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.

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

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

[1] Institut Pasteur, National Reference Centre and World Health Organization Collaborating Center for *Listeria*, Paris, France;

[2] Institut Pasteur, Biology of Infection Unit, Paris, France;

[3] Inserm U1117, Paris France;

[4] Institut Pasteur, Microbial Evolutionary Genomics Unit, Paris, France;

[5] CNRS, UMR 3525, Paris, France;

[6] Institut Pasteur – Hub Bioinformatique et Biostatistique – C3BI, USR 3756 IP CNRS – Paris, France;

[7] Applied-Maths, Sint-Martens-Latem, Belgium;

[8] Paris Diderot University, Sorbonne Paris Cité, Cellule Pasteur, Paris, France;

[9] Centers for Disease Control and Prevention, Atlanta, Georgia, United States;

[10] Statens Serum Institut, Copenhagen, Denmark;

[11] Public Health England, London, United Kingdom;

[12] Public Health Agency of Canada, Winnipeg, Canada;

[13] Institut Pasteur, Pasteur International Bioresources network (PIBnet), Mutualized Microbiology Platform (P2M), Paris, France;

[14] CNRS, UMS 3601 IFB-Core, Gif-sur-Yvette, France;

[15] Public Health France, Saint-Maurice, France;

[16] Paris Descartes University, Sorbonne Paris Cité, Institut Imagine, Necker-Enfants Malades University Hospital, Division of Infectious Diseases and Tropical Medicine, APHP, Paris, France;

* Correspondence to sylvain.brisse@pasteur.fr and marc.lecuit@pasteur.fr

Abstract

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

Introduction

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

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

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

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

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

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

Results

Universal Lm cgMLST

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

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

Definition of cgMLST types

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

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

Comparison of cgMLST and PFGE genotyping

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

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

Phylogenetic structure and nomenclature of Lm sublineages

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

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

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

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

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

Evidence for international spread of Lm strains

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

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

Temporal accumulation of variation within Lm lineages and outbreaks

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

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

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

The phylogeny of SL1 (**Fig. 5A**) showed that outbreaks strains were dispersed in multiple branches, suggesting that all SL1 isolates have the potential to cause outbreaks. Moreover, it demonstrated that the multiple outbreaks caused by this sublineage, previously called 'epidemic clone ECI',[30] are actually independent epidemiological events. The most recent common ancestor of SL1 was estimated to have existed around 1876 (95% HPD=[1861-1891]), reinforcing the idea that 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.[10,33,34] In contrast, PCR-serogroup variant IVb-v1[35] was found in
280 various branches. Likewise, and as expected, serogroup L[35] 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[36] was almost
286 exclusively detected within lineage I. The recently described LIPI-4[25] 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,[37] 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[38,39], 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,[40] 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.[41]
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

Discussion

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

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

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

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

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

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

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

Methods

Bacterial isolates

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

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

411 A previously defined *Lm* core genome with 1,791 loci[25] 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 *Asc*I and *Apa*I 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

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

Phylogenetic and clustering analyses

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

Phylogeography and temporal analysis

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

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

(**Supplementary Information 2.8**, **Supplementary Tables 1 and 5**). Phylogenetic analyses were performed using FastME on p-distances estimated from either concatenated multiple sequence alignments or cgMLST profiles. Linear regression of the root-to-tip distances against the year of isolation was carried out using Path-O-Gen v1.4 (http://tree.bio.ed.ac.uk/software/pathogen/). The rate of evolution of SL1 and SL9 genomes were independently estimated from the concatenated multiple sequence alignments of the 1,748 loci using BEAST v.2.3.1[28]. For this analysis, Gubbins[29] was used to detect recombination within the alignments. Isolates with recombinant regions were discarded from the alignments (**Supplementary Information 2.8**). Subsequently, the mean of the rates of SL1 and SL9 ($2.5x10^{-7}$ substitutions per site per year) was used to estimate the age of all major sublineages using BEAST v.2.3.1[28]. Details of the temporal analysis methods are given in **Supplementary Information 2.8**. Genetic divergence as a function of the time span between the first and last isolate of outbreak sets and international clusters was evaluated using regression analysis.

Determination of PCR serogroups, virulence and resistance genes profiles

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

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

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

Accession numbers

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

References

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

Microbiol. **53,** 2869–2876 (2015).
20. Stasiewicz, M. J., Oliver, H. F., Wiedmann, M. & den Bakker, H. C. Whole genome sequencing allows for improved identification of persistent *Listeria monocytogenes* in food associated environments. *Appl. Environ. Microbiol.* **81,** 6024–6037 (2015).
21. Fretz, R. *et al.* Update: Multinational listeriosis outbreak due to 'quargel', a sour milk curd cheese, caused by two different *L. monocytogenes* serotype 1/2a strains, 2009-2010. *Eurosurveillance* **15,** 2–3 (2010).
22. Pightling, A. W., Petronella, N. & Pagotto, F. Choice of reference sequence and assembler for alignment of *Listeria monocytogenes* short-read sequence data greatly influences rates of error in SNP analyses. *PLoS One* **9,** e104579 (2014).
23. Jolley, K. A. & Maiden, M. C. J. BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* **11,** 595 (2010).
24. 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).
25. Maury, M. *et al.* Uncovering *Listeria monocytogenes* hypervirulence by harnessing its biodiversity. *Nat. Genet.* **48,** 308–313 (2016).
26. Bollback, J. P. SIMMAP: stochastic character mapping of discrete traits on phylogenies. *BMC Bioinformatics* **7,** 88 (2006).
27. Revell, L. J. phytools: An R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* **3,** 217–223 (2012).
28. Bouckaert, R. *et al.* BEAST 2: A software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **10,** 1–6 (2014).
29. 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).
30. Cantinelli, T. *et al.* 'Epidemic clones' of *Listeria monocytogenes* are widespread and ancient clonal groups. *J. Clin. Microbiol.* **51,** 3770–3779 (2013).
31. Chenal-Francisque, V. *et al.* Clonogrouping, a rapid multiplex PCR method to identify major clones of *Listeria monocytogenes*. *J. Clin. Microbiol.* **53,** 3335–3358 (2015).
32. Ferreira, V., Wiedmann, M., Teixeira, P. & Stasiewicz, M. J. *Listeria monocytogenes* persistence in food-associated environments: epidemiology, strain characteristics, and implications for public health. *J. Food Prot.* **77,** 150–70 (2014).
33. Piffaretti, J. C. *et al.* Genetic characterization of clones of the bacterium *Listeria monocytogenes* causing epidemic disease. *Proc. Natl. Acad. Sci. U. S. A.* **86,** 3818–3822 (1989).
34. Wiedmann, M. *et al.* Ribotypes and virulence gene polymorphisms suggest three distinct *Listeria monocytogenes* lineages with differences in pathogenic potential. *Infect. Immun.* **65,** 2707–16 (1997).
35. Leclercq, A. *et al.* Characterization of the novel *Listeria monocytogenes* PCR 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

Acknowledgments

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

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

Author contributions

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

Additional information

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

Declaration of interests

HP and BP are co-developers of the BioNumerics software mentioned in the manuscript. The remaining authors declare no competing interests.

Legends to figures

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

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

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

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

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

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

Figure 6. Virulence and resistance profiles across the phylogeny of the 1,696 *Lm* isolates. A) Cluster analysis based on cgMLST profiles. The dotted vertical bar indicates the cgMLST mismatch cut-off for sublineages (SL). The 10 most frequent sublineages are highlighted. B) Pattern of gene presence (color line) or absence (white). The first and last columns corresponds to the serogroup and sample source, respectively, represented by color codes (upper left key). The presence/absence gene matrix represents, from left to right, 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, mdrM, mdrT, brtA),* resistance to detergents (*qac, bcrABC, ermE*) and biofilm formation and virulence *(comK).*

Figure 1

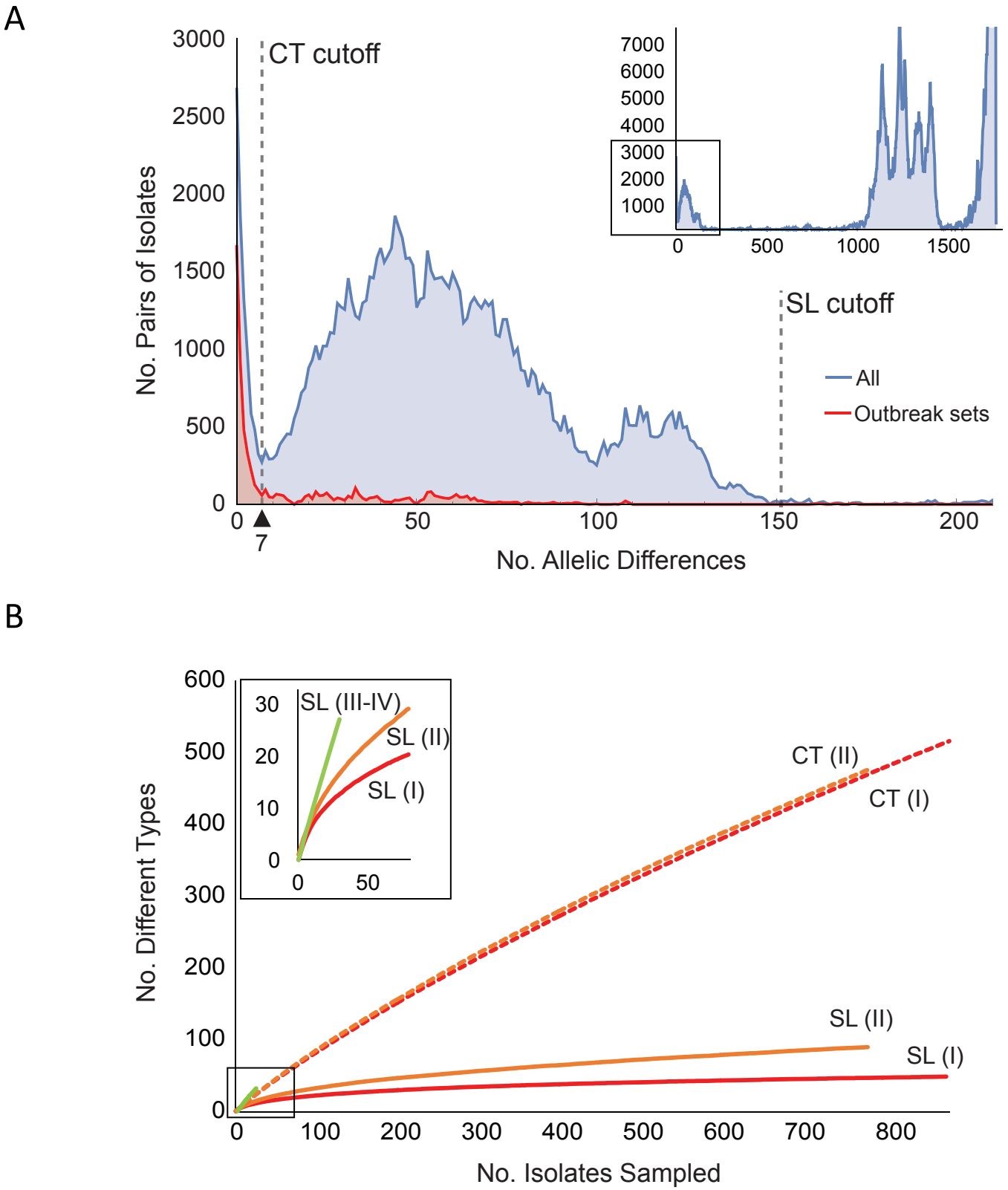
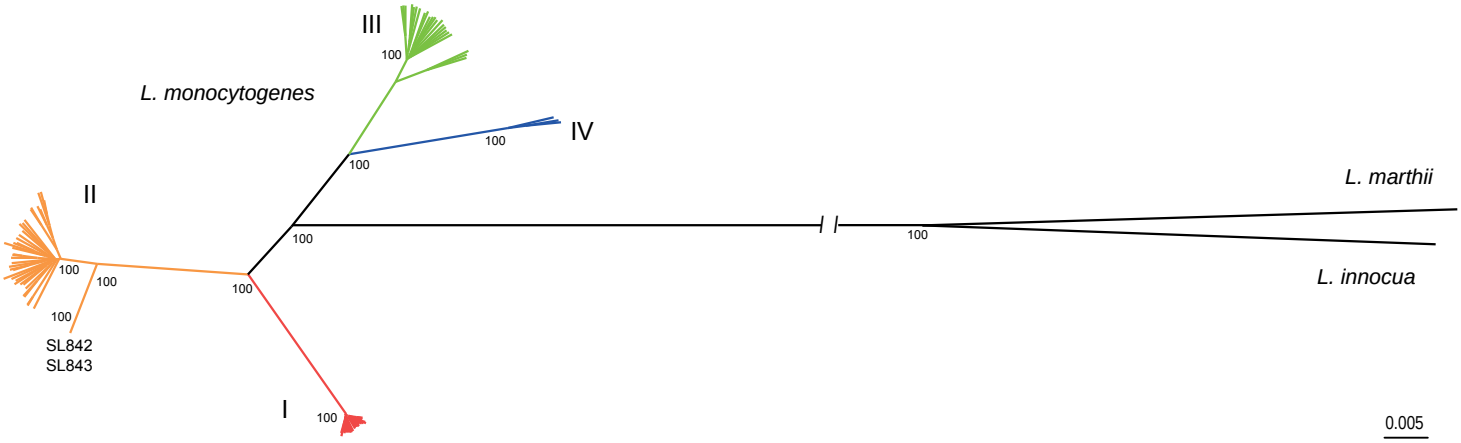

A
CT cutoff
SL cutoff
No. Pairs of Isolates
No. Allelic Differences
All
Outbreak sets
B
No. Different Types
No. Isolates Sampled
CT (II)
CT (I)
SL (II)
SL (I)
SL (III-IV)

Figure 2

A


III
100
L. monocytogenes
IV
100
100
II
100
100
100
100
100
L. marthii
100
L. innocua
SL842
SL843
I
100
0.005

B

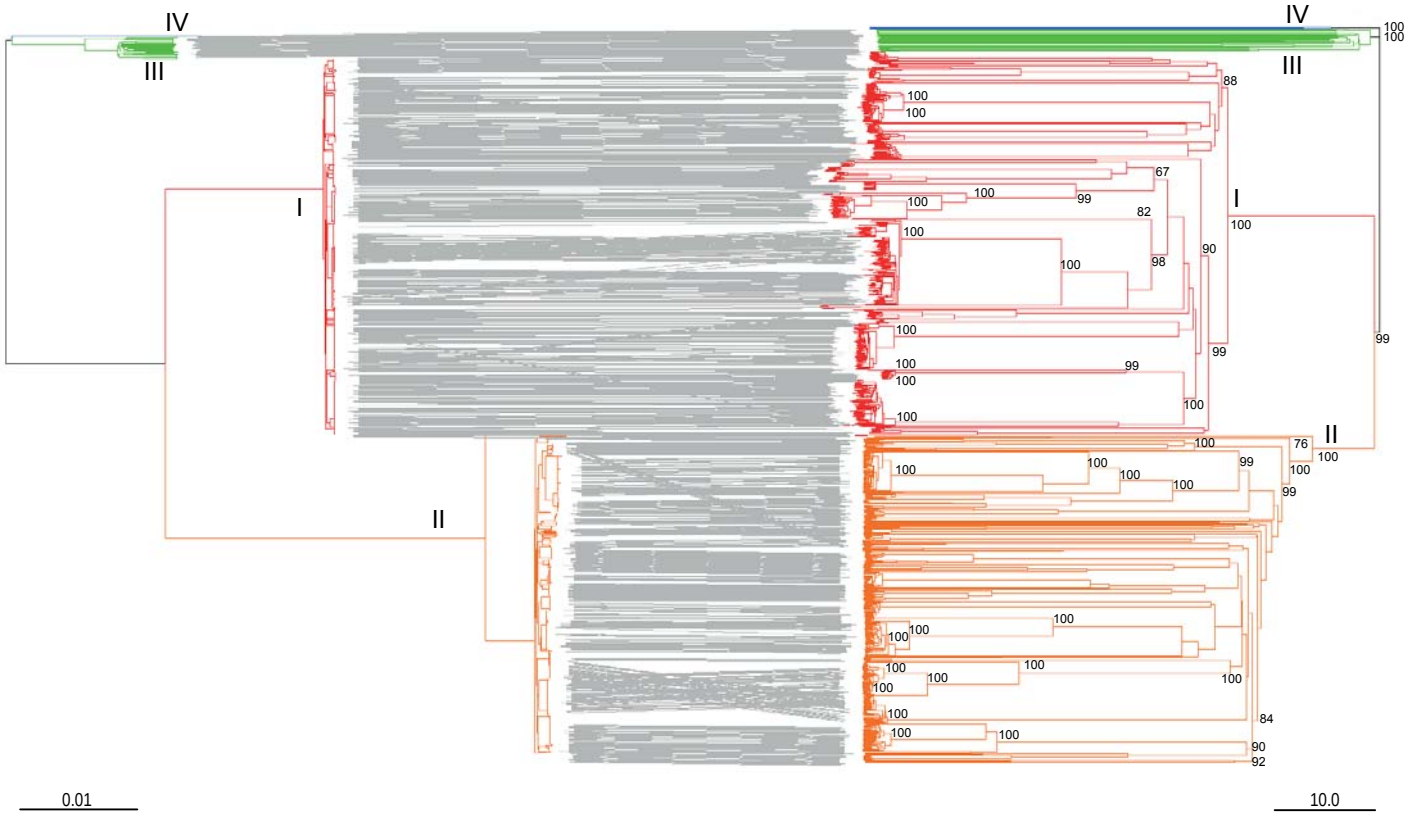

IV
III
I
II
0.01
10.0

Figure 3

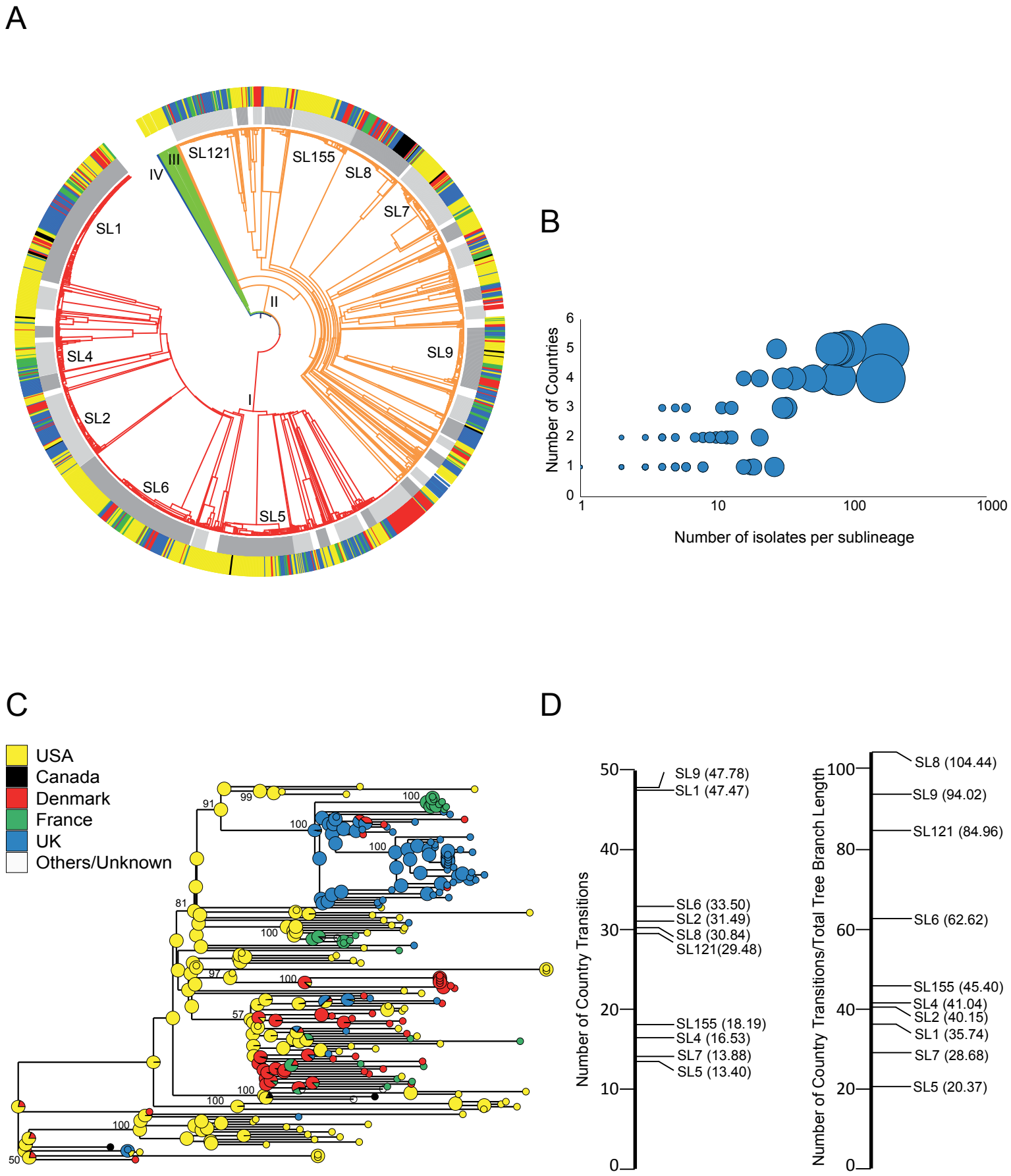

A
SL121
SL155
SL8
SL7
III
IV
SL1
II
SL4
SL9
I
SL2
SL6
SL5
B
Number of Countries
0
1
2
3
4
5
6
1
10
100
1000
Number of isolates per sublineage
C
USA
Canada
Denmark
France
UK
Others/Unknown
91
99
100
100
100
81
100
97
100
57
100
100
100
50
D
Number of Country Transitions
50
40
30
20
10
0
SL9 (47.78)
SL1 (47.47)
SL6 (33.50)
SL2 (31.49)
SL8 (30.84)
SL121(29.48)
SL155 (18.19)
SL4 (16.53)
SL7 (13.88)
SL5 (13.40)
Number of Country Transitions/Total Tree Branch Length
100
80
60
40
20
0
SL8 (104.44)
SL9 (94.02)
SL121 (84.96)
SL6 (62.62)
SL155 (45.40)
SL4 (41.04)
SL2 (40.15)
SL1 (35.74)
SL7 (28.68)
SL5 (20.37)

Figure 4

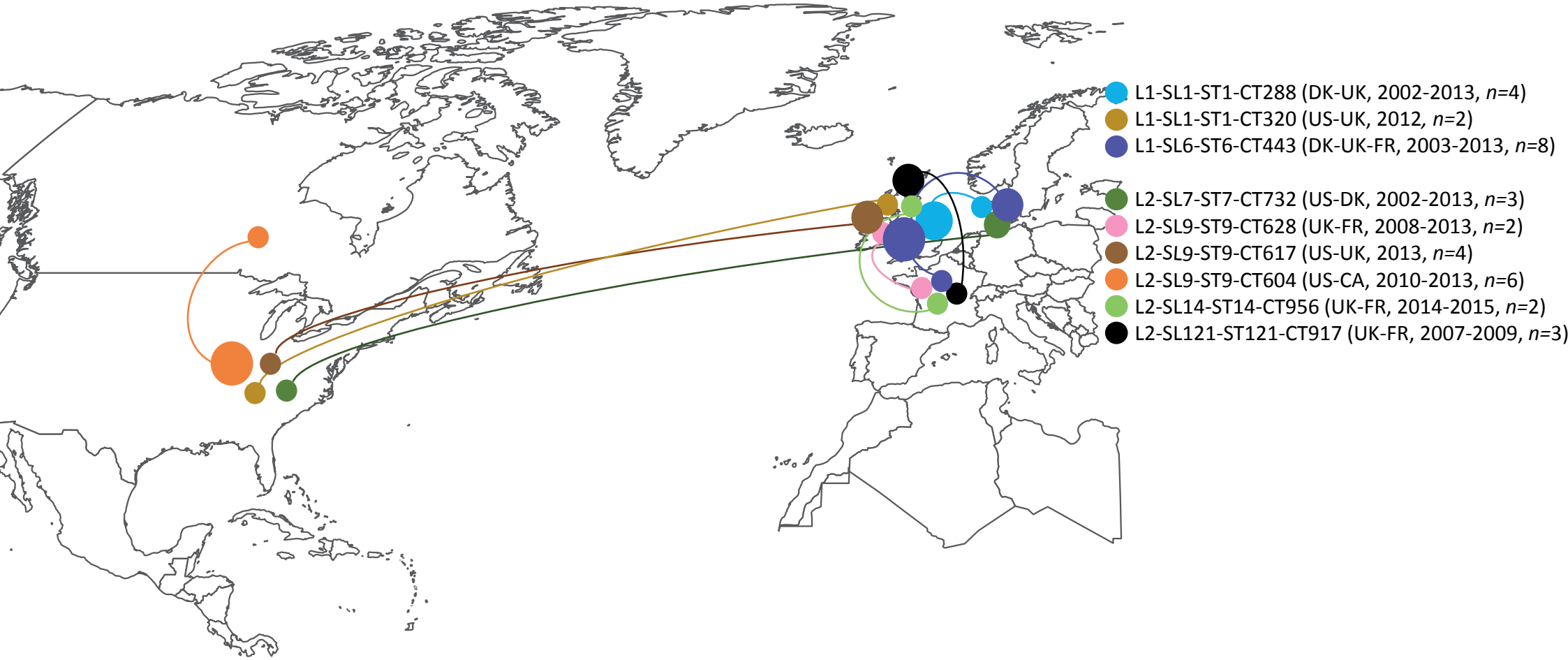

L1-SL1-ST1-CT288 (DK-UK, 2002-2013, n=4)
L1-SL1-ST1-CT320 (US-UK, 2012, n=2)
L1-SL6-ST6-CT443 (DK-UK-FR, 2003-2013, n=8)
L2-SL7-ST7-CT732 (US-DK, 2002-2013, n=3)
L2-SL9-ST9-CT628 (UK-FR, 2008-2013, n=2)
L2-SL9-ST9-CT617 (US-UK, 2013, n=4)
L2-SL9-ST9-CT604 (US-CA, 2010-2013, n=6)
L2-SL14-ST14-CT956 (UK-FR, 2014-2015, n=2)
L2-SL121-ST121-CT917 (UK-FR, 2007-2009, n=3)

Figure 5

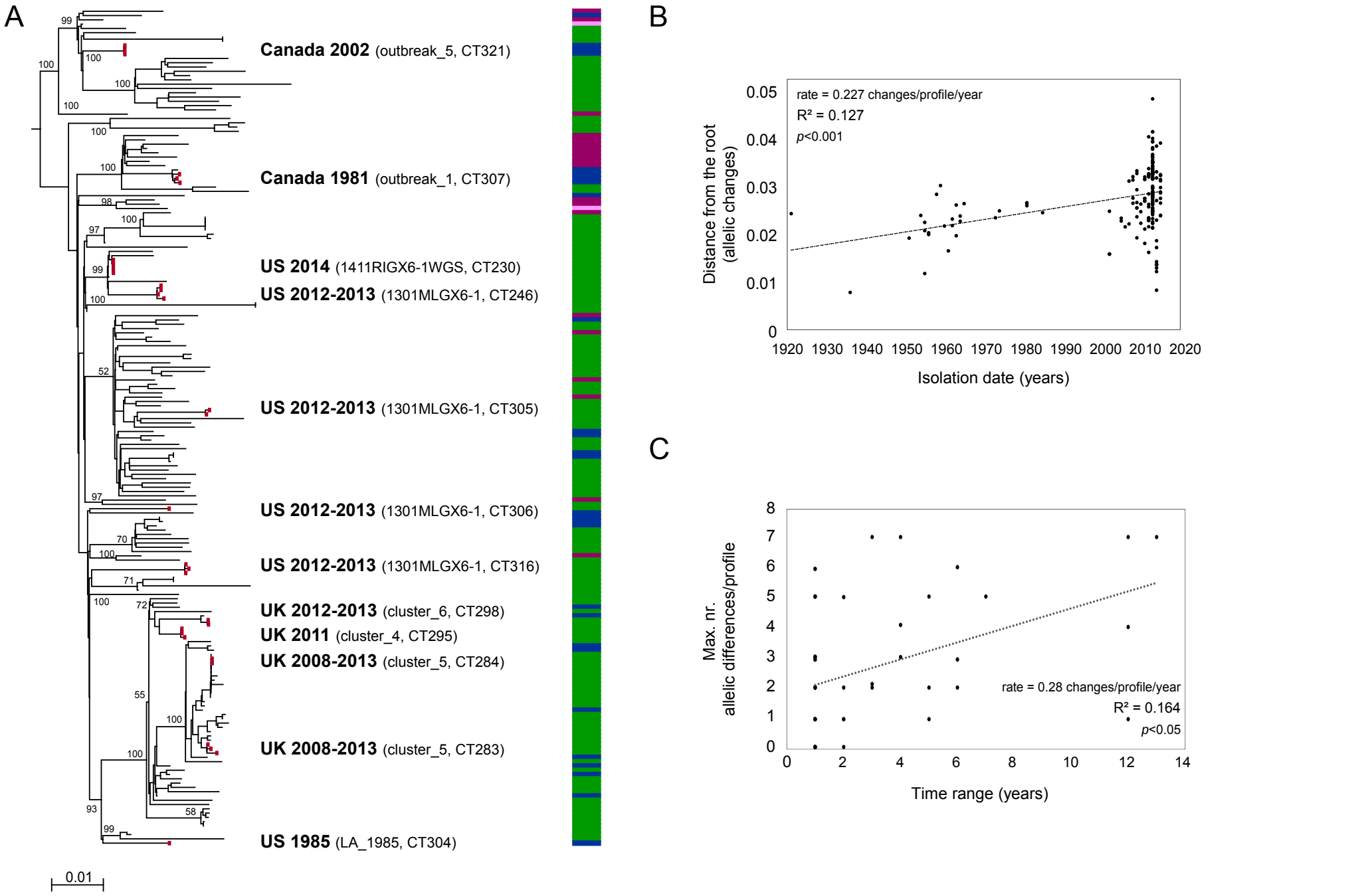
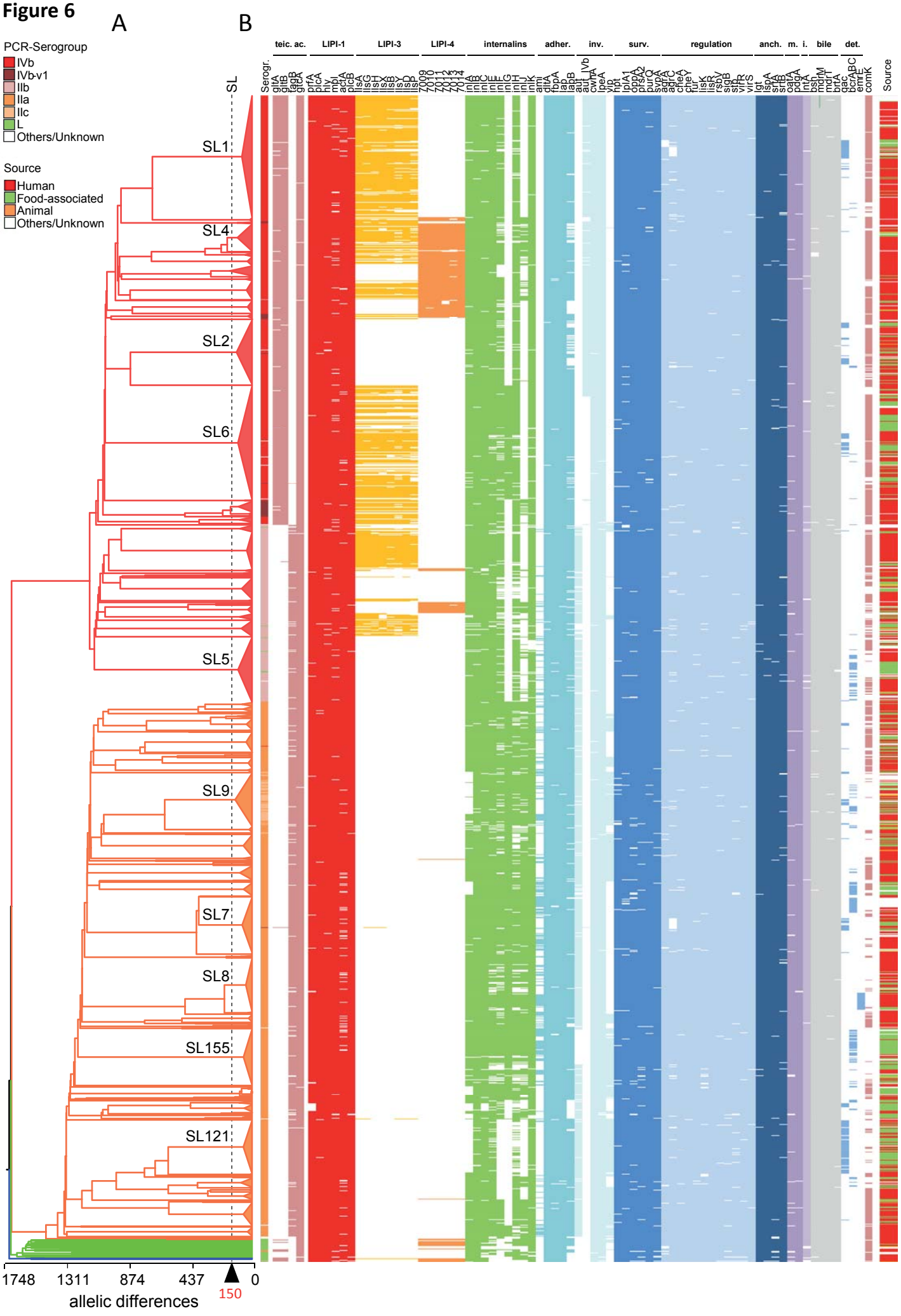

A
Canada 2002 (outbreak_5, CT321)
Canada 1981 (outbreak_1, CT307)
US 2014 (1411RIGX6-1WGS, CT230)
US 2012-2013 (1301MLGX6-1, CT246)
US 2012-2013 (1301MLGX6-1, CT305)
US 2012-2013 (1301MLGX6-1, CT306)
US 2012-2013 (1301MLGX6-1, CT316)
UK 2012-2013 (cluster_6, CT298)
UK 2011 (cluster_4, CT295)
UK 2008-2013 (cluster_5, CT284)
UK 2008-2013 (cluster_5, CT283)
US 1985 (LA_1985, CT304)
0.01
B
rate = 0.227 changes/profile/year
R² = 0.127
p<0.001
Distance from the root (allelic changes)
Isolation date (years)
C
rate = 0.28 changes/profile/year
R² = 0.164
p<0.05
Max. nr. allelic differences/profile
Time range (years)

Figure 6


A
B
PCR-Serogroup
IVb
IVb-v1
IIb
IIa
IIc
L
Others/Unknown
Source
Human
Food-associated
Animal
Others/Unknown
SL
SL1
SL4
SL2
SL6
SL5
SL9
SL7
SL8
SL155
SL121
1748
1311
874
437
0
150
allelic differences
teic. ac.
LIPI-1
LIPI-3
LIPI-4
internalins
adher.
inv.
surv.
regulation
anch.
m.
i.
bile
det.
Serogr.
Source