes, and could  
376 therefore provide a quantitative estimate of the widely recognized view that *Lm*  
377 genomes are highly stable.<sup>12,44,45</sup> Because cgMLST types diversify slowly (roughly  
378 0.2 alleles per year), greater discrimination may be needed to decipher short-term  
379 patterns of transmission.<sup>17,18</sup> Therefore, to fully harness the power of genomic  
380 sequencing for *Lm* epidemiology, multi-approach strategies can be applied,  
381 including the use of pan-genomic MLST and reference-based SNP-calling. However,  
382 in contrast to MLST, genome-wide SNP-based approaches do not rely on  
383 predefined genomic loci and require *ad-hoc* reference sequences, thus being more  
384 complex to standardize. In this context, the genotyping method and publicly shared  
385 nomenclature developed herein will represent a pivot element of collaborative  
386 approaches to control the burden of *Lm* infections at the global scale.

387

388 **Methods**

389 ***Bacterial isolates***

390 A total of 1,696 *Lm* genomes were included in the main dataset (1,055 human  
391 isolates, 475 isolates from food and food-processing environments, and 166  
392 isolates from other or unknown sources; **Supplementary Table 1;**  
393 **Supplementary Figure 1**), comprising isolates collected between 1960 and 2015,  
394 mostly from North America and Europe. This set included the 104 genomes  
395 representative of the clonal diversity of *Lm* used for core genome definition,<sup>25</sup>  
396 genomic sequences from isolates collected in the context of *Lm* surveillance  
397 programs in Canada ( $n=36$  isolates), Denmark ( $n=224$ ), France ( $n=112$ ), the United  
398 Kingdom ( $n=448$ ) and the United States ( $n=758$ ), and 14 genomes from a German-  
399 Austrian outbreak.<sup>17</sup> This collection included (i) prospectively collected isolates as  
400 well as (ii) isolates collected in the frame of outbreak investigations or mother-  
401 child transmission cases (**Supplementary Table 1; Supplementary Figure 1**). In  
402 addition, 34 historical isolates (**Supplementary Table 5**), were included for  
403 analysis of *Lm* evolutionary rates. DNA extraction, library preparation and Illumina  
404 sequencing using MiSeq, NextSeq or HiSeq instruments were performed locally in  
405 each reference center. Sequence assembly was performed using BioNumerics v.7.5  
406 (Applied Maths NV, Sint-Martens-Latem, Belgium) or CLC Assembly Cell 4.3.0  
407 (Qiagen, Aarhus, Denmark). Provenance data and genomic assembly details of the  
408 1,696 isolates are listed in **Supplementary Table 1**.

409

#### 410 ***Validation of a universal cgMLST scheme for Lm genotyping***

411 A previously defined *Lm* core genome with 1,791 loci<sup>25</sup> was further refined by  
412 removing genes present in less than 95% of 957 high-quality genome sequences  
413 (**Supplementary Figure 1**), genes with close paralogs and genes belonging to the  
414 seven MLST scheme (**Supplementary Information 2.1**). This filtering procedure  
415 led to a final subset of 1,748 core genes, here referred as the *Lm* cgMLST scheme  
416 (**Supplementary Tables 2 and 3**). The levels of diversity, selection and  
417 recombination were quantified for each cgMLST locus (**Supplementary**  
418 **Information 2.2**). The robustness of cgMLST genotyping was tested using both  
419 assembly-free and *de novo* assembly-based methods, to control that allelic profiles  
420 generated by the two approaches are consistent and to exclude potential assembly  
421 artefacts. The performance of different assemblers was also tested at different  
422 sequencing coverage depths (**Supplementary Information 2.4**).

423

#### 424 ***Comparison of cgMLST and PFGE genotyping***

425 To compare cgMLST with PFGE for *Lm* strain typing, we analyzed in parallel 100  
426 isolates (57 human isolates, 33 food isolates and 10 isolates from food production  
427 environments) prospectively collected between January and April 2015 in the  
428 frame of the French listeriosis surveillance system by the National Reference  
429 Center for *Listeria* (Institut Pasteur, France). PFGE restriction profiles were  
430 obtained using the enzymes *AscI* and *Apal* according to PulseNet standardized  
431 procedures (<http://www.cdc.gov/pulsenet/PDF/listeria-pfge-protocol-508c.pdf>)  
432 and were analyzed using BioNumerics. PFGE and cgMLST typing results were

433 compared using Simpson's index of diversity and the adjusted Wallace index of  
434 concordance (see **Supplementary Information 2.5** for details).

435

#### 436 ***Phylogenetic and clustering analyses***

437 The phylogenetic relationships of the 1,696 isolates were inferred based either on  
438 the allelic profiles or on the recombination-purged multiple sequence alignments  
439 of the 1,748 loci (see **Supplementary Information 2.6** for details). Single-linkage  
440 clustering analysis was performed from the *p*-distances among allelic profiles  
441 (cgMLST allelic distances, *i.e.* proportion of mismatched loci among those that are  
442 called in both strains). Clustering efficiency (*i.e.* optimizing both compactness  
443 within clusters and separateness among clusters) was assessed with Dunn's index  
444 (**Supplementary Information 2.7**) using different allelic mismatch thresholds.

445

#### 446 ***Phylogeography and temporal analysis***

447 Geographical transitions within major sublineages were inferred from FastME  
448 v.2.07 trees using discrete trait transition modelling based on 100 simulations with  
449 the make.simmap tool in the phytools R package.<sup>26,27</sup> Once the ancestral states  
450 were estimated, the total number of character changes was computed from the  
451 resulting set of trees, using the count.simmap function within the same R  
452 package.<sup>27</sup>

453 To estimate the evolutionary rate of sequences and cgMLST profiles, 22 historical  
454 isolates belonging to MLST clonal complex CC1 and 12 isolates from clonal  
455 complex CC9, collected between 1921 and 1974, were analyzed jointly with the  
456 isolates from sublineages SL1 and SL9 (see below) from the main dataset

457 **(Supplementary Information 2.8, Supplementary Tables 1 and 5).**  
458 Phylogenetic analyses were performed using FastME on p-distances estimated  
459 from either concatenated multiple sequence alignments or cgMLST profiles. Linear  
460 regression of the root-to-tip distances against the year of isolation was carried out  
461 using Path-O-Gen v1.4 (<http://tree.bio.ed.ac.uk/software/pathogen/>). The rate of  
462 evolution of SL1 and SL9 genomes were independently estimated from the  
463 concatenated multiple sequence alignments of the 1,748 loci using BEAST v.2.3.1<sup>28</sup>.  
464 For this analysis, Gubbins<sup>29</sup> was used to detect recombination within the  
465 alignments. Isolates with recombinant regions were discarded from the alignments  
466 **(Supplementary Information 2.8)**. Subsequently, the mean of the rates of SL1  
467 and SL9 ( $2.5 \times 10^{-7}$  substitutions per site per year) was used to estimate the age of  
468 all major sublineages using BEAST v.2.3.1<sup>28</sup>. Details of the temporal analysis  
469 methods are given in **Supplementary Information 2.8**. Genetic divergence as a  
470 function of the time span between the first and last isolate of outbreak sets and  
471 international clusters was evaluated using regression analysis.

472

#### 473 ***Determination of PCR serogroups, virulence and resistance genes profiles***

474 To investigate the biological differences among sublineages, the PCR-serogroup  
475 and the presence of 76 loci involved in virulence or resistance were deduced *in*  
476 *silico* from genomic sequences using the BIGSdb platform<sup>23</sup> for each of the 1,696  
477 genomes (see **Supplementary Information 2.9** for details).

478

#### 479 ***Online implementation of an open bioinformatics platform for Lm strain*** 480 ***nomenclature and genome analysis***

481 To make the cgMLST-based nomenclature sharable and expandable, the *Lm*  
482 cgMLST scheme was implemented in an integrative database and analysis platform  
483 (BIGSdb-*Lm*) powered by the BIGSdb v.1.10<sup>23</sup> bioinformatics tool. To unify *Lm*  
484 genotyping resources, the classical 7-gene MLST scheme was transferred into the  
485 BIGSdb-*Lm* platform. Openly accessible predefined schemes for molecular  
486 serogrouping and for virulence and resistance gene analyses were also  
487 incorporated in the BIGSdb-*Lm* platform. BIGSdb-*Lm* is publicly accessible at  
488 <http://bigsdbs.pasteur.fr/listeria>.

489

#### 490 ***Accession numbers***

491 FASTQ data files were deposited in NCBI-SRA and EBI-ENA public archives under  
492 the project's accession numbers PRJEB12738 (Institut Pasteur), PRJEB14476  
493 (Statens Serum Institut), PRJNA248549 (Public Health England) and PRJNA212117  
494 (Centers for Disease Control and Prevention). The accession numbers of all isolates  
495 are indicated in Supplementary Table 1.

496

497 **References**

- 498 1. Mutreja, A. *et al.* Evidence for several waves of global transmission in the  
 499 seventh cholera pandemic. *Nature* **477**, 462–465 (2011).
- 500 2. Grad, Y. H. *et al.* Genomic epidemiology of the *Escherichia coli* O104:H4  
 501 outbreaks in Europe, 2011. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 3065–3070  
 502 (2012).
- 503 3. Woolhouse, M. E. J., Rambaut, A. & Kellam, P. Lessons from Ebola: Improving  
 504 infectious disease surveillance to inform outbreak management. *Sci. Transl.*  
 505 *Med.* **7**, 1–9 (2015).
- 506 4. van Belkum, A. *et al.* Guidelines for the validation and application of typing  
 507 methods for use in bacterial epidemiology. *Clin. Microbiol. Infect.* **13**, 1–46  
 508 (2007).
- 509 5. Bogner, P., Capua, I., Cox, N. J., Lipman, D. J. & Others. A global initiative on  
 510 sharing avian flu data. *Nature* **442**, 981–981 (2006).
- 511 6. Gerner-Smidt, P. *et al.* PulseNet USA: a five-year update. *Foodborne Pathog.*  
 512 *Dis.* **3**, 9–19 (2006).
- 513 7. Grundmann, H. *et al.* Geographic distribution of *Staphylococcus aureus*  
 514 causing invasive infections in Europe: a molecular-epidemiological analysis.  
 515 *PLoS Med.* **7**, e1000215 (2010).
- 516 8. Control, E. C. for D. P. and. *Surveillance of seven priority food- and waterborne*  
 517 *diseases in the EU/EEA.* (ECDC, 2015).
- 518 9. Dalton, C. B. *et al.* An outbreak of gastroenteritis and fever due to *Listeria*  
 519 *monocytogenes* in milk. *N. Engl. J. Med.* **336**, 100–5 (1997).
- 520 10. Ragon, M. *et al.* A new perspective on *Listeria monocytogenes* evolution. *PLoS*  
 521 *Pathog.* **4**, e1000146 (2008).
- 522 11. Chenal-Francisque, V. *et al.* Worldwide distribution of major clones of  
 523 *Listeria monocytogenes*. *Emerg. Infect. Dis.* **17**, 1110–1112 (2011).
- 524 12. Haase, J. K., Didelot, X., Lecuit, M., Korkeala, H. & Achtman, M. The ubiquitous  
 525 nature of *Listeria monocytogenes* clones: a large-scale multilocus sequence  
 526 typing study. *Environ. Microbiol.* **16**, 405–16 (2014).
- 527 13. Harris, S. R. *et al.* Evolution of MRSA during hospital transmission and  
 528 intercontinental spread. *Science (80-. )*. **327**, 469–474 (2010).
- 529 14. Maiden, M. C. J. *et al.* MLST revisited: the gene-by-gene approach to bacterial  
 530 genomics. *Nat. Rev. Microbiol.* **11**, 728–36 (2013).
- 531 15. Orsi, R. H. *et al.* Short-term genome evolution of *Listeria monocytogenes* in a  
 532 non-controlled environment. *BMC Genomics* **9**, 539 (2008).
- 533 16. Bergholz, T. M. *et al.* Evolutionary relationships of outbreak-associated  
 534 *Listeria monocytogenes* strains of serotypes 1/2a and 1/2b determined by  
 535 whole genome sequencing. *Appl. Environ. Microbiol.* **82**, 928–938 (2015).
- 536 17. Schmid, D. *et al.* Whole genome sequencing as a tool to investigate a cluster  
 537 of seven cases of listeriosis in Austria and Germany, 2011–2013. *Clin.*  
 538 *Microbiol. Infect.* **20**, 431–436 (2014).
- 539 18. Kwong, J. C. *et al.* Prospective whole genome sequencing enhances national  
 540 surveillance of *Listeria monocytogenes*. *J. Clin. Microbiol.* **54**, 333–342 (2015).
- 541 19. Ruppitsch, W. *et al.* Defining and evaluating a core genome MLST scheme for  
 542 whole genome sequence-based typing of *Listeria monocytogenes*. *J. Clin.*

- 543 *Microbiol.* **53**, 2869–2876 (2015).
- 544 20. Stasiewicz, M. J., Oliver, H. F., Wiedmann, M. & den Bakker, H. C. Whole  
545 genome sequencing allows for improved identification of persistent *Listeria*  
546 *monocytogenes* in food associated environments. *Appl. Environ. Microbiol.*  
547 **81**, 6024–6037 (2015).
- 548 21. Fretz, R. *et al.* Update: Multinational listeriosis outbreak due to ‘quargel’, a  
549 sour milk curd cheese, caused by two different *L. monocytogenes* serotype  
550 1/2a strains, 2009-2010. *Eurosurveillance* **15**, 2–3 (2010).
- 551 22. Pightling, A. W., Petronella, N. & Pagotto, F. Choice of reference sequence and  
552 assembler for alignment of *Listeria monocytogenes* short-read sequence data  
553 greatly influences rates of error in SNP analyses. *PLoS One* **9**, e104579  
554 (2014).
- 555 23. Jolley, K. A. & Maiden, M. C. J. BIGSdb: Scalable analysis of bacterial genome  
556 variation at the population level. *BMC Bioinformatics* **11**, 595 (2010).
- 557 24. Pightling, A. W., Petronella, N. & Pagotto, F. The *Listeria monocytogenes* Core-  
558 Genome Sequence Typer (LmCGST): a bioinformatic pipeline for molecular  
559 characterization with next-generation sequence data. *BMC Microbiol.* **15**,  
560 224 (2015).
- 561 25. Maury, M. *et al.* Uncovering *Listeria monocytogenes* hypervirulence by  
562 harnessing its biodiversity. *Nat. Genet.* **48**, 308–313 (2016).
- 563 26. Bollback, J. P. SIMMAP: stochastic character mapping of discrete traits on  
564 phylogenies. *BMC Bioinformatics* **7**, 88 (2006).
- 565 27. Revell, L. J. phytools: An R package for phylogenetic comparative biology  
566 (and other things). *Methods Ecol. Evol.* **3**, 217–223 (2012).
- 567 28. Bouckaert, R. *et al.* BEAST 2: A software platform for Bayesian evolutionary  
568 analysis. *PLoS Comput. Biol.* **10**, 1–6 (2014).
- 569 29. Croucher, N. J. *et al.* Rapid phylogenetic analysis of large samples of  
570 recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids*  
571 *Res.* **43**, e15 (2015).
- 572 30. Cantinelli, T. *et al.* ‘Epidemic clones’ of *Listeria monocytogenes* are  
573 widespread and ancient clonal groups. *J. Clin. Microbiol.* **51**, 3770–3779  
574 (2013).
- 575 31. Chenal-Francisque, V. *et al.* Clonogrouping, a rapid multiplex PCR method to  
576 identify major clones of *Listeria monocytogenes*. *J. Clin. Microbiol.* **53**, 3335–  
577 3358 (2015).
- 578 32. Ferreira, V., Wiedmann, M., Teixeira, P. & Stasiewicz, M. J. *Listeria*  
579 *monocytogenes* persistence in food-associated environments: epidemiology,  
580 strain characteristics, and implications for public health. *J. Food Prot.* **77**,  
581 150–70 (2014).
- 582 33. Piffaretti, J. C. *et al.* Genetic characterization of clones of the bacterium  
583 *Listeria monocytogenes* causing epidemic disease. *Proc. Natl. Acad. Sci. U. S. A.*  
584 **86**, 3818–3822 (1989).
- 585 34. Wiedmann, M. *et al.* Ribotypes and virulence gene polymorphisms suggest  
586 three distinct *Listeria monocytogenes* lineages with differences in pathogenic  
587 potential. *Infect. Immun.* **65**, 2707–16 (1997).
- 588 35. Leclercq, A. *et al.* Characterization of the novel *Listeria monocytogenes* PCR  
589 serogrouping profile IVb-v1. *Int. J. Food Microbiol.* **147**, 74–7 (2011).

- 590 36. Cotter, P. D. *et al.* Listeriolysin S, a novel peptide haemolysin associated with  
591 a subset of lineage I *Listeria monocytogenes*. *PLoS Pathog.* **4**, e1000144  
592 (2008).
- 593 37. Jacquet, C. *et al.* A molecular marker for evaluating the pathogenic potential  
594 of foodborne *Listeria monocytogenes*. *J. Infect. Dis.* **189**, 2094–2100 (2004).
- 595 38. Verghese, B. *et al.* *comK* prophage junction fragments as markers for *Listeria*  
596 *monocytogenes* genotypes unique to individual meat and poultry processing  
597 plants and a model for rapid niche-specific adaptation, biofilm formation,  
598 and persistence. *Appl. Environ. Microbiol.* **77**, 3279–3292 (2011).
- 599 39. Rabinovich, L., Sigal, N., Borovok, I., Nir-Paz, R. & Herskovits, A. a. Prophage  
600 excision activates *Listeria* competence genes that promote phagosomal  
601 escape and virulence. *Cell* **150**, 792–802 (2012).
- 602 40. Müller, A. *et al.* The *Listeria monocytogenes* transposon Tn6188 provides  
603 increased tolerance to various quaternary ammonium compounds and  
604 ethidium bromide. *FEMS Microbiol. Lett.* **361**, 166–73 (2014).
- 605 41. Schmitz-Esser, S., Müller, A., Stessl, B. & Wagner, M. Genomes of sequence  
606 type 121 *Listeria monocytogenes* strains harbor highly conserved plasmids  
607 and prophages. *Front. Microbiol.* **6**, 380 (2015).
- 608 42. Acciari, V. A. *et al.* Tracing sources of *Listeria* contamination in traditional  
609 Italian cheese associated with a US outbreak: investigations in Italy.  
610 *Epidemiol. Infect.* **2**, 1–9 (2015).
- 611 43. Leclercq, A., Charlier, C. & Lecuit, M. Global burden of listeriosis: the tip of  
612 the iceberg. *Lancet Infect. Dis.* **14**, 1027–8 (2014).
- 613 44. Kuenne, C. *et al.* Reassessment of the *Listeria monocytogenes* pan-genome  
614 reveals dynamic integration hotspots and mobile genetic elements as major  
615 components of the accessory genome. *BMC Genomics* **14**, 47 (2013).
- 616 45. Holch, A. *et al.* Genome sequencing identifies two nearly unchanged strains  
617 of persistent *Listeria monocytogenes* isolated at two different fish processing  
618 plants sampled 6 years apart. *Appl. Environ. Microbiol.* **79**, 2944–51 (2013).
- 619

620 **Acknowledgments**

621 The authors wish to thank Keith Jolley (Oxford University) for assistance with BIGSdb  
622 implementation, PulseNet International Network members for continuous surveillance  
623 and data sharing, the Genomics platform (PF1, Institut Pasteur) for assistance with  
624 sequencing, Damien Mornico (Institut Pasteur) for assistance with the submission of raw  
625 data, Jana Haase and Mark Achtman (Environmental Research Institute, Ireland) for  
626 providing cultures of historical isolates of SL1. The authors are also grateful to Nathalie  
627 Tessaud-Rita, Guillaume Vales and Pierre Thouvenot (National Reference Centre for  
628 Listeria, Institut Pasteur) for recovering and extracting DNA from historical isolates of SL9.

629 This work was supported by Institut Pasteur, INSERM, Public Health France, French  
630 government's Investissement d'Avenir program Laboratoire d'Excellence 'Integrative  
631 Biology of Emerging Infectious Diseases' (grant ANR-10-LABX-62-IBEID), European  
632 Research Council, Swiss National Fund for Research and the Advanced Molecular  
633 Detection (AMD) initiative at CDC.

634

635 **Author contributions**

636 This study was designed by SB, ML, PGS and BP. Selection of isolates was carried out by  
637 EMN, CN, VCF, AL, AR, KG, TD and LSK. DNA preparation and sequencing was performed  
638 by HBD, VCF, AL, CT, HC, SS, ZK, JTB, AR, CN, KG, MW and VE. PFGE analysis was performed  
639 by HBD, VCF, AL and AM. Sequence analysis was carried out by AM, HP, TC, LK, HC, JTB.  
640 Definition of core genome was done by MMM, EPCR, MTouchon. Validation and  
641 reproducibility of cgMLST loci was performed by AM, HP and EL. Phylogenetic and  
642 clustering analyses were carried out by AM and AC. Online database implementation was  
643 done by LJ, AM and SB. Epidemiological data analysis was performed by MTourdjman, AL,  
644 AM, TD, KG, EMN and CT. AM and SB wrote the manuscript, with contributions and  
645 comments from all authors.

646

647 **Additional information**

648 Correspondence and requests for materials should be addressed to ML and SB.

649

650 **Declaration of interests**

651 HP and BP are co-developers of the BioNumerics software mentioned in the manuscript.

652 The remaining authors declare no competing interests.

653

654 **Legends to figures**

655 **Figure 1. Nomenclature of *Lm* cgMLST profiles.** A) Distribution of the number of  
656 cgMLST allelic differences between pairs of isolates among the 1,696 genomes (blue) and  
657 within 49 sets of epidemiologically related isolates (426 isolates in total; red). Dashed bars  
658 represent cut-off values for cgMLST types (CT, 7 allelic mismatches) and sublineages (SL,  
659 150 allelic mismatches). Inset: global dataset; Main figure: up to 200 allelic mismatches. B)  
660 Rarefaction curves of the number of sublineages and cgMLST types identified, broken  
661 down per main phylogenetic lineage (I-IV). Curves were estimated using 100 random  
662 samples per point. Inset: zoom on the 0-50 X-axis values. Lineages III and IV were pooled  
663 but must be sampled more extensively to determine the shape of the curve.

664

665 **Figure 2. Phylogenetic structure of the global *Lm* dataset.** A) Phylogeny of the four  
666 phylogenetic lineages (I, red; II, orange; III, green; IV, blue). Representative isolates of the  
667 four lineages were used to determine the location of the root, using *L. innocua* and *L.*  
668 *marthii* as outgroups. The tree was obtained using FastME on the p-distance of the 1,748  
669 concatenated alignments. B) Comparison of the phylogeny obtained from 1,748  
670 recombination-purged sequence alignments (left) and from cgMLST allelic profile  
671 distances (right). To reduce redundancy, only one strain per outbreak set was used. Scale  
672 bars indicate the % of nucleotide substitutions (A right and B left) and the % of allelic  
673 mismatches (B right). For practical reasons, bootstrap values (based on 500 replicates)  
674 are shown only for long internal branches.

675

676 **Figure 3. International distribution of *Lm* sublineages.** A) Clustering of 1,696 *Lm*  
677 isolates based on single-linkage analysis of the cgMLST profiles. Lineage branch colours  
678 are as in Fig. 2. Light and dark grey alternation (inner circle) delimitates sublineages with  
679 more than 10 isolates (main sublineages are labelled). Source country is represented in  
680 the external ring using the colour key from panel C. B) Number of countries from which a  
681 sublineage was isolated, as a function of number of isolates per sublineage. Disk size is a  
682 function of number of isolates per sublineage. C) Inferred geographical origin of ancestral  
683 nodes of the phylogeny of sublineage 1. Pie charts represent the likelihood proportion of  
684 geographical origins. The tree was constructed using minimum evolution based on  
685 cgMLST profiles. Bootstrap values above 50% (based on 500 replicates) are shown for the  
686 major nodes. D) Absolute number of geographical transitions (left) and number of  
687 geographical transitions normalized by total branch length (right) within the 10 most  
688 frequent sublineages, as inferred by stochastic ancestral state reconstructions (numbers in  
689 parentheses indicate the precise values inferred for each sublineage).

690

691 **Figure 4. International groups of isolates classified into the same cgMLST type.** The 9  
692 groups of isolates are indicated by a specific colour. The genotype is indicated as a string  
693 consisting of a succession of lineage (e.g., L1), sublineage (e.g., SL1), sequence type (e.g.,  
694 ST1) and cgMLST type (e.g., CT288). Countries of isolation, isolation year range and total  
695 number of isolates are given after the genotype string. The circles on the map indicate the  
696 country where a particular CT was isolated and their size is related to the number of  
697 isolates from that country. The details of each CT are given in Supplementary Information.

698 Abbreviations: L, lineage; ST, sequence type; SL, sublineage; CT, cgMLST type; US, United  
699 States of America; CA, Canada; DK, Denmark; UK, United Kingdom; FR, France.

700

701 **Figure 5. Temporal analysis of cgMLST profiles evolution.** A) Best-fitting rooted  
702 phylogeny of SL1 isolates ( $n=195$ ), including the historical isolates. The tree was obtained  
703 using FastME on cgMLST profiles. Coloured blocks represent the isolation time range  
704 (1921-1950, pink; 1951-1980, purple; 1981-2010, blue; 2011-2015, green). Outbreak  
705 reference strains are indicated by red dots. Outbreak identifier, country, year and cgMLST  
706 type are provided on the right. The scale bar indicates the number of allelic substitutions  
707 per locus. Statistical significance was assessed using F-test. B) Linear regression of  
708 isolation year with root-to-tip cgMLST distance. C) Accumulation of cgMLST variation over  
709 time, determined based on the international CTs ( $n=9$ ) and outbreak sets ( $n=49$ ).  
710 Statistical significance was assessed using F-test.

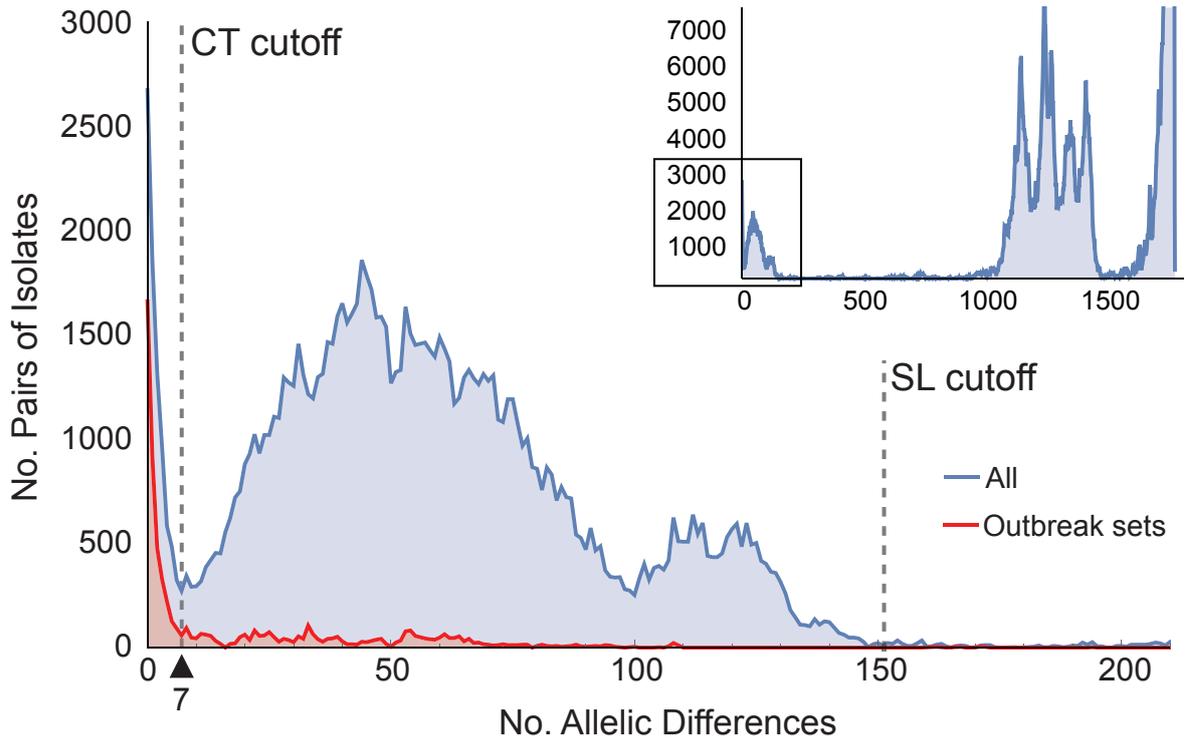
711

712 **Figure 6. Virulence and resistance profiles across the phylogeny of the 1,696 *Lm***  
713 **isolates.** A) Cluster analysis based on cgMLST profiles. The dotted vertical bar indicates  
714 the cgMLST mismatch cut-off for sublineages (SL). The 10 most frequent sublineages are  
715 highlighted. B) Pattern of gene presence (color line) or absence (white). The first and last  
716 columns corresponds to the serogroup and sample source, respectively, represented by  
717 color codes (upper left key). The presence/absence gene matrix represents, from left to  
718 right, genes involved in teichoic acid biosynthesis (*gltAB*, *tagB*, *gtcA*), genes located in the  
719 pathogenicity islands LIPI-1 (*prfA*, *plcA*, *hly*, *mpl*, *actA*, *plcB*), LIPI-3 (*llsAGHXBYDP*) and  
720 LIPI-4 (LM9005581\_70009 to LM9005581\_70014), genes coding for internalins  
721 (*inlABCEFGHJK*), and other genes involved in adherence (*ami*, *dltA*, *fbpA*, *lap*, *lapB*),  
722 invasion (*aut*, *aut\_IVb*, *cwhA*, *lpeA*, *vip*), intracellular survival (*hpt*, *lplA1*, *oppA*, *prsA2*, *purQ*,  
723 *svpA*), regulation of transcription and translation (*agrAC*, *cheAY*, *fur*, *lisKR*, *rsbV*, *sigB*, *stp*,  
724 *virRS*), surface protein anchoring (*lgt*, *lspA*, *srtAB*), peptidoglycan modification (*oatA*,  
725 *pdgA*), immune modulation (*IntA*), bile-resistance (*bsh*, *mdrM*, *mdrT*, *brtA*), resistance to  
726 detergents (*qac*, *bcrABC*, *ermE*) and biofilm formation and virulence (*comK*).

727

Figure 1

A



B

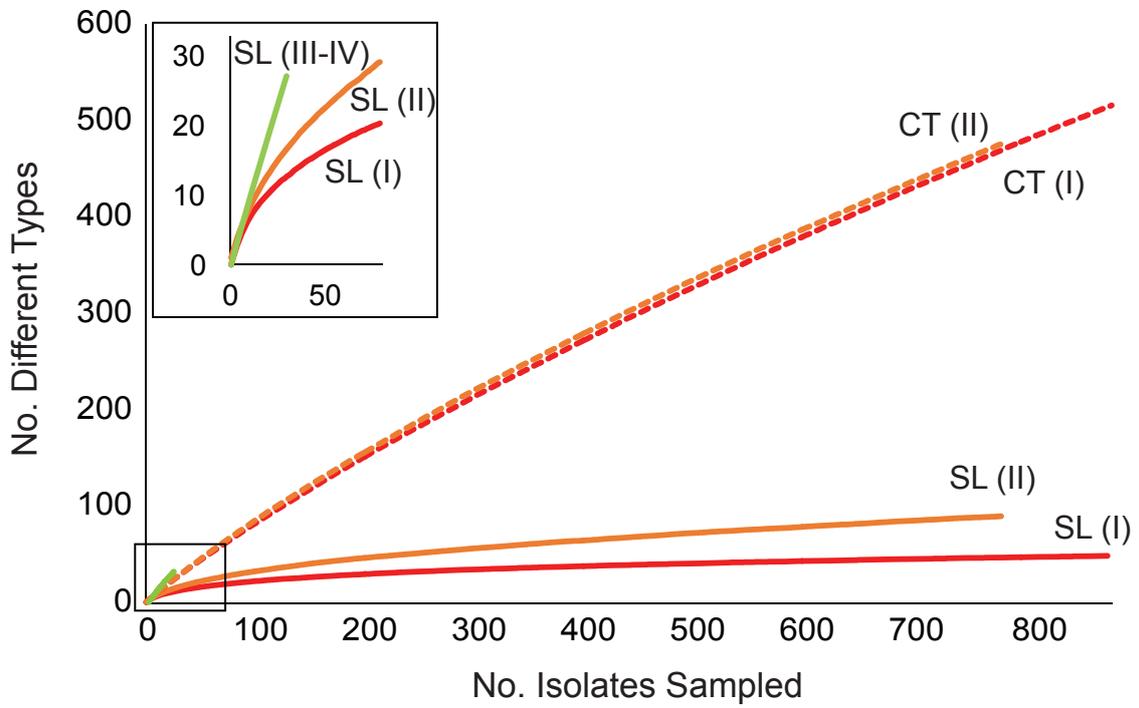
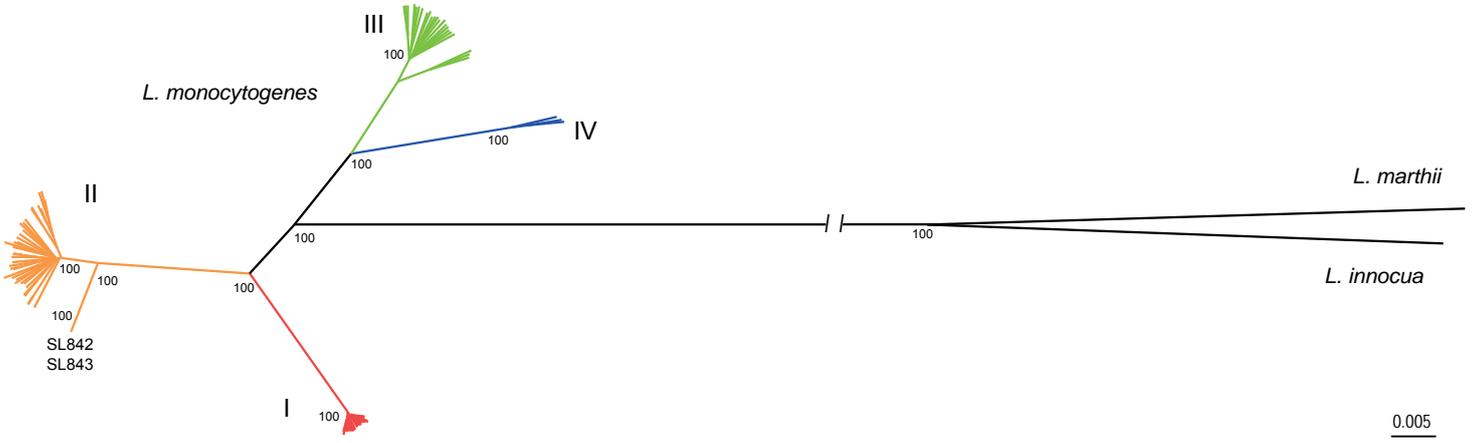
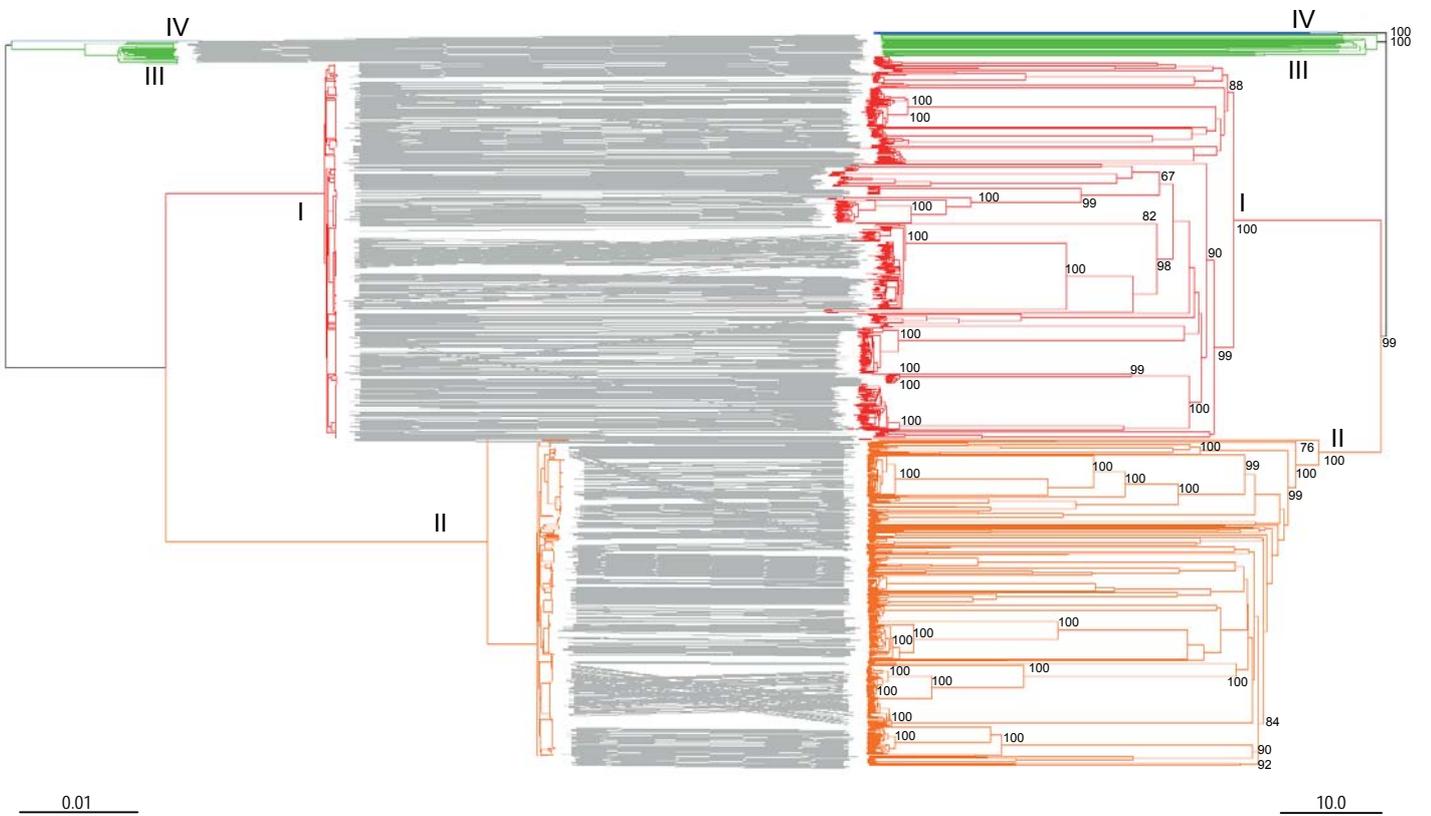


Figure 2

A

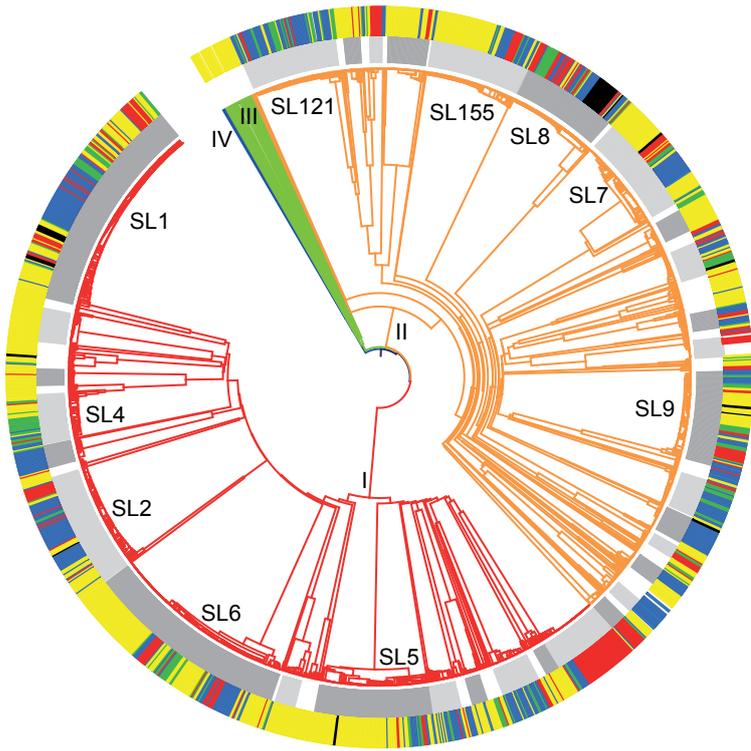


B

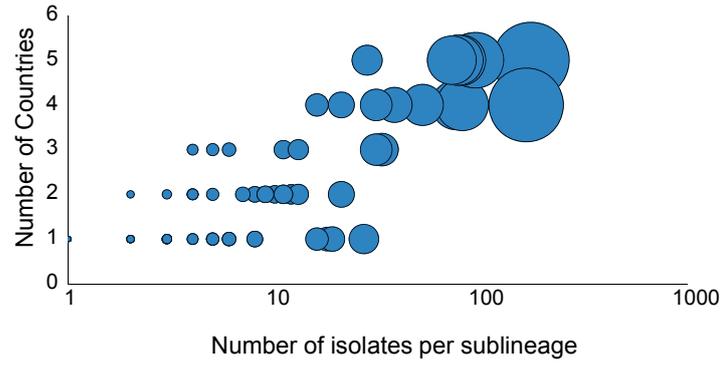


**Figure 3**

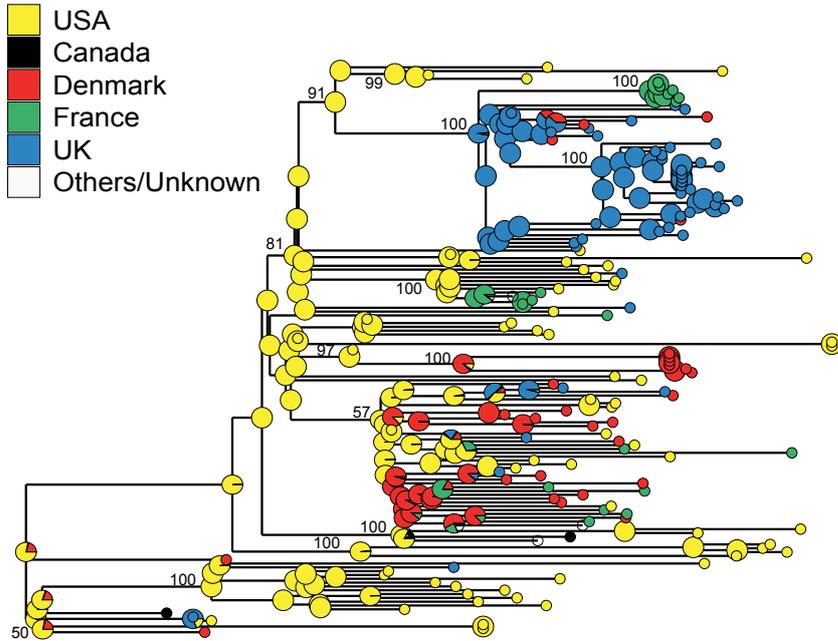
**A**



**B**



**C**



**D**

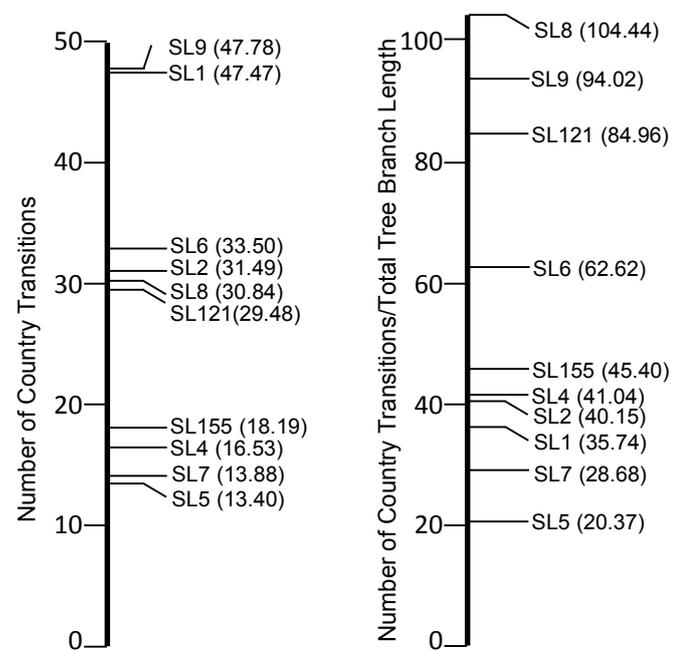
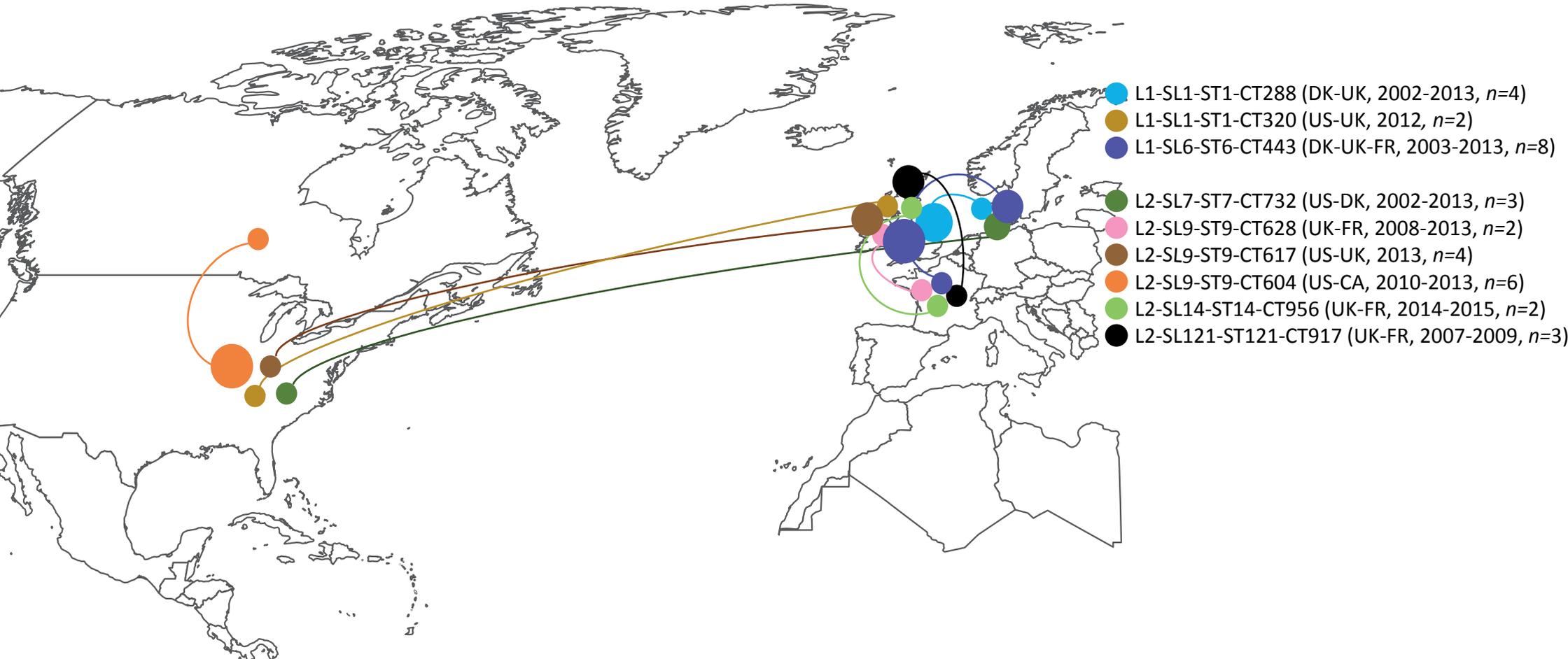
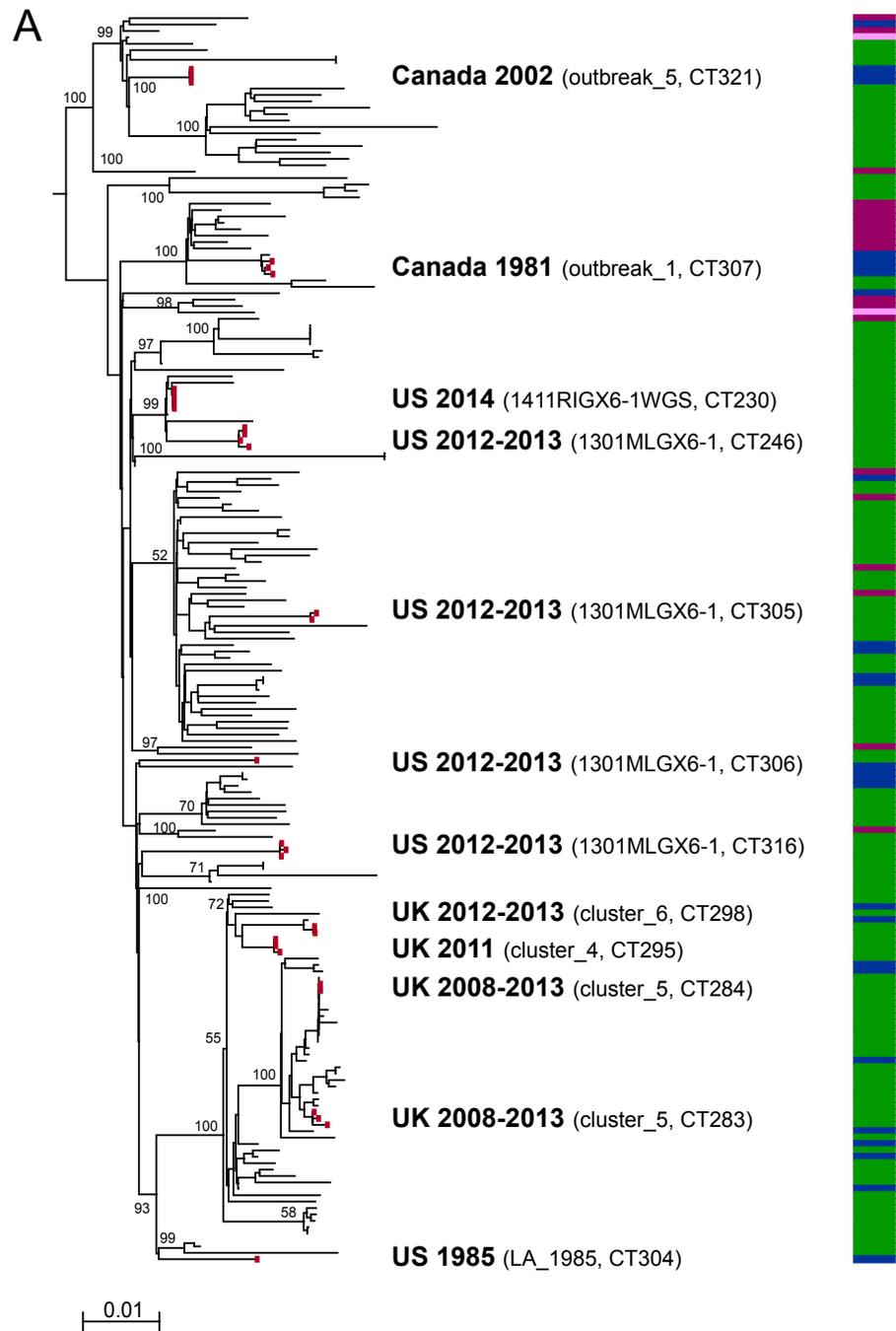


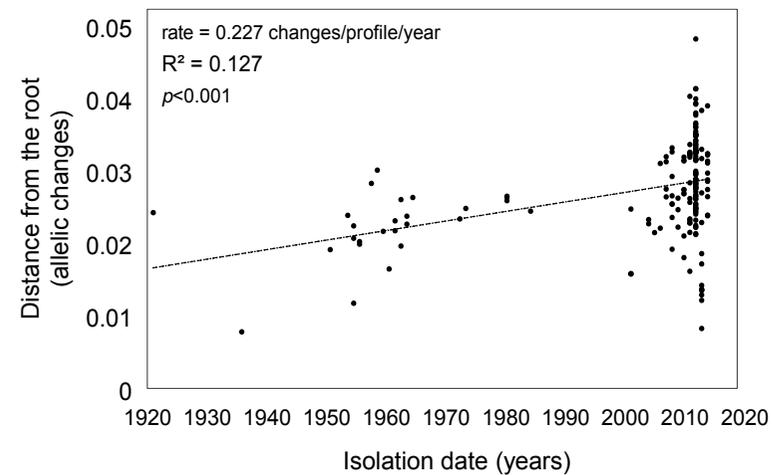
Figure 4



**Figure 5**



**B**



**C**

