JOURNAL OF CLINICAL MICROBIOLOGY, Oct. 2006, p. 3742–3751 Vol. 44, No. 10
0095-1137/06/$08.00+0 doi:10.1128/JCM.00618-06

Novel Method To Identify Source-Associated Phylogenetic Clustering Shows that *Listeria monocytogenes* Includes Niche-Adapted Clonal Groups with Distinct Ecological Preferences

K. K. Nightingale,[1,2]* K. Lyles,[1] M. Ayodele,[1] P. Jalan,[3] R. Nielsen,[4] and M. Wiedmann[1]

Department of Food Science, Cornell University, Ithaca, New York[1]; Department of Animal Sciences, Colorado State University, Fort Collins, Colorado[2]; Department of Biostatistics and Computational Biology, Cornell University, Ithaca, New York[3]; and Center for Bioinformatics and Institute of Biology, University of Copenhagen, Copenhagen, Denmark[4]

Received 22 March 2006/Returned for modification 30 April 2006/Accepted 27 May 2006

While phylogenetic and cluster analyses are often used to define clonal groups within bacterial species, the identification of clonal groups that are associated with specific ecological niches or host species remains a challenge. We used *Listeria monocytogenes*, which causes invasive disease in humans and different animal species and which can be isolated from a number of environments including food, as a model organism to develop and implement a two-step statistical approach to the identification of phylogenetic clades that are significantly associated with different source populations, including humans, animals, and food. If the null hypothesis that the genetic distances for isolates within and between source populations are identical can be rejected (SourceCluster test), then particular clades in the phylogenetic tree with significant overrepresentation of sequences from a given source population are identified (TreeStats test). Analysis of sequence data for 120 *L. monocytogenes* isolates revealed evidence of clustering between isolates from the same source, based on the phylogenies inferred from *actA* and *inlA* ($P = 0.02$ and $P = 0.07$, respectively; SourceCluster test). Overall, the TreeStats test identified 10 clades with significant ($P < 0.05$) or marginally significant ($P < 0.10$) associations with defined sources, including human-, animal-, and food-associated clusters. Epidemiological and virulence phenotype data supported the fact that the source-associated clonal groups identified here are biologically valid. Overall, our data show that (i) the SourceCluster and TreeStats tests can identify biologically meaningful source-associated phylogenetic clusters and (ii) *L. monocytogenes* includes clonal groups that have adapted to infect specific host species or colonize nonhost environments.

Recent advances in DNA sequencing technologies have made DNA sequence-based subtyping methods such as multilocus sequence typing (MLST) widely available and affordable (2, 4), and as a result, a rich source of data for evolutionary and population genetics analyses of pathogenic and nonpathogenic bacteria is available (30). Traditional phylogenetic methods are frequently used to re-create the evolutionary history of a group of bacterial isolates and identify clonal groups of isolates that share a recent common ancestor. These methods also allow preliminary identification of clonal groups that show obvious associations with specific ecological niches, hosts, or disease symptoms in affected hosts. For example, MLST studies with *Campylobacter jejuni* showed that some clonal complexes within this species are exclusively or predominantly associated with specific host populations, including clonal complexes that were associated with human hosts and specific animal host species, such as sheep or poultry (5, 6). Similarly, Luan et al. (13) used MLST analysis to describe a specific clonal complex among invasive group B *Streptococcus* isolates which was associated with neonatal infections but which rarely caused invasive disease in adults. A recent MLST study of *Candida albicans* also showed that isolates from different infection sites varied in their distribution among phylogenetic clades (29), indicating that clonal complexes in this fungal pathogen may have evolved to demonstrate tissue specificity in infected hosts. While previous studies have used traditional statistical and phylogenetic methods to make inferences about the distribution of isolates within a phylogeny, these studies usually evaluated only specific clades which showed exclusive or almost exclusive associations with isolation from a particular niche. There is thus a clear need for the development of an unbiased statistical approach to assess the significance of phylogenetic clustering between isolates from the same source as a first step for the identification of niche-adapted clonal groups.

Listeria monocytogenes, a facultative intracellular pathogen that may cause severe invasive infections in humans and more than 40 species of animal hosts (25, 32), was chosen as a model organism for the development and implementation of a novel two-step statistical approach for the unbiased identification of phylogenetic clades that are significantly associated with different source populations, i.e., humans, animals, and food. *L. monocytogenes* was chosen as a model since it not only causes human and animal food-borne infections but also is commonly isolated from different environmental sources (e.g., soil, surface water, vegetation, and manure) and from food (7, 8). Consequently, a number of epidemiological studies have previously been performed to identify the *L. monocytogenes* subtypes and clonal groups that may differ in their ability to cause disease (9, 19, 20, 33). In general, the results of these previous studies supported the finding that *L. monocytogenes* represents

* Corresponding author. Mailing address: Colorado State University, Department of Animal Sciences, 108B Animal Science Building, Fort Collins, CO 80523-1171. Phone: (970) 491-1556. Fax: (970) 491-5326. E-mail: kendra.nightingale@colostate.edu.

two major genetic lineages (termed lineages I and II) and a third minor lineage (lineage III) that appear to differ in their abilities to cause human disease (3, 9, 21, 33). *L. monocytogenes* isolates representing lineage I, which includes serotypes 1/2b, 4b, and 3b (17), are more prevalent among human clinical isolates (9, 11, 16, 31), while lineage II, which includes isolates belonging to serotypes 1/2a, 1/2c, and 3a (17), even though they are regularly isolated from human clinical cases, are overrepresented among *L. monocytogenes* strains isolated from food (9, 31). Lineage III includes isolates belonging to serotypes 4a, 4b, and 4c; and animal clinical isolates are overrepresented among this third rare lineage (11, 17, 31, 33). Modeling of the dose-response relationships for these *L. monocytogenes* lineages also supported the finding that lineage I strains show higher levels of virulence for humans than lineage II strains (3). Previous studies (9, 20, 33) performed by use of a tissue culture plaque assay also showed that lineage I isolates, on average, formed larger plaques than lineage II isolates, indicative of an enhanced ability of lineage I isolates to spread intracellularly between host cells.

While previous studies clearly indicate that the genetic lineages within *L. monocytogenes* differ in their virulence characteristics and their association with different source populations (e.g., human, animal, and food), each *L. monocytogenes* lineage contains considerable subtype diversity and subtypes that are associated with isolation from specific source populations (9). Further studies are thus needed to probe the evolution of niche adaptation and to define biologically meaningful clonal groups within the major *L. monocytogenes* genetic lineages. Specifically, there is a clear need to differentiate *L. monocytogenes* clonal groups that are adapted to infect humans and that may have virulence enhanced over that of clonal groups that are adapted to colonize the environment and that may have limited virulence. We thus used *L. monocytogenes* as a model organism to (i) develop and implement a novel statistical approach for the unbiased identification of phylogenetic clades that are significantly associated with a particular source population (i.e., humans, animals, or food) and (ii) validate the biological significance of clades associated with specific source populations using tissue culture pathogenicity data and epidemiological information.

MATERIALS AND METHODS

Bacterial isolates. A set of 120 geographically and temporally matched *L. monocytogenes* isolates from human ($n = 60$) and animal ($n = 30$) clinical cases as well as food samples ($n = 30$) was used in this study. These isolates were selected from a larger set of *L. monocytogenes* isolates collected in New York State between January 1999 and December 2001 ($n = 354$), as described in detail in a previous study (18). All isolates were previously characterized by EcoRI ribotyping (24) and were assigned to one of three genetic lineages as described by Wiedmann et al. (33). In addition, all 120 isolates were previously characterized by an MLST scheme that included partial sequencing of *actA*, *inlA*, *gap*, *prs*, *purM*, *ribC*, and *sigB* (18). A detailed description from a previous study (18) of the isolate set investigated here (including isolate source, serotype, EcoRI ribotype, allelic types for all seven genes, and MLST type) is available elsewhere (http://www.foodscience.cornell.edu/wiedmann/Nightingale%20Supplementary.txt). DNA sequence data for these 120 *L. monocytogenes* isolates are available for download at www.pathogentracker.com. From the seven genes sequenced in our previous MLST study (18), we selected partial sequences for the two virulence genes (*actA* and *inlA*) and a concatenated sequence composed of partial sequences for the housekeeping genes *gap* and *prs* and stress response gene *sigB* to evaluate niche adaptation in *L. monocytogenes*. These two housekeeping genes and the stress response gene were previously shown to have evolved primarily by point mutation and thus provide a data set suitable as a reliable probe for *L. monocytogenes* niche adaptation at the core genetic level (18). A concatenated housekeeping and stress response gene sequence rather than individual gene sequences were used here, as the concatenated sequence provided a more robust phylogeny due to the limited number of polymorphisms found in each individual gene (18). The virulence genes *actA* and *inlA* were selected because we expected that these genes would be the most likely to show evidence of adaptive evolution, as these genes play key roles in host-pathogen interactions (12).

Frozen stocks of *L. monocytogenes* isolates were maintained at −80°C in brain heart infusion (Difco, Detroit, MI) broth with 15% (wt/vol) glycerol.

Mouse cell plaque assay. The in vitro virulence phenotype of all 120 *L. monocytogenes* isolates was evaluated by a plaque assay with mouse L cells, which was performed essentially as described previously (9, 26, 27). Briefly, *L. monocytogenes* isolates were grown in brain heart infusion broth overnight (18 h) at 30°C without shaking. Overnight bacterial cultures (1 ml) were pelleted and resuspended in phosphate-buffered saline (PBS; pH 7.4), and 1:10 serial dilutions were performed in PBS. Mouse L cells were grown to confluence in treated flat-bottom tissue culture six-well plates (Corning; Acton, MA), and approximately 1.5×10^5 or 4.0×10^6 CFU of a given *L. monocytogenes* isolate was inoculated into one well containing an L-cell monolayer. The lineage II, EcoRI ribotype DUP-1030A, standard laboratory control strain 10403S (1) was included as an internal control in each plaque assay. Plaques were visualized by staining with neutral red (Sigma Chemical, St. Louis, MO), as described previously (9, 26, 27), and images of infected L cells were captured with a digital scanner (Perfection 1650; Epson, Long Beach, CA). The area of approximately 25 plaques, selected to represent each *L. monocytogenes* isolate, was measured with SigmaScan Pro software (version 5.0; Statistical Solutions, Saugus, MA). The ability of each *L. monocytogenes* isolate to spread from cell to cell was expressed as the average plaque size for a given *L. monocytogenes* isolate relative to the average plaque size for 10403S, which was set equal to 100%. Two independent plaque assays were performed for all 120 *L. monocytogenes* isolates studied here.

Phylogenetic analysis. In a previous study (18), the 120 *L. monocytogenes* isolates described above were assigned sequence types (STs) on the basis of MLST analyses that included partial gene sequences for two key virulence genes (*actA* and *inlA*), a stress response gene (*sigB*), two hypervariable (15) housekeeping genes (*purM* and *ribC*), and two slowly evolving (18) housekeeping genes (*gap* and *prs*). While we previously reported phylogenetic trees inferred from a single isolate selected to represent each of the 52 unique STs (18), phylogenetic trees based on DNA sequence data for all 120 *L. monocytogenes* isolates were constructed here to allow the implementation of the novel statistical approach to detection of the same-source clustering described below.

MODELTEST (22) was used to optimize the input parameters to infer maximum-likelihood phylogenetic trees in PAUP* (28), based on partial sequences of two key virulence genes (i.e., *actA* and *inlA*) and a concatenated sequence representing two housekeeping genes (*gap* and *prs*) and the stress response gene *sigB*. Heuristic searches were performed by using equal weights for all sites, and the tree-bisection-reconnection branch-swapping algorithm was used. While unrooted maximum-likelihood phylogenies were used as inputs for the TreeStats test (described below), phylogenies were displayed here as rooted phylograms (see Fig. 1A to C). A homologous concatenated gene sequence from *Bacillus subtilis* (http://genolist.pasteur.fr/SubtiList/) was used as the outgroup for the concatenated housekeeping and stress response gene phylogeny (Fig. 1C), while the three lineage III sequences (i.e., FSL F2-525, FSL F2-655, and FSL-F2-695) were defined as the outgroups for the *inlA* and *actA* phylogenies (Fig. 1A and B), as described previously (18), because homologous *B. subtilis* sequences are not available for these sequences.

Identification of phylogenetic clades that are significantly associated with different source populations. To test if isolates from different source populations, i.e., humans, animals, and food samples, were evolutionarily related, we used a systematic two-step approach. We first tested the null hypothesis that the genetic distances among isolates within and between source populations were identical (SourceCluster test). If this null hypothesis can be rejected, then there is significant evidence for overall clustering between isolates from the same source population within a phylogeny, and specific clades with a significant overrepresentation of sequences from a given source population can be reliably identified (TreeStats test).

The SourceCluster test was performed by using uncorrected pairwise distance matrices (raw distance) generated with PAUP* software (28) from (i) *actA* sequences, (ii) *inlA* sequences, and (iii) a concatenated sequence of two housekeeping and a stress response gene (i.e., *gap*, *prs*, and *sigB*) for the 120 *L. monocytogenes* isolates described above. The SourceCluster test used the average genetic distance (raw distance) between sequences from the same source (i.e.,

human [H], animal [A], and food [F] samples) as a test statistic (T) calculated as follows:

$$T = \sum_{k\epsilon\{H,A,F\}} \frac{\sum_{i,j:i\neq j} d_{ij} I(S_i = k) I(S_j = k)}{\sum_{i,j:i\neq j} I(S_i = k) I(S_j = k)} \quad (1)$$

where d_{ij} is the raw genetic distance between isolates i and j, S_i is the source of isolate i, and $I(S_i = k)$ is an indicator function that returns a value of 1 if isolate i is from source population k. Significance is determined by a procedure that permutes the labeling of the source population among isolates to generate 1,000 new data sets. For each resampled data set, a new value of T, T^*, is calculated; and if T falls in the 5th percentile of the distribution of T^*, the null hypothesis of no clustering between isolates from the same source population is rejected. The relative position of T in the distribution of T^* was expressed as a P value, which was calculated as the number of T^* values that exceeded T divided by the number of resamplings. For example, if 10 of 1,000 T^* estimates were greater than T, the P value would be 0.01. The program SourceCluster is publicly available at http://www.foodscience.cornell.edu/wiedmann/SourceCluster.txt.

The TreeStats test was performed by using unrooted maximum-likelihood phylogenetic trees for the 120 *L. monocytogenes* isolates inferred from (i) *actA* sequences, (ii) *inlA* sequences, and (iii) a concatenated sequence of two housekeeping and a stress response gene (i.e., *gap*, *prs*, and *sigB*). In the TreeStats test, for each clade on a given phylogenetic tree, the expected number of isolates belonging to group A, F, and H, under the null hypothesis of no clustering between isolates from the same source population, was calculated by permuting the labeling of the leaf clades 10,000 times. A chi-square goodness-of-fit test was then used to compare the observed to the expected numbers for each clade on the tree. This procedure is implemented in the program TreeStats, which is publicly available at http://www.foodscience.cornell.edu/wiedmann/TreeStats.htm. Clades with P values <0.05 and <0.10 were considered to show statistically significant and marginally significant evidence for clustering among isolates belonging to the same source population, respectively. No correction for multiple testing was performed, since the SourceCluster test was used to first test the null hypothesis of no clustering between isolates from the same source population. Significant and marginally significant P values from TreeStats analysis were superimposed on the maximum -likelihood trees that were constructed and rooted with an appropriate outgroup as described above.

Statistical analysis. Chi-square tests of independence or Fisher's exact tests (if appropriate; i.e., if more than 25% of expected values in a table were less than five) were used to probe associations between *L. monocytogenes* source populations (i.e., human, animal, or food) and *L. monocytogenes* genotypes (i.e., genetic lineage and EcoRI ribotype). Human and animal clinical isolate were also pooled to create a categorical variable termed "host" to evaluate associations between this source population and *L. monocytogenes* genotypes. Categorical analyses were performed only for the EcoRI ribotypes that were observed at least four times among the 120 *L. monocytogenes* isolates studied here.

One-way analysis of variance was used to determine the relationship between the measure "plaque size" and the categorical variables "lineage," "source," and "ribotype." Analyses were performed only for ribotypes that were observed at least four times among the 120 isolates. All analyses were performed with Statistical Analysis Systems software (SAS Institute, Cary, NC). While P values of <0.05 were considered statistically significant, P values of <0.10 are also reported and were considered marginally significant.

RESULTS AND DISCUSSION

We combined MLST, epidemiology, and virulence phenotype data for 120 *L. monocytogenes* isolates from human and animal clinical cases as well as food samples to develop, implement, and biologically validate a novel two-step statistical approach for the unbiased identification of phylogenetic clades that are significantly associated with different source populations. Our results indicate (i) that the combined SourceCluster and TreeStats approach identifies biologically meaningful source-associated clonal groups, providing evidence for niche adaptation within the major *L. monocytogenes* genetic lineages, and (ii) that *L. monocytogenes* contains both host- and non-host-adapted clonal groups. The power of the combined SourceCluster and TreeStats approach for the identification of niche-adapted clonal groups is further supported by the fact that most of the source-associated clades identified in our study represent previously described subtypes with unique epidemiological or virulence characteristics, including (i) the major human epidemic clones (ECs) and (ii) subtypes with mutations that result in premature stop codons in the key *L. monocytogenes* virulence gene *inlA*, which are further associated with a significantly reduced ability to invade human intestinal epithelial cells (19). While it is well known that *L. monocytogenes* is an environmental pathogen (8), our data support the finding that clonal groups within this species may differ in their ability to cause human disease. Significant associations between clades in a given gene tree and a source population do not necessarily indicate that specific polymorphisms in a given gene are responsible for the source association. In fact, the evolution of linked genetic traits by different mechanisms (e.g., gene acquisition, gene loss, and polymor-

TABLE 1. Distribution of *L. monocytogenes* molecular subtypes among isolate sources

Molecular subgroup (lineage)	No. (%) of isolates representing molecular subtype from each source[a]				
	Host			Food	Total (n = 120)
	Human	Animal	Total		
Lineage I	39 (56.5)#	17 (24.6)	56 (81.1)#	13 (18.8)	69 (57.5)
Lineage II	18 (37.5)	13 (28.1)	31 (64.6)	17 (35.4)*	48 (40.0)
Lineage III	3 (5.0)	0 (0.0)	3 (100.0)	0 (0.0)	3 (2.5)
DUP-1038B (I)	8 (72.7)	3 (27.3)	11 (100.0)*	0 (0.0)	11 (9.2)
DUP-1039C (II)	3 (27.3)	4 (36.4)	7 (63.6)	4 (36.4)	11 (9.2)
DUP-1042B (I)	11 (57.9)	7 (36.8)	18 (94.7)*	1 (5.3)	19 (15.8)
DUP-1044A (I)	5 (71.4)	1 (14.3)	6 (85.7)	1 (14.3)	7 (5.8)
DUP-1052A (I)	5 (62.5)	1 (12.5)	6 (75.0)	2 (25.0)	8 (6.7)
DUP-1053A (II)	7 (100.0)**	0 (0.0)	7 (100.0)	0 (0.0)	7 (5.8)
DUP-1062A (II)	0 (0.0)	0 (0.0)	0 (0.0)	4 (100.0)**	4 (3.3)
Total	60 (50.0)	30 (25.0)	90 (75.0)	30 (25.0)	120 (100.0)

[a] Molecular subgroups that are significantly associated with a given source population (i.e., human, animal, host, or food) compared to their associations with all other source populations (e.g., food versus human and animal). Associations were determined by chi-square tests of independence or Fisher's exact test (if more than 25% of the expected values in a given table were less than five). Significance was marginal (#; $P < 0.10$) or P values of <0.05 (*) and <0.01 (**).

TABLE 2. Summary of associations between plaque size and categorical variables

Categorical variable	Group	No. of isolates	Mean plaque size[b]
Source			
	Human	60	110.9
	Animal	30	115.3
	Food	30	100.7*
Genetic lineage[a]			
	Lineage I	69	116.5**
	Lineage II	48	98.8***
	Lineage III	3	116.5
Ribotype			
	DUP-1038B	11	124.6*
	DUP-1039C	11	99.7
	DUP-1042B	19	120.2*
	DUP-1044A	7	118.6
	DUP-1052A	8	94.6#
	DUP-1053A	6	91.0#
	DUP-1062A	4	94.9

[a] Assigned as previously described by Wiedmann et al. (33).

[b] Isolate groups shown to have mean plaque sizes that are signficantly different from those for all isolates belonging to other groups within each categorical variable analyzed (e.g., lineage I versus lineages II and III), as determined by one-way analysis of variance, are indicated by marginal significance (#; $P < 0.10$) and P values of <0.05 (*), <0.01 (**), and <0.0001 (***). Statistical analyses were performed only for ribotypes that were observed at least four times among the 120 isolates studied here.

phisms within linked genes) may allow host tropism and niche adaptation.

The combined SourceCluster and TreeStats approach identifies biologically relevant phylogenetic clades, which provide evidence for niche adaptation within *L. monocytogenes* lineages. Initial categorical analyses showed that the distribution of the 120 *L. monocytogenes* isolates studied here among source populations, genetic lineages, and EcoRI ribotypes was consistent with that found in previous studies, which were often based on larger *L. monocytogenes* isolate sets (9, 11, 16, 31). Specifically, *L. monocytogene*s lineage I was marginally overrepresented ($P = 0.07$) among isolates from human listeriosis cases, while lineage II was significantly ($P = 0.04$) overrepresented among food isolates (Table 1). Determination and analyses of mean plaque sizes, a measure for the ability of a given *L. monocytogenes* strain to spread intracellularly between mammalian host cells, indicated that isolates representing lineage II produced significantly ($P < 0.0001$) smaller plaques than isolates belonging to lineages I and III and, similarly, that isolates belonging to lineage I formed significantly ($P < 0.01$) larger plaques than lineage II and III isolates (Table 2), also consistent with previous studies (9, 20, 33). Interestingly, *L. monocytogenes* isolated from food samples also formed significantly ($P = 0.03$) smaller plaques than human and animal clinical isolates (Table 2). Since isolates that form larger plaques have accelerated rates of intracellular spread (26), these data further suggest that lineage I isolates are more virulent than those belonging to lineage II and, similarly, that isolates from clinical cases of listeriosis (i.e., both human and animal cases) are typically more virulent than isolates from food, providing preliminary evidence for niche adaptation within *L. monocytogenes*. Our analyses also support the finding that the collection of 120 *L. monocytogenes* isolates studied here is representative of the previously reported genetic and phenotypic diversity for this pathogen and is thus appropriate for use for the development and biological validation of a new statistical approach to the identification of the phylogenetic clades associated with specific source populations.

In order to allow the unbiased identification of phylogenetic clades that are associated with different sources populations, rather than using clades defined a priori (e.g., the major *L. monocytogenes* genetic lineages), we developed a two-step statistical approach to the analysis of phylogenetic trees constructed for the 120 *L. monocytogenes* isolates included in our study for evidence of clustering between *L. monocytogenes* isolates from a given source population (i.e., humans, animals, and food). The SourceCluster test was first used to test the null hypothesis that the genetic distances for isolates within and between source populations were identical. Results from SourceCluster analysis indicated that the genetic distances for isolates within and between source populations were not identical, based on *actA* sequences ($P = 0.02$), providing overall evidence of clustering between isolates from the same source. While marginally significant evidence for overall clustering of *L. monocytogenes* isolates from the same source population ($P = 0.07$) was observed for *inlA* sequences, no evidence for overall clustering between isolates from same source population ($P = 0.29$) was found on the basis of the concatenated housekeeping and stress response gene sequence data. Subsequent analyses by use of the TreeStats test identified a number of clades within each phylogeny that showed significant ($P < 0.05$) clustering of isolates from a specific source population. Importantly, the results from the TreeStats analyses found that *L. monocytogenes* lineages I and II represent clades that show significant associations with specific source populations ($P < 0.05$ for *inlA* and housekeeping gene-derived phylogenies and $P < 0.10$ for the *actA* phylogeny; Fig. 1A to C). These findings not only are consistent with our plaque assay data and the results of the categorical analyses, which show that lineages I and II differ in their distributions among source populations and in their ability to spread from cell to cell, but are also consistent with the findings of previous prevalence studies (9, 11, 16, 31) that showed that human clinical isolates are overrepresented among lineage I isolates, while lineage II isolates are more commonly isolated from food. While one might expect a stronger association between source population and lineages in the *actA* phylogeny, particularly since the lineages differ in their ability to spread from cell to cell, as indicated by the average plaque sizes (9), variation in virulence genes other than *actA* (e.g., *prfA*) or ActA amino acid residues outside the region sequenced here may also have contributed to cell-to-cell-spread phenotype. The strong association between the source population and lineage in the *inlA* tree suggests that variation in InlA-mediated attachment and invasion is fundamental for niche adaptation in *L. monocytogenes* and is consistent with the fact that naturally occurring polymorphisms in *inlA* (e.g., nonsense mutations that lead to premature stop codons) have a significant effect on the invasiveness of *L. monocytogenes* for human intestinal epithelial cells (19). It is, however, important to recognize that significant associations of clades in a given gene tree with a specific source population do not necessarily indicate that the given gene is solely responsible

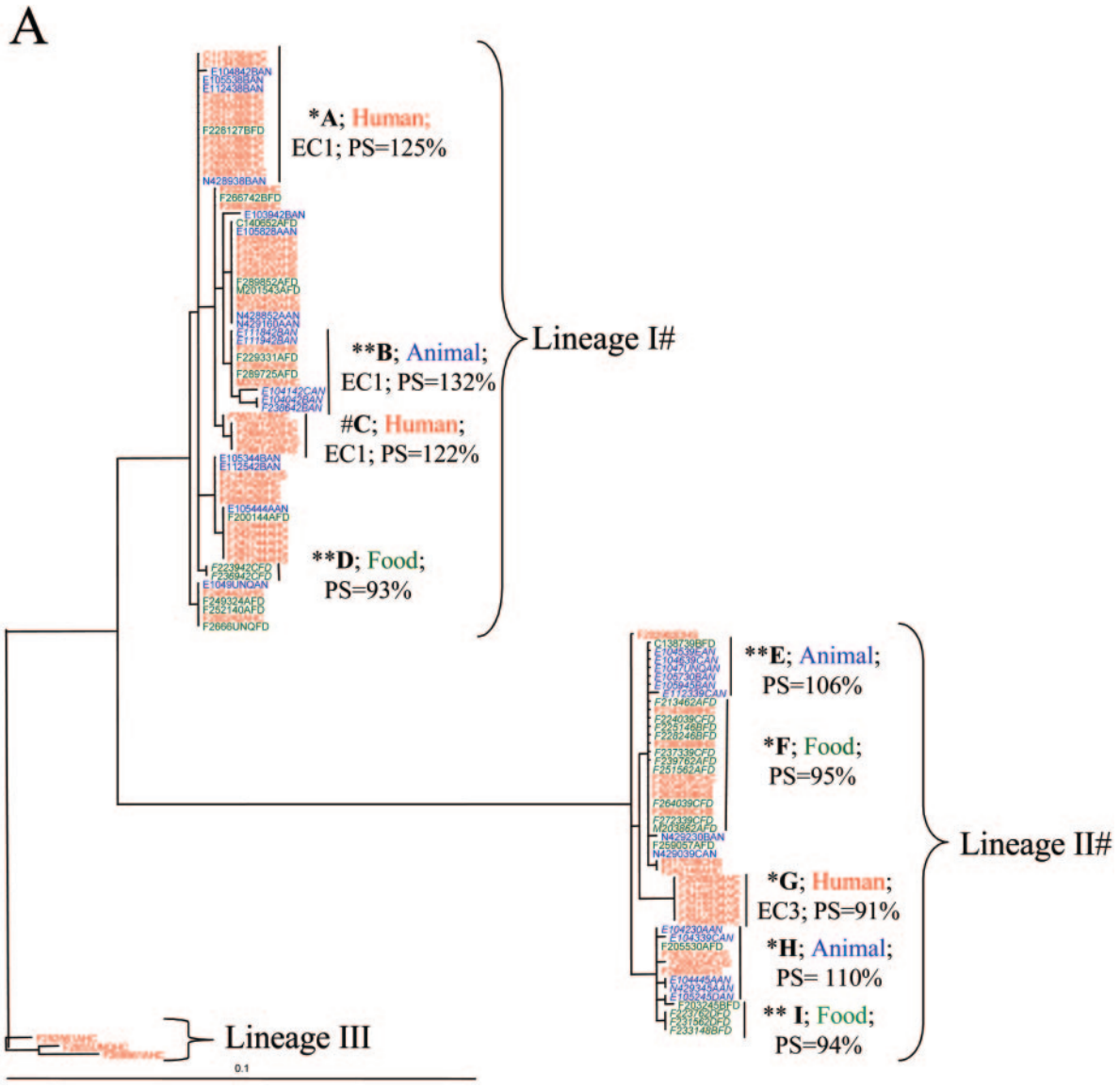

FIG. 1. Maximum-likelihood phylogenetic trees inferred from (A) *actA*, (B) *inlA*, and (C) concatenated *gap*, *prs*, and *sigB* sequences for a sample set containing 60 human clinical (red), 30 animal clinical (blue), and 30 food (green) *L. monocytogenes* isolates. Taxon labels include the Cornell Food Safety Laboratory Culture Collection isolate name (e.g., F2655 represents isolate FSL F2-655), EcoRI ribotype (e.g., 44A represents ribotype DUP-1044A), and source (e.g., human isolate from New York State Department of Health [HS], human isolate from New York City Department of Health and Mental Hygiene [HC], animal isolate [AN], and food isolate [FD]). Isolates in clades that show significant or marginally significant associations with source populations (i.e., humans, animals, and food), as identified by the TreeStats test, are indicated by italics and are marked with a vertical bar; significance levels are indicated by $*$ and $**$ (which represent $P < 0.05$ and $P < 0.01$, respectively) and # (which indicates marginally significant associations; $P < 0.10$). Clades that are significantly associated with source populations in at least one phylogeny and that contain the same or overlapping isolates are designated with the same letter (A to J) across all three phylogenies (Table 3); the predominant source population for each clade (i.e., human, animal, or food) is also indicated. The mean plaque size (PS) for isolates in each cluster as well as classification into ECs (as described by Kathariou [12]) is also indicated at each significant source-associated clade.

for the source association. Analyses of additional virulence gene phylogenies (e.g., phylogenies based on other internalin genes, such as *inlB*) as well as analysis of potentially linked gene indels or islands (e.g., through microarray analysis) are likely to provide further insight into the genetic mechanisms underlying niche adaptation in *L. monocytogenes*.

Implementation of SourceCluster and TreeStats tests can identify biologically meaningful same-source clades in phylogenetic trees. Because the SourceCluster and TreeStats tests identified the same, biologically validated, source-associated lineages as previous studies (9, 11, 16, 31), this approach has the potential to identify specific clades representing clonal groups that may have common biologically relevant characteristics, consistent with niche adaptation within the major genetic lineages. Specifically, the TreeStats test identified eight, six, and five clades with significant ($P < 0.05$) source associations

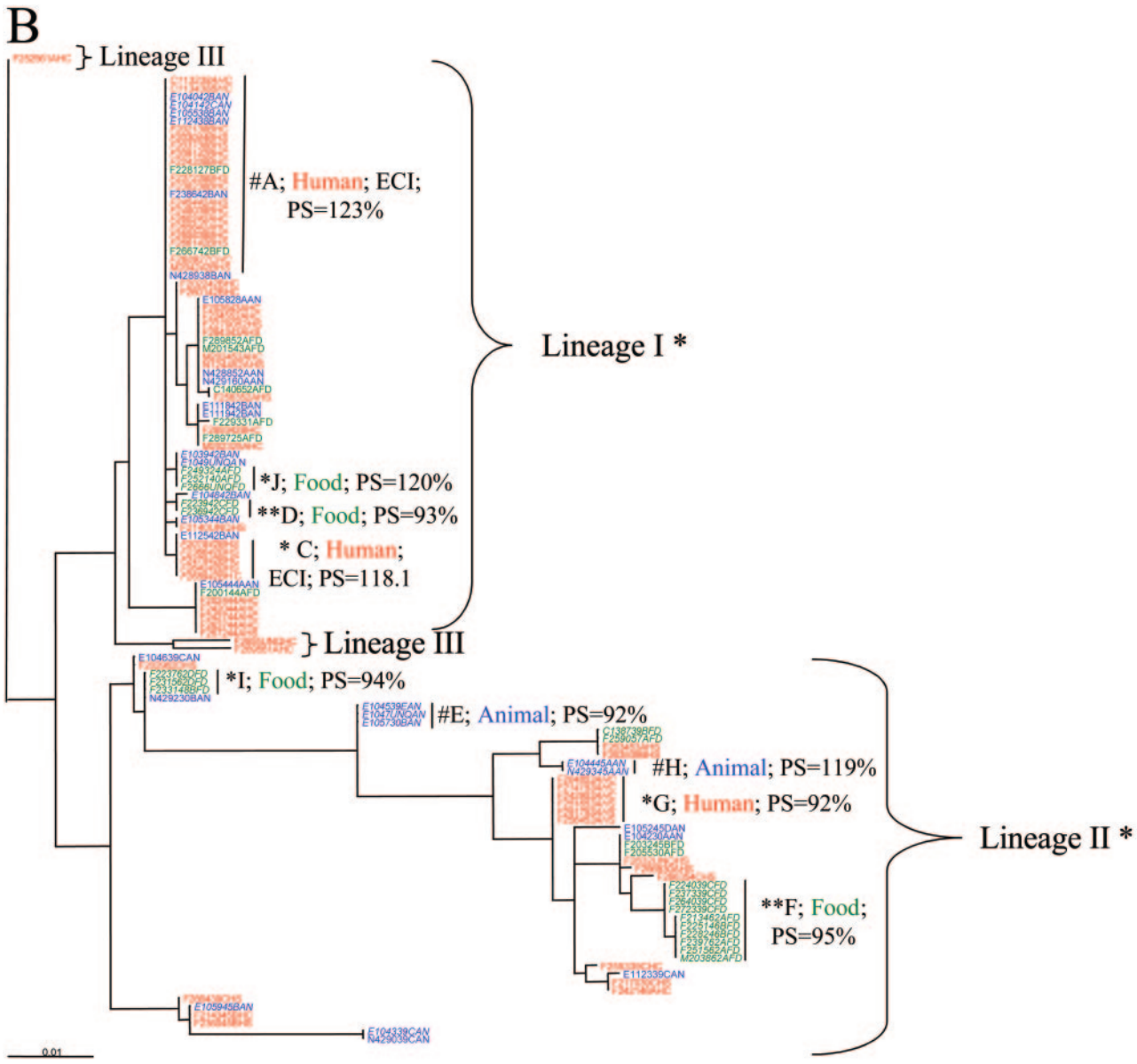

FIG. 1—*Continued.*

in the *actA* (Fig. 1A), *inlA* (Fig. 1B), and concatenated housekeeping and stress response gene (Fig. 1C) phylogenies, respectively. In addition, one clade each in the *actA* and the concatenated housekeeping and stress response gene phylogeny as well as three clades in the *inlA* phylogeny showed marginal ($P < 0.10$) evidence for associations with specific source populations. Overall, a total of 10 clades covering the same or overlapping isolates showed significant associations with specific source populations across the three phylogenies analyzed here, and all 10 of these clades were identified independently in at least two phylogenies (Table 3). For example, clade F (Table 3; Fig. 1) was identified as a significant ($P < 0.05$) source-associated clade in all three phylogenies. Additionally, each of the 10 clades listed in Table 3 was identified as a significant ($P < 0.05$) source-associated clade in at least one phylogeny. For example, while clade A was identified only as a marginally significant ($P < 0.10$) source-associated clade in the *inlA* phylogeny, this clade was significantly ($P < 0.05$) overrepresented by isolates obtained from human clinical cases in the *actA* phylogeny (Fig. 1). While we appreciate that the source-associated clades identified by TreeStats analysis of the phylogeny inferred from the concatenated housekeeping and stress response gene sequence should be interpreted carefully, since the overall SourceCluster statistic was not significant, the overlap between significant clades identified in the different trees supports the robustness of the TreeStats approach, particularly when it is applied to pathogens with a highly clonal population structure, such as *L. monocytogenes* (18).

Of the 10 clades that were identified as being significantly associated with different source populations, three were characterized by an overrepresentation of isolates from human clinical cases, including two clades within lineage I (clades A

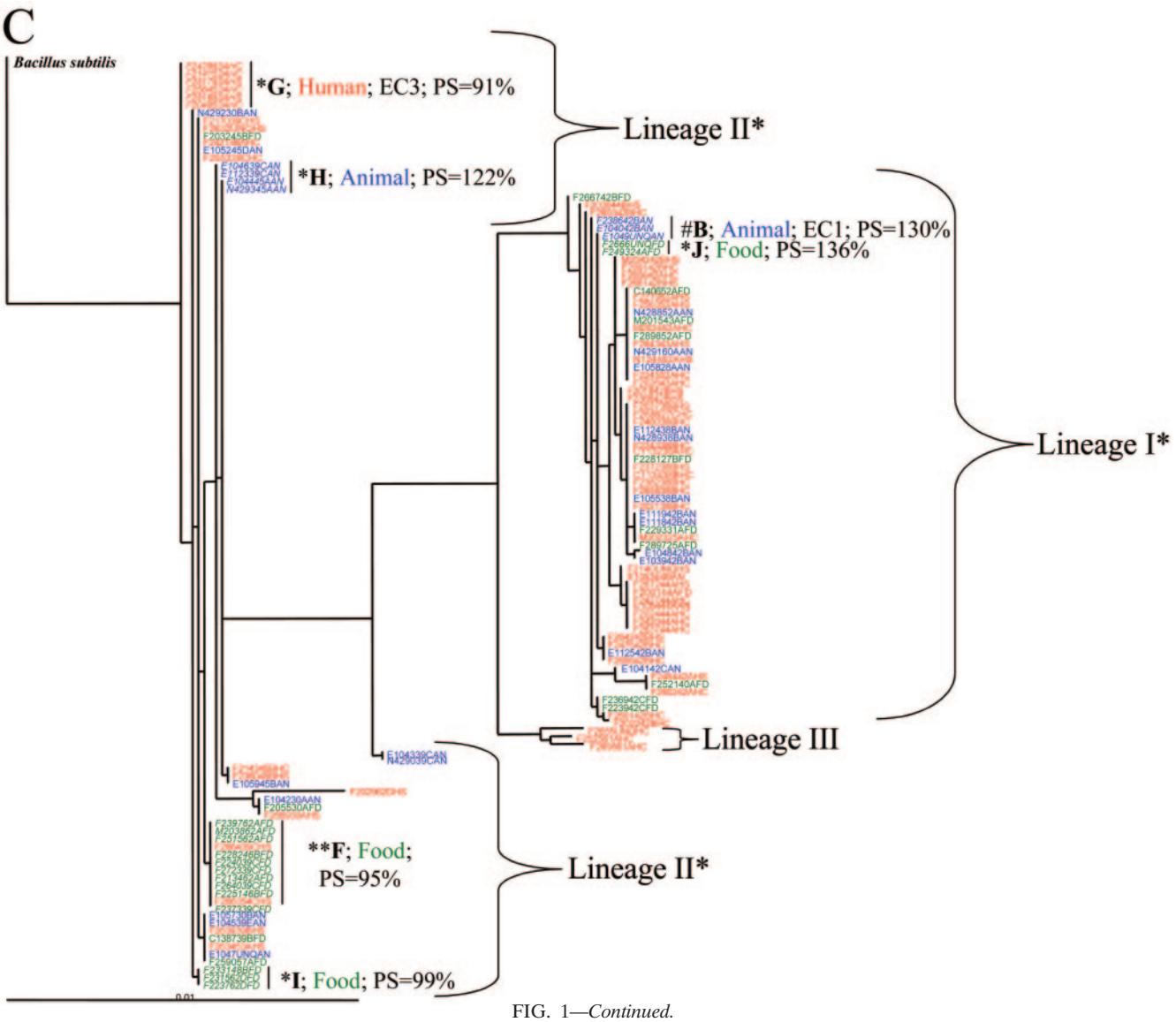

FIG. 1—*Continued.*

and C) and one clade within lineage II (clade G) (Fig. 1A and B and Table 3). Interestingly, all three of these clades not only correlate with previously identified human ECs (12), including ECI (clade A and C) and ECIII (clade G), but also correlate with EcoRI ribotypes previously identified to be associated with human listeriosis epidemics (9, 11), including ribotypes DUP-1038B (clade A), DUP-1042B (clade C), and DUP-1053A (clade G). Our results also suggest that ECI represents two distinct clones (clades A and C), as supported by previous reports that ECI represents two distinct ribotypes (9). Interestingly, *L. monocytogenes* isolates classified into clades A and C showed some of the largest mean plaque sizes observed among all 10 of the source-associated clades identified here (Fig. 1A and B and Table 3), consistent with the observation that isolates of the ribotypes associated with these clades (DUP-1038B and DUP-1042B) also formed significantly ($P < 0.05$) larger plaques than isolates belonging to other ribotypes (Table 2). Our findings not only further support the biological significance of the human-associated clades identified here, but they also indicate that different clades associated with human clinical cases may show distinct phenotypic characteristics, as isolates in the human-associated clade G (which belongs to lineage II and ribotype DUP-1053A) produced relatively smaller (91.0%) plaques (Table 3). Interestingly, lineage II ribotype DUP-1053A isolates appear to be characterized by enhanced invasiveness for human intestinal epithelial Caco-2 cells (K. Nightingale and M. Wiedmann, unpublished data) compared to the invasiveness of the standard lineage II laboratory control strain 10403S, providing a potential biological explanation for their association with human hosts. These findings support the importance of future studies on the host-associated clades identified here and serve as a reminder that while in vitro assays such as the plaque assay used here may provide useful information on the virulence characteristics of *L. monocytogenes* isolates, virulence for humans cannot be fully measured by an in vitro assay.

TABLE 3. Description of significant same-source isolate clusters in *L. monocytogenes* phylogenies inferred from key virulence genes (*actA* and *inlA*) and a concatenated housekeeping gene sequence composed of *gap*, *prs*, and *sigB*

Clade	Lineage[a]	Source[b]	Cluster description[c]	Cluster observed in								
				actA			*inlA*			Concatenated housekeeping gene		
				Serotype (no. of isolates)	Ribotype (no. of isolates)	Avg plaque size	Serotype (no. of isolates)	Ribotype (no. of isolates)	Avg Plaque size	Serotype (no. of isolates)	Ribotype (no. of isolates)	Avg Plaque size
A	I	Human	Epidemic clone 1	4b (11)	DUP-1027B (1) DUP-1038A (1) DUP-1038B (8) DUP-1044B (1)	125.3	1/2b (1), 4b (15), NT (1)	DUP-1027C (1) DUP-1038B (9) DUP-1042A (3) DUP-1042B (3) DUP-1044B (1)	123.3			
B	I	Animal	Epidemic clone I	1/2b (5)	DUP-1042B (4) DUP-1042C (1)	131.5				1/2b (3)	DUP-1042B (2) Unique (1)	129.7
C	I	Human	Epidemic clone I	4b (6)	DUP-1042A (1) DUP-1042B (5)	122.0	4b (5)	DUP-1042B (5)	118.1			
D	I	Food	NA	1/2b (2)	DUP-1042C (2)	92.9	1/2b (2)	DUP-1042C (2)	92.9			
E	II	Animal	NA	1/2a (6)	DUP-1030B (1) DUP-1039C (2) DUP-1039E (1) DUP-1045B (1) Unique (1)	105.7	1/2a (3)	DUP-1030B (1) DUP-1039E (1) Unique (1)	92.2			
F	II	Food	Premature *inlA* stop codons	1/2a (9), NT[d] (1)	DUP-1039C (4) DUP-1046B (2) DUP-1062A (4)	94.6	1/2a (9), NT[d] (1)	DUP-1039C (4) DUP-1046B (2) DUP-1062A (4)	94.6	1/2a (9), NT (1)	DUP-1039C (4) DUP-1046B (2) DUP-1062A (4)	94.6
G	II	Human	Epidemic clone III	1/2a (6)	DUP-1053A (6)	91.0	1/2a (6)	DUP-1053A (6)	91.0	1/2a (6)	DUP-1053A (6)	91
H	II	Animal	NA	1/2a (5)	DUP-1030A (1) DUP-1039C (1) DUP-1045A (2) DUP-1045D (1)	110.9	1/2a (2)	DUP-1045A (2)	118.9	1/2a (4)	DUP-1039C (2) DUP-1045A (2)	121.5
I	II	Food	Premature *inlA* stop codons	1/2a (4)	DUP-1045B (1) DUP-1048B (1) DUP-1062D (2)	98.5	1/2a (3)	DUP-1048B (1) DUP-1062D (2)	93.7	1/2a (3)	DUP-1048B (1) DUP-1062D (2)	93.7
J	I	Food	NA				1/2b (3)	1024A (1) 1040A (1) Unique (1)	119.8	1/2b (2)	DUP-1024A (1) Unique (1)	135.5

[a] Assigned as described by Wiedmann et al. (33); lineage I generally includes serotypes 1/2b, 4b, 3b, and 3c, while lineage II generally includes serotypes 1/2a, 1/2c, and 3a.
[b] Indicates predominant source population for each clade (e.g., human, animal, or food).
[c] Epidemic clone assigned as described by Kathariou (12). Ribotypes associated with premature stop codons in *inlA* identified by Nightingale et al. (19), Orsi and Wiedmann (unpublished), or Nightingale and Wiedmann (unpublished). NA, no epidemiology information available.
[d] NT, not typeable.

Overall, a total of three clades were significantly associated with isolation from animal clinical cases, including one clade within lineage I (clade B) and two clades within lineage II (clades E and H) (Table 3). Interestingly, isolates in clade B predominantly represent ribotype DUP-1042B, which, on average, shows an enhanced ability to spread intracellularly between host cells (Table 2). Our data suggest that DUP-1042B isolates represent two distinct clonal groups that may have adapted to infect specific host species. This observation is further supported by the fact that DUP-1042B isolates in the human- and animal-associated clades (clades C and B, respectively) represent distinct serotypes, with isolates in clades C and B belonging to serotypes 4b and 1/2b, respectively, consistent with the fact that most human listeriosis epidemics are caused by serotype 4b strains (14). While isolates in clades E and H do not show any apparent phenotypic or epidemiological characteristics that explain their association with animal clinical cases, further characterization of isolates belonging to these clades may provide new insights into the evolution of host specificity in *L. monocytogenes*.

Overall, a total of four clades were significantly associated with isolation from contaminated food, including two clades within lineage I (clades D and J) and two clades belonging to lineage II (clades F and I) (Table 3). Interestingly, isolates in clade F belong to three *L. monocytogenes* ribotypes (i.e., DUP-1062A, DUP-1039C, and DUP-1046B) previously shown to carry nonsense mutations that lead to premature stop codons in the virulence gene *inlA*, express a truncated and secreted form of InlA, and demonstrate attenuated invasiveness in Caco-2 cells (19; R. Orsi and M. Wiedmann, unpublished data). We recently identified two similar additional mutations that lead to premature stop codons in *inlA* in isolates belonging to ribotypes DUP-1045B, DUP-1062D, and DUP-1048B, which also demonstrated attenuated invasion of Caco-2 cells (Nightingale and Wiedmann, unpublished), supporting the finding that the second lineage II food-associated clade identified here, clade I, may also represent a human virulence-attenuated clonal group. These findings not only provide a clear biological explanation for the association of clades F and I isolates with food but also are consistent with the findings of a previous study (19), which found that isolates with premature stop codons in *inlA* are significantly underrepresented among human clinical isolates compared to their prevalence in food. On very rare occasions, *L. monocytogenes* isolates containing premature *inlA* stop codons have been linked to human disease (<2% of more than 1,000 human listeriosis cases), suggesting that these strains may have the potential to cause disease in extremely immunosuppressed individuals (19). Isolates in clades I and F also showed, on average, a smaller plaque size (Table 3), indicative of a reduced ability to spread intracellularly in mammalian hosts cells (Table 3), further supporting the hypothesis that isolates in these clades may represent virulence-attenuated clonal groups. Interestingly, ribotype DUP-1042C isolates belonging to food-associated clade D also formed smaller (92.9%) plaques (Table 3). The classification of clade D as a food-associated clade is also consistent with the results from a previous study (9), which included a larger number of DUP-1042C isolates ($n = 14$) and which showed that DUP-1042C isolates were significantly more common among food isolates and were not observed among nearly 500 human clinical isolates. Some DUP-1042C isolates from food were also recently found to be characterized by attenuated invasiveness in Caco-2 cells (Nightingale and Wiedmann, unpublished), providing a possible biological explanation for their classification into a food-associated clade. Interestingly, isolates in food-associated clade J, on average, formed large (120%) plaques (Table 3), further indicating that different clades associated with food may show distinct phenotypic characteristics. Future studies are thus required to fully understand the biological underpinnings for the observed sources associations of specific clades.

Conclusions. While a number of previous studies have provided preliminary evidence that the main *L. monocytogenes* lineages (i.e., lineages I and II) differ in their ability to cause human disease, only a few clonal groups with specific host or niche preferences (e.g., epidemic clones and virulence-attenuated strains with *inlA* mutations) have been described (10, 12, 19, 23). Our findings not only support the finding that previously described epidemic clones and virulence-attenuated strains represent distinct clades within *L. monocytogenes* but also identified a number of additional clades that are significantly associated with specific source populations and may thus represent host- or niche-adapted ecotypes. Most importantly, our data, obtained by using *L. monocytogenes* as a model system, show that a novel two-step statistical approach (SourceCluster and TreeStats analyses) for the unbiased identification of phylogenetic clades that are significantly associated with particular source populations can reliably identify biologically relevant clonal groups that may represent niche-adapted *L. monocytogenes* ecotypes. This approach will thus provide a valuable set of tools for the characterization of other bacterial pathogens as well as nonpathogenic bacteria, allowing the rapid identification of meaningful subtypes and clades that differ in their relevant phenotypic characteristics, without requiring considerable information a priori on the biology of the organism being characterized. As larger sequence data sets for *L. monocytogenes* and other pathogens become available, this approach may also provide an opportunity to identify associations between clades and more narrowly defined hosts and niches (e.g., specific animal species and environments).

ACKNOWLEDGMENTS

This work was supported by National Institutes of Health (NIH) grant R01GM63259 awarded to M. Wiedmann.

We thank Qi Sun (Computational Biology Service Unit, Cornell University) for his expertise in setting up the computing system to perform evolutionary analyses. We also thank Wan-Lin Su for helpful discussions. We are also indebted to Esther Fortes and Alphina Ho for their technical assistance with DNA sequencing.

REFERENCES

1. **Bishop, D. K., and D. J. Hinrichs.** 1987. Adoptive transfer of immunity to *Listeria monocytogenes*: the influence of in vitro stimulation on lymphocyte subset requirements. J. Immunol. **139:**2005–2009.
2. **Cai, S., D. Y. Kabuki, A. Y. Kuaye, T. G. Cargioli, M. S. Chung, R. Nielsen, and M. Wiedmann.** 2002. Rational design of DNA sequence-based strategies for subtyping *Listeria monocytogenes*. J. Clin. Microbiol. **40:**3319–3325.
3. **Chen, Y., W. H. Ross, M. J. Gray, M. Wiedmann, R. C. Whiting, and V. N. Scott.** 2006. Attributing risk to *L. monocytogenes* subgroups: dose response in relation to genetic lineages. J. Food Prot. **69:**335–344.
4. **Clarke, S. C.** 2002. Nucleotide sequence-based typing of bacteria and the impact of automation. BioEssays **24:**858–862.
5. **Colles, F. M., K. Jones, R. M. Harding, and M. C. J. Maiden.** 2003. Genetic diversity of *Campylobacter jejuni* isolates from farm animals and the farm environment. Appl. Environ. Microbiol. **69:**7409–7413.

6. **Dingle, K. E., F. M. Colles, R. Ure, J. A. Wagenaar, B. Duim, F. J. Bolton, A. J. Fox, D. R. A. Wareing, and M. C. J. Maiden.** 2002. Molecular characterization of *Campylobacter jejuni* clones: a basis for epidemiological investigation. Emerg. Infect. Dis. **8:**949–955.
7. **Farber, J. M., and P. I. Peterkin.** 1991. *Listeria monocytogenes*, a food-borne pathogen. Microbiol. Rev. **55:**476–511.
8. **Fenlon, D. R.** 1999. *Listeria monocytogenes* in the natural environment, p. 21–39. *In* E. T. Ryser and E. H. Marth (ed.), *Listeria* listeriosis and food safety, 2nd ed. rev. and expanded. Marcel Decker, Inc., New York, N.Y.
9. **Gray, M. J., R. N. Zadoks, E. D. Fortes, B. Dogan, S. Cai, Y. Chen, V. N. Scott, D. E. Gombas, K. J. Boor, and M. Wiedmann.** 2004. Food and human isolates of *Listeria monocytogenes* form distinct but overlapping populations. Appl. Environ. Microbiol. **70:**5833–5841.
10. **Jacquet, C., M. Doumith, J. I. Gordon, P. M. V. Martin, P. Cossart, and M. Lecuit.** 2004. A molecular marker for evaluating the pathogenic potential of foodborne *Listeria monocytogenes*. J. Infect. Dis. **189:**2094–2100.
11. **Jeffers, G. T., J. L. Bruce, P. L. McDonough, J. Scarlett, K. J. Boor, and M. Wiedmann.** 2001. Comparative genetic characterization of *Listeria monocytogenes* isolates from human and animal listeriosis cases. Microbiology **147:** 1095–1104.
12. **Kathariou, S.** 2002. *Listeria monocytogenes* virulence and pathogenicity, a food safety perspective. J. Food Prot. **11:**1811–1829.
13. **Luan, S., M. Granlund, M. Sellin, T. Lagergard, B. G. Spratt, and M. Norgren.** 2005. Multilocus sequence typing of Swedish invasive group B *Streptococcus* isolates indicates a neonatally associated genetic lineage and capsule switching. J. Clin. Microbiol. **43:**3727–3733.
14. **McLauchlin, J.** 1990. Distribution of serovars of *Listeria monocytogenes* isolated from different categories of patients with listeriosis. Eur. J. Clin. Microbiol. Infect. Dis. **9:**210–213.
15. **Meinersmann, R. J., R. W. Phillips, M. Wiedmann, and M. E. Berrang.** 2004. Multilocus sequence typing of *Listeria monocytogenes* by use of hypervariable genes reveals clonal and recombination histories of three lineages. Appl. Environ. Microbiol. **70:**2193–2203.
16. **Mereghetti, L., P. Lanotte, V. Savoye-Marczuk, N. Marquet-Van Der Mee, A. Audurier, and R. Quentin.** 2002. Combined ribotyping and random multiprimer DNA analysis to probe the population structure of *Listeria monocytogenes*. Appl. Environ. Microbiol. **68:**2849–2857.
17. **Nadon, C. A., D. L. Woodward, C. Young, F. G. Rodgers, and M. Wiedmann.** 2001. Correlations between molecular subtyping and serotyping of *Listeria monocytogenes*. J. Clin. Microbiol. **39:**2704–2707.
18. **Nightingale, K. K., K. Windham, and M. Wiedmann.** 2005. Evolution and molecular phylogeny of *Listeria monocytogenes* from human and animal cases and food. J. Bacteriol. **187:**5537–5551.
19. **Nightingale, K. K., K. Windham, K. E. Martin, M. Yeung, and M. Wiedmann.** 2005. Selected *Listeria monocytogenes* subtypes commonly found in food show reduced invasion in human intestinal cells due to distinct nonsense mutations in *inlA* leading to expression of truncated and secreted internalin A. Appl. Environ. Microbiol. **12:**8764–8772.
20. **Norton, D. M., J. M. Scarlett, K. Horton, D. Sue, J. Thimothe, K. J. Boor, and M. Wiedmann.** 2001. Characterization and pathogenic potential of *Listeria monocytogenes* isolates from the smoked fish industry. Appl. Environ. Microbiol. **67:**646–653.
21. **Piffaretti, J. C., H. Kressebuch, M. Aeschbacher, J. Bille, E Bannerman J. M. Musser, R. K. Selander, and J. Rocourt.** 1989. Genetic characterization of clones of the bacterium *Listeria monocytogenes* causing epidemic disease. Proc. Natl. Acad. Sci. USA **10:**3818–3822.
22. **Posada, D., and K. A. Crandall.** 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics **14:**817–818.
23. **Rousseaux, S., M. Olier, J. P. Lamaitre, P. Piveteau, and J. Guzzo.** 2004. Use of PCR-restriction fragment polymorphism of *inlA* for rapid screening of *Listeria monocytogenes* strains deficient in the ability to invade Caco-2 cells. Appl. Environ. Microbiol. **70:**2180–2185.
24. **Sauders, B. D., K. Mangione, C. Vincent, J. Schermerhorn, C. M. Farchione, N. B. Duman, D. Bopp, L. Kornstein, E. D. Fortes, K. Windham, and M. Wiedmann.** 2004. Distribution of *Listeria monocytogenes* molecular subtypes among human and food isolates from New York State shows persistence of human-disease associated *Listeria monocytogenes* stubypes in retail environments. J. Food Prot. **67:**1417–1428.
25. **Schlech, W. F.** 2000. Foodborne listeriosis. Clin. Infect. Dis. **31:**770–775.
26. **Smith, G. A., J. A. Theriot, and D. A. Portnoy.** 1996. The tandem repeat domain in the *Listeria monocytogenes* ActA protein controls the rate of actin-based mobility, the percentage of moving bacteria, and the localization of vasodilator-stimulated phosphoprotein and profilin. J. Cell Biol. **135:**647–660.
27. **Sun, A. N., A Camilli, and D. A. Portnoy.** 1990. Isolation of *Listeria monocytogenes* small-plaque mutants defective for intracellular growth and cell-to-cell spread. Infect. Immun. **58:**3770–3778.
28. **Swofford, D.** 1997. Phylogenetic analysis using parsimony (*and other methods). Sinauer Associates, Sunderland, Mass.
29. **Tavanti, A., A. D. Davisdon, M. J. Fordyce, N. A. R. Gow, M. C. J. Maiden, and F. C. Odds.** 2005. Population structure and properties of *Candida albicans*, as determined by multilocus sequence typing. J. Clin. Microbiol. **43:**5601–5613.
30. **Urwin, R., and M. C. J. Maiden.** 2003. Multi-locus sequence typing: a tool for global epidemiology. Trends Microbiol. **11:**479–487.
31. **Ward, T. J., L. Gorski, M. K. Borucki, R. E. Mandrell, J. Hutchins, and K. Pupedis.** 2004. Intraspecific phylogeny and lineage group identification based on the *prfA* virulence gene cluster of *Listeria monocytogenes*. J. Bacteriol. **15:**4994–5002.
32. **Wesley, I. V.** 1999. Listeriosis in animals, p. 39–73. *In* E. T. Ryser and E. H. Marth (ed.), *Listeria* listeriosis and food safety, 2nd ed. rev. and expanded. Marcel Decker, Inc., New York, N.Y.
33. **Wiedmann, M., J. L. Bruce, C. Keating, A. E. Johnson, P. L. McDonough, and C. A. Batt.** 1997. Ribotypes and virulence gene polymorphisms suggest three distinct *Listeria monocytogenes* lineages with differences in pathogenic potential. Infect. Immun. **65:**2707–2716.