Recombination and positive selection contribute to evolution of *Listeria monocytogenes* *inlA*

R. H. Orsi,1  
D. R. Ripoll,2  
M. Yeung,3  
K. K. Nightingale4  
and M. Wiedmann1

1Department of Food Science, Cornell University, Ithaca, NY, USA  
2Computational Biology Service Unit, Cornell Theory Center, Cornell University, Ithaca, NY, USA  
3Department of Microbiology and Immunology, Cornell University, Ithaca, NY, USA

The surface molecule InlA interacts with E-cadherin to promote invasion of *Listeria monocytogenes* into selected host cells. DNA sequencing of *inlA* for 40 *L. monocytogenes* isolates revealed 107 synonymous and 45 nonsynonymous substitutions. A frameshift mutation in a homopolymeric tract encoding part of the InlA signal peptide was identified in three lineage II isolates, which also showed reduced ability to invade human intestinal epithelial cells. Phylogenies showed clear separation of *inlA* sequences into lineages I and II. Thirteen *inlA* recombination events, predominantly involving lineage II strains as recipients (12 events), were detected and a number of amino acid residues were shown to be under positive selection. Four of the 45 nonsynonymous changes were found to be under positive selection with posterior probabilities > 95%. Mapping of polymorphic and positively selected amino acid sites on the partial crystal structure for InlA showed that the internalin surface of the leucine-rich repeat (LRR) region that faces the InlA receptor E-cadherin does not include any polymorphic sites; all polymorphic and positively selected amino acids mapped to the outer face of the LRR region or to other InlA regions. The data show that (i) *inlA* is highly polymorphic and evolution of *inlA* involved a considerable number of recombination events in lineage II isolates; (ii) positive selection at specific amino acid sites appears to contribute to evolution of *inlA*, including fixation of recombinant events; and (iii) single-nucleotide deletions in a lineage II-specific 39 homopolymeric tract in *inlA* lead to complete loss of InlA or to production of truncated InlA, which conveys reduced invasiveness. In conclusion, *inlA* has a complex evolutionary history, which is consistent with *L. monocytogenes’* natural history as an environmental pathogen with broad host-range, including its adaptation to environments and hosts where different *inlA* alleles may provide a selective advantage or where *inlA* may not be required.

**INTRODUCTION**

*Listeria monocytogenes* is a facultative intracellular foodborne pathogen, which can cause septicaemia, encephalitis, meningoencephalitis and abortion in different mammalian hosts, most commonly in humans and ruminants (Vazquez-Boland et al., 2001). *L. monocytogenes* can be divided into at least three distinct genetic lineages as determined by various molecular subtyping methods, including multilocus sequence typing (MLST) (Ward et al.,
2004; Wiedmann et al., 1997; Nightingale et al., 2005a). Lineage I isolates have been responsible for the majority of listeriosis outbreaks and are more commonly isolated from human clinical cases than foods (Gray et al., 2004; Jeffers et al., 2001); lineage I thus has been hypothesized to represent a human-host-adapted lineage (Nightingale et al., 2005a). While lineage II strains have been isolated from sporadic human clinical cases, they are rarely responsible for human listeriosis outbreaks (Jeffers et al., 2001). Furthermore, lineage II strains are significantly more common among food isolates than among human clinical isolates (Gray et al., 2004), suggesting that these isolates may be better adapted to a non-host environment. Lineage III strains are predominantly isolated from animal clinical cases and are rarely isolated from environmental and food samples or human clinical cases (Jeffers et al., 2001; Roberts et al., 2006). *L. monocytogenes* lineages thus may have adapted to different host-and non-host-associated ecological niches.

Bacterial surface molecules involved in attachment to and invasion of host cells may play important roles in host specificity (e.g. Wilson et al., 2000). In *L. monocytogenes*, a group of cell surface proteins termed internalins, which are characterized by the presence of variable numbers of leucine-rich repeats (LRRs), appear to play important roles in attachment to host cells. While more than 25 internalin genes have been identified in *L. monocytogenes*, internalin A (InlA), encoded by *inlA*, has been shown to be particularly important for host-cell invasion (Lecuit et al., 2001). InlA is typically anchored to the bacterial cell wall via an LPXTG anchor (Dhar et al., 2000) and has been shown to bind to the host protein E-cadherin, which is an outer-membrane protein expressed in epithelial cells of several organisms, including humans (Mengaud et al., 1996). While InlA has been shown to be critical for invasion of intestinal epithelial cells and thus for crossing of the intestinal epithelial barrier, internalin A–E-cadherin interactions appear to be host-specific; InlA binds to human and guinea pig E-cadherin; however, a single amino acid difference in the murine E-cadherin (presence of glutamate in murine E-cadherin at aa 16 compared to a proline in humans and guinea pig E-cadherin) is sufficient to prevent InlA binding (Lecuit et al., 1999). In addition to its role in intestinal invasion, InlA may also contribute to crossing of the maternofetal barrier in pregnant hosts (Lecuit et al., 2004), even though these contributions were not apparent in a pregnant guinea pig model (Bakardjiev et al., 2004). InlA is 800 aa long and can be divided into eight distinct regions. Starting at the N-terminal end, internalin A has a signal peptide followed by an a-helix motif and fifteen and a half 22-residue LRRs, forming a right-handed solenoid with each repeat adding a helical turn (Schubert et al., 2002). C-terminal of the LRRs, the inter-repeat (IR) or immunoglobulin (Ig)-like region is the most flexible part of the internalin domain (Schubert et al., 2002). This region is followed by the B repeats, which consist of two repeats of 70 aa (B repeats 1 and 2) and a third repeat of 49 aa (B repeat 3) (Dramsi et al., 1993; Gaillard et al., 1991). The most C-terminal functionally important domain of InlA is the transmembrane region, which includes an LPXTG anchoring motif and a short positively charged tail. Additionally, a small region with no described function is located between the IR region and the B repeats and a proline/glycine-rich segment (PGS), also with no known function, is located between the B repeats and the LPXTG motif (Dramsi et al., 1993; Gaillard et al., 1991).

In this work, we analysed the evolutionary history of *inlA* using 40 *L. monocytogenes* isolates obtained from diverse sources, including human and bovine clinical cases, foods, and natural environments. An approach using a combination of evolutionary analyses and
phenotypic characterization using tissue culture invasion assays showed that *inlA* has a complex evolutionary history involving positive selection, recombination among lineage II strains and single-nucleotide deletions that lead to reduced invasiveness in selected lineage II strains.

**METHODS**

**Bacterial isolates and lysate preparation.** A collection of 132 *L. monocytogenes* isolates collected in New York State between 2000 and 2002, including 120 isolates from foods, human clinical cases and bovine clinical cases (previously described by Nightingale et al., 2005a) and 12 isolates from natural environments (Sauders et al., 2006) was stratified by source (human clinical cases, bovine clinical cases, food or natural environment) and ten isolates were randomly selected for each source. All isolates had previously been characterized by EcoRI ribotyping, and ribotyping data had also been used to classify isolates to lineage (Nightingale et al., 2005a; Sauders et al., 2006). The 40 isolates selected represented the three major *L. monocytogenes* genetic lineages: lineage I (18 isolates), lineage II (21 isolates) and lineage III (one isolate) (Supplementary Table S1, available with the online version of this paper, shows detailed information for all 40 isolates).

**DNA sequencing.** Four pairs of primers were used to amplify four fragments covering the whole *inlA* ORF (Supplementary Table S2, available with the online version of this paper, shows all primer sequences). PCR reactions were performed using 16 PCR buffer, 1.5 μM MgCl2, 50 μM dNTPs, 0.5 mM of each primer and 0.05 U Taq DNA polymerase ml⁻¹ (Perkin Elmer-Applied Biosystems) as well as 1–2 ml bacterial lysate. PCR products were purified using the QIAquick PCR purification kit (Qiagen) and DNA sequencing was performed at Macrogen using the same primers used for PCR. *inlA* sequences were assembled using Seqman (DNASTAR, Lasergene).

**Descriptive sequence analysis.** *inlA* contigs for the 40 *L. monocytogenes* isolates were aligned with CLUSTAL W (MEGALIGN, DNASTAR, Lasergene) and all variable sites were proofread for errors. Descriptive analyses, including G+C content, percentage of polymorphisms and number of synonymous and nonsynonymous substitutions, were carried out using DnaSP (Rozas & Rozas, 1999).

**Phylogenetic analysis.** MODELTEST (Posada & Crandall, 1998) was used to find the most likely model of DNA substitution. One sequence representing each of the 23 *inlA* allelic types was used to construct rooted and unrooted maximum-likelihood phylogenetic trees in PAUP (http://paup.csit.fsu.edu/) using the nucleotide substitution model HKY85+G+I, which was chosen by MODELTEST. Bootstrap analysis was performed with 100 repetitions using PAUP. The rooted tree that was constructed assuming a molecular clock and the unrooted tree that did not assume a molecular clock were used to test the molecular clock assumption using a likelihood ratio test (LRT) with *n*-2 degrees of freedom (with *n* being the number of alleles).

**Recombination analysis.** The program GENECONV (http://www.math.wustl.edu/~sawyer) was used to identify fragments that were likely to have
originated by recombination. Recombination analyses were performed using an alignment that contained one sequence representing each inlA allelic type. The GENECONV analysis was repeated with four different 'gscale' values, including G0 (the default), G1, G2 and G3. Mismatch penalties increase from G1 to G3 and are highest for G0. GENECONV outputs were further analysed and fragments with the same breakpoints were considered to represent the same recombination event, as previously described (Nightingale et al., 2005a).

**Positive selection analysis.** The program codeml in the software package PAML version 3.14 (Yang, 1997) was used to test whether inlA is likely to include codons that are under positive selection and to identify specific amino acid sites that showed statistically significant evidence for positive selection (Yang et al., 2005). The naive empirical Bayes (NEB) inference approach was used for M3 while the new Bayes empirical Bayes (BEB) inference approach was used for M2a and M8. Analyses of positive selection by lineage were carried out as previously described (Zhang et al., 2005).

**3D modelling.** The available structure for the N-terminal, the LRR and IR domains of InlA bound to E-cadherin (pdb code: 1O6S; Schubert et al., 2002) was retrieved from the Protein Data Bank (Berman et al., 2000) and used for analysis and modelling purposes. The structure was used to map nonsynonymous mutations and in particular residues under positive selection.

**Invasion assay.** Selected isolates, including two isolates with an inlA frameshift mutation (Table 3), were tested for their ability to invade the human intestinal epithelial Caco-2 cell line. Caco-2 invasion assays were performed in 24-well plates essentially as previously described (Nightingale et al., 2005b). Briefly, 1 ml of L. monocytogenes overnight culture (grown at 30 °C without shaking) was pelleted and resuspended in 1 ml PBS. Confluent Caco-2 monolayers were inoculated with approximately 2 x 10⁷ L. monocytogenes; the inoculum was enumerated on BHI agar plates. Inoculated Caco-2 monolayers were incubated for 30min at 37 °C, followed by three washes with PBS and addition of fresh media without antibiotics. Medium containing 150 μg gentamicin ml⁻¹ was added 45 min post-inoculation to kill extracellular bacteria. At 90 min post-inoculation, Caco-2 monolayers were washed three times with PBS and Caco-2 cells were subsequently lysed by addition of 500 μl cold sterile distilled water and vigorous pipetting. Intracellular L. monocytogenes were enumerated by spread plating on BHI agar plates. Each isolate was tested in three different wells on each of three different days (representing three independent replicates). Invasion efficiency was reported as the percentage of inoculum recovered by enumeration of intracellular bacteria. A standard laboratory control strain (10403S) and an isogenic ΔinlA strain were included as controls in each invasion assay (Table 2). Data from the three replicates were used for statistical analysis using the Wilcoxon rank sum test (as data were not normally distributed).

**Western blot analysis.** Western immunoblot analysis was performed as previously described (Nightingale et al., 2005b) to probe for the presence of InlA in bacterial cell wall and supernatant fractions from isolate FSL F2-723, which has an inlA frame shift mutation, and control strain 10403S, which encodes a full-length InlA. Briefly, bacteria were grown in
LB supplemented with 50 mM MOPS adjusted to pH 7.3, 25 mM glucose 1-phosphate and 0.2 % activated charcoal. Chloramphenicol was added to a final concentration of 10 mg ml\(^{-1}\) 10 min prior to cell harvest. Cell wall and supernatant fractions were prepared as described previously (Snyder & Marquis, 2003). Equivalent amounts of culture optical density units were loaded per lane. InlA was detected by Western immunoblotting using a mouse anti-InlA antibody.

**Isolate information and nucleotide sequence accession numbers.** Isolate information and subtyping data from this study are archived and freely available through the Pathogen Tracker 2.0 database (http://www.pathogentracker.net). *inlA* sequences have also been deposited in GenBank (accession nos EF445899–EF445938).

**RESULTS**

**Descriptive analysis**

Sequencing of the full 2400 bp *inlA* ORF in 40 *L. monocytogenes* isolates yielded 23 different *inlA* allelic types (Supplementary Table S1). Three isolates, including two lineage I, ribotype DUP-1044A isolates (both obtained from human clinical cases) and one lineage II, ribotype DUP-1039C isolate (obtained from a bovine clinical case), presented an in-frame deletion of nine nucleotides (nt 2220–2228), resulting in a sequence of 2391 bp. One food isolate (FSL F2-515; lineage II, ribotype DUP-1062A) had a nonsense mutation at nt 2100 that creates a stop codon after aa 699. Another three food isolates (all classified as lineage II, ribotype DUP-1039C) carried a single nucleotide deletion in a homopolymeric tract of seven adenine residues (located at nt 6–12), resulting in a premature stop codon after aa 8 (Fig. 1); all three of these isolates had an identical *inlA* sequence. With the exception of the frameshift mutation, the *inlA* sequences for these three isolates were identical to the *inlA* sequences of other three ribotype DUP-1039C isolates, which had been collected from natural environments. Among the 21 lineage II isolates, 14 had a homopolymeric tract with seven adenines, while four had six adenines with a guanine after the second adenine. All lineage I isolates and the lineage III isolate had a guanine after the second adenine in that region (Fig. 1).

The average G+C content for *inlA* was 37.3 mol%. A total of 6.2% of the *inlA* sites were polymorphic; these polymorphic sites included 107 sites with synonymous substitutions and 44 sites with nonsynonymous substitutions (representing a total of 45 nonsynonymous changes). The average number of nucleotide differences per site between two sequences (\(p\)) was 0.02134; if only nonsynonymous sites were considered, \(p\) was 0.007. Lineage II isolates showed a larger *inlA* nucleotide diversity (\(\pi_{II}=0.01322\)) than lineage I isolates (\(\pi_{I}=0.00526\)).

**Phylogenetic analysis**

MODELTEST identified the evolutionary model HKY+G+I as the most likely model explaining the nucleotide substitution patterns found in the 40 *inlA* sequences analysed. The molecular clock hypothesis was rejected (\(P<0.001\)), indicating that all branches in the tree did not evolve at the same rate. The *inlA* phylogenetic tree (Fig. 2) constructed using the HKY+G+I model without the molecular clock assumption showed that the *inlA* sequences represented two distinct clusters, which were consistent with the lineage
groupings of isolates based on EcoRI ribotyping; the two main clusters represented lineages I and II (Fig. 2). When phylogenetic trees were generated containing only isolates from lineage I or lineage II, the molecular clock assumption could be rejected for lineage II isolates ($P<0.001$), but not for lineage I isolates ($P>0.05$). These results suggest that different evolutionary forces have driven the evolution of inlA of lineage I and lineage II isolates and that lineage II inlA is probably exposed to evolutionary forces other than point mutations (e.g. recombination, positive selection).

Recombination analysis

GENECONV was used to infer possible fragments that do not follow a history of vertical evolution. When the analysis was performed with the GENECONV default setting (G0), which does not allow mismatches within the inferred recombinant fragment, many recombinant fragments had similar breakpoints and involved the same isolates. Visual examination suggested that these fragments represented the same recombination event and that a single recombination event was identified by GENECONV as multiple fragments. Hence, G0 results were not further analysed and GENECONV analysis was repeated using three modifications of the default setting (G1, G2 and G3), which allow for different levels of mismatches within the fragment; the mismatch penalty increases from G1 to G3 and is lowest for G1; mismatch penalties for G2 and G3 are about two and three times higher, respectively, than for G1. G1, G2 and G3 identified a total of 8, 58 and 56 global inner fragments, respectively; multiple fragments often showed the same 59 or 39 breakpoint, likely indicating that multiple fragments represented a single recombination event (as also described by Nightingale et al., 2005a). Based on visual analyses, all recombinant fragments detected by GENECONV could be classified into seven different recombination events (events 1–7; Table 1). Visual evaluation of an alignment containing only the polymorphic sites identified lineage II isolates as recipients of the recombinant fragment for six events; in one event, the lineage III isolate was identified as recipient. Donors in these seven events included lineage I and II isolates (three events each), as well as the lineage III isolate (one event). Events 5–7 (Table 1; Fig. 3) involved related isolates and appear to represent an ancestral recombination event (event 5, Fig. 3) followed by two other recombination events (events 6 and 7, Fig. 3).

Analysis of the inlA phylogenetic tree (Fig. 2) revealed that seven lineage II inlA sequences (isolates F2-590, F2-634, F2-639, E1-123 and S4-304) and two lineage I sequences (F2-637 and F2-672) that were not identified as recipients in the GENECONV analysis were highly divergent and have long-terminal branches, indicating an accumulation of polymorphisms in these branches. Hence, polymorphic sites alignments for these isolates as well as for all lineage II isolates (due to the large number of long-terminal branches in this lineage) were visually analysed for evidence of additional recombination events. These analyses identified six additional putative recombination events (events 8–13, Table 1), which all appear to have lineage II isolates as recipients and lineage I isolates as donors. To further identify and confirm recombination events, a fourth approach (termed ‘no singletons’, NS; Table 1) was used; in this approach, all singletons (non-informative sites) in the inlA alignment were removed and GENECONV analysis was performed using the default settings (G0). This procedure is justified because singletons may represent mutations that occurred after the recombination event in either the donor or recipient sequence; this approach differs from G1, G2 and G3 as it specifically removes
non-informative sites. This 'NS' procedure identified seven recombination events, including five events previously identified by G1, G2 and G3 (events 3–7, Table 1) and two events previously identified by the visual analysis (events 8 and 9, Table 1, Fig. 4). In summary, we identified a total of 13 inlA recombination events, 12 of them involving lineage II isolates as recipients. Fourteen out of the 21 lineage II isolates showed evidence for at least one horizontal gene transfer event, including 4/4 animal, 6/9 environmental, 2/3 human and 1/5 food isolates classified in lineage II. While we detected 13 recombination events, it is feasible that some recombination events (e.g. between closely related sequences in lineage I) may not have been detected, and that additional recombination events were involved in the evolution of the inlA sequences studied here.

**Positive selection in inlA**

Nested models of heterogeneous codon substitution (Yang et al., 2000, 2005) were used to determine whether inlA has evolved by positive selection. In these models, \( \omega \) represents the \( d_N/d_S \) ratio [i.e. (no. of nonsynonymous changes/no. of nonsynonymous sites) / (no. of synonymous changes/no. of synonymous sites)]; \( \omega > 1 \) is indicative of positive selection, \( \omega = 1 \) is indicative of purifying selection. The comparison between M0 and M3 may be interpreted as a test for variability in the \( \omega \) ratio among sites (Yang & Nielsen, 2002). M0 was rejected in favour of M3 (\( P < 0.001 \)), indicating that the assumption of a single \( \omega \) for all sites along a sequence is not valid and that \( \omega \) varies along the length of the inlA sequence.

Four models were used to determine whether positive selection is likely to have occurred in inlA. Model 1a (M1a) is a neutral model, which assumes two classes of \( \omega \), including (i) \( \omega_0 \) (0 < \( \omega_0 < 1 \); representing codons under negative selection) and (ii) \( \omega_1 \) (\( \omega_1 = 1 \); representing codons under neutral selection). Model 2a (M2a) allows for a third class of \( \omega_1 \) (\( \omega_2 > 1 \); representing codons under positive selection). Model 7 (M7) uses a discrete approximation of a beta distribution and does not allow a site class where \( \omega > 1 \); model 8 (M8) allows for an additional site class where \( \omega > 1 \). M1a was rejected in favour of M2a (\( P < 0.005 \)) and M7 was rejected in favour of M8 (\( P < 0.005 \)), indicating that positive selection has likely played a role in the evolution of inlA. M2a identified two sites with posterior probabilities of being under positive selection greater than 95%, while M8 identified four sites, which included the two sites found by M2a (Table 2).

Model A as described by Zhang et al. (2005) was used to test the hypothesis that positive selection has occurred in specific lineages (lineage I or lineage II). Test 2 compares model A, which allows for positive selection along one specific branch or set of branches, and model A null, which does not allow for positive selection along the same branch or set of branches (Zhang et al., 2005). Each lineage was tested independently (see Fig. 2; branches tested by these models are indicated by Greek letters). First, the branch that originated each lineage was tested using the models and then all the branches that belong to each lineage were also tested. We found no evidence for positive selection in lineage I. Conversely, when all branches derived from branch b were tested (meaning all lineage II branches), model A null could be rejected (\( P < 0.001 \)), indicating positive selection restricted to lineage II isolates. Model A identified a single site with BEB posterior probability 0.95 as being under positive selection in lineage II (site 594; \( P = 0.989 \)); this site was also identified as being under positive selection by M2a and M8.
As we were aware that the $d_N/d_S$ tests implemented in PAML assume no recombination and knowledge of the true phylogeny (Anisimova et al., 2003; Nielsen, 2001), we repeated the analysis using an alignment without the inlA sequences that were identified to be recipients of fragments by horizontal gene transfer. In this analysis, we could not reject any of the null models (M0, M1a, M7) in favour of the alternative models (M3, M2a and M8) and no sites were identified to be under positive selection. However, 16 of the 45 nonsynonymous substitutions were found only in the 14 isolates that showed evidence for recombination in inlA. By comparison, eight nonsynonymous substitutions were only found in the 26 isolates without evidence for recombination in inlA; 21 nonsynonymous substitutions were shared between these two groups of isolates. While our analyses thus may indicate that recombination events may have affected our positive selection analyses, the observation that no evidence of positive selection was found when analyses were performed only on 26 inlA sequences without any apparent history of recombination may also reflect reduced power of these analyses as this dataset contains a considerably smaller number of nonsynonymous substitutions as compared to all 40 inlA sequences.

Positive selection by functional InlA regions

InlA can be divided into eight regions, which have distinct characteristics and functions. The LRR was found to be the most conserved region (only 1.4 % of amino acid sites are polymorphic), while PGS is the most variable region, with 10 % of amino acid sites being variable (Fig. 5). To determine how selection acts on these eight regions, the $d_N/d_S$ overall ratio was calculated for each region. Nested models of evolution, including (i) M0 and M3, (ii) M1a and M2a and (iii) M7 and M8 were also analysed as described in detail above. The signal peptide (SP) and interdomain (ID) regions were not analysed because the first is not present in the mature InlA and the second is too short (23 aa) to provide reliable estimates of parameters. For two regions (i.e. the LRR and B-repeat region), M0 was rejected and M3 was accepted. M1a could not be rejected for any region ($P>0.05$), even though for the LRR region, M8 was accepted over M7, providing evidence for positive selection in the LRR region. One site (codon 187) was identified by M8 as being under positive selection in the LRR region (Fig. 5); this site had also been identified as being under positive selection by M2a and M8 using the complete inlA sequence (Table 2). Consistent with the fact that M3 was accepted when the whole sequence was analysed, these results suggest that the different regions of InlA are under different selective pressures.

Mapping of the polymorphic and positively selected sites using the available crystallized structure of InlA

The partial crystallized structure of InlA, which includes the LRR region as well as the α-helical and the Ig-like domains (Schubert et al., 2002), was used to map the localization of the polymorphic amino acid sites as well as positively selected sites identified by M8 (see above). Among the 44 polymorphic amino acids sites (including the four positively selected sites, which were identified by M8), 14 polymorphic amino acids sites (including two sites under positive selection) were located in the InlA regions that are included in the partial crystallized structure (Fig. 6). All positively selected amino acid were found to be located on the outer surface of the LRR domain and no polymorphic amino
acids were identified in the inner surface of the LRR domain, including the specific region that interacts with E-cadherin.

**Invasiveness of isolates with an inlA frameshift mutation**

Invasion assays on two representative ribotype DUP-1039C isolates with a premature stop codon after InlA aa 8 were carried out to investigate the invasiveness of these isolates. The two isolates bearing the frameshift mutation were significantly (P<0.05; invasion efficiencies of 0.010 and 0.013 % for isolates FSL F2-723 and FSL F2-640, respectively) less invasive than strain 10403S (a laboratory control strain that encodes a full-length InlA; invasion efficiency of 0.139 %) and isolate S4-497 (a ribotype DUP1039C isolate that encodes a full-length InlA; invasion efficiency of 1.48 %). The two isolates with the inlA frameshift mutation did not differ significantly (P<0.05) in their invasiveness from strain FSL K4-006, a DinlA mutant derived from strain 10403S (invasion efficiency of 0.004 %).

**Western blot analysis of a selected isolate with an inlA frameshift mutation**

Western blot analysis of the cytoplasmic protein fraction and culture supernatant of an isolate with an inlA frameshift mutation (i.e. isolate FSL F2-723) was performed to confirm the premature stop codon after InlA aa 8 identified in three isolates based on inlA nucleotide sequence data. While no InlA was detected in the cell wall (not shown) and supernatant fraction of isolate FSL F2-723 (Fig. 7), a weak band with the expected size for InlA was detected in the cytoplasmic extract of isolate FSL F2-723 (Fig. 7). We hypothesized that this weak band is due to weak translation of InlA from a second potential start codon [located at nt 169–171, which encodes aa 57 (methionine) in the full-length InlA]. This hypothesis would also explain why InlA was not identified in the cell wall or supernatant, since this potential start codon is just downstream of the signal peptide region, which would not be present in a truncated protein translated from this start codon.

**DISCUSSION**

Overall, our data show that (i) inlA is highly polymorphic and evolution of inlA involved a considerable number of recombination events in lineage II isolates; (ii) positive selection at specific amino acid sites appears to contribute to evolution of inlA, including fixation of recombinant events; and (iii) a single nucleotide deletion in a lineage II-specific 59 homopolymeric tract in inlA leads to complete loss of InlA or to production of truncated InlA, which is associated with reduced invasiveness. The identification of both polymorphic and positively selected sites in inlA as reported here also provides an opportunity for future studies that include generation and characterization of isogenic strains carrying different inlA allelic variants to explore the phenotypic consequences of these mutations.

**inlA is highly polymorphic and evolution of inlA involved a considerable number of recombination events in lineage II isolates**

Overall nucleotide diversity (π=0.02134) as well as the number of nonsynonymous sites were high for inlA, but were within the range previously described for six other internalin genes (inlB, inlC2, inlD, inlE, inlF and inlG) that had previously been sequenced in the same 40 isolates characterized here (Tsai et al., 2006). Thus, this group of *L. monocytogenes* surface proteins is characterized by considerable overall nucleotide and
amino acid diversity. Similar to the observation that all of the six other *L. monocytogenes* internalins previously characterized by Tsai et al. (2006) showed evidence for recombination, we have identified a number of recombination events among the 23 *inlA* allelic types identified here. Interestingly, all but one of the *inlA* recombination events identified involved lineage II strains as recipients. This is consistent with previous observations that lineage II strains show evidence for considerably more horizontal gene transfer than lineage I isolates (Meinersmann et al., 2004; Nightingale et al., 2005a) and that most recombination events in other internalin genes involved lineage II strains as recipients (Tsai et al., 2006). The important role for recombination in generating diversity in selected *L. monocytogenes* genes is further supported by Nightingale et al. (2005a), who showed that *L. monocytogenes* genes with weak (i.e. *prs*) or no evidence for recombination (i.e. *gap* and *sigB*) showed lower diversity than housekeeping genes with strong evidence for recombination (i.e. *purM, ribC*). Our data are also consistent with a number of studies that have shown that recombination plays an important role in the diversification of genes encoding surface proteins in many pathogens, including the *Escherichia coli* intimin gene (McGraw et al., 1999) and *fimA* (Peek et al., 2001), the *Plasmodium falciparum* AMA1 gene (Polley & Conway, 2001) as well as *Neisseria meningitidis* *pilE* (Andrews & Gojobori, 2004) and *porB* (Urwin et al., 2002). Recombination within genes encoding surface proteins, including *inlA*, may specifically be important by providing rapid diversification that allows for evasion of the host immunity system and for adaptive shifts in host or tissue tropism (Andrews & Gojobori, 2004; McGraw et al., 1999; Peek et al., 2001; Polley & Conway, 2001; Urwin et al., 2002).

**Positive selection at specific amino acid sites appears to contribute to evolution of *inlA*, including fixation of recombinant events**

Our analysis indicated that specific amino acid sites in *inlA* are under positive selection and that positive selection appears to specifically act on lineage II *inlA*. While we are aware that our dataset violated one of the assumptions of the positive selection analysis used (i.e. absence of recombination) (Anisimova et al., 2003), there are currently no methods available that can be used to test for evidence of positive selection in genes that have a history of recombination. While recombination, particularly in early internal branches, can affect the reliability of the LRT (Anisimova et al., 2003), Bayes’ identification of positively selected sites seems to be less sensitive to the tree topology (Anisimova et al., 2003), and therefore can be used even when recombination is present. As Bayes’ identification of positively selected sites found a number of positively selected sites, we are confident that our results correctly suggest that *inlA* has undergone positive selection. We specifically hypothesize that positive selection may contribute to fixation of recombinant *inlA* genes that encode for proteins that provide for increased fitness in the *L. monocytogenes* population sampled. This is consistent with the hypothesis that, in order to maintain recombinant fragments in a population, such fragments should be advantageous to the organism (Andrews & Gojobori, 2004; Milkman et al., 2003). Recombination has previously been suggested to act together with positive selection in several genes in different organisms (Andrews & Gojobori, 2004; Peek et al., 2001; Polley & Conway, 2001; Urwin et al., 2002), including many pathogens (Andrews & Gojobori, 2004; Milkman et al., 2003; Polley & Conway, 2001). For example, strong positive selection and recombination seem to drive the antigenic variation of the *PilE* protein of the human pathogen *N.*
meningitidis (Andrews & Gojobori, 2004). In E. coli, recombination in fimA has been proposed to allow for diversifying and purifying selection to occur in different regions of the same gene by uncoupling these regions (Peek et al., 2001). Similar mechanisms may be at work in inlA, where one area, which contacts the human receptor E-cadherin, seems to be under strong purifying selection while other regions may be under diversifying selection, possibly in response to selective pressure imposed by the host immune response, which may select for changes in epitopes found on the outer surface of pathogen proteins (Andrews & Gojobori, 2004; Peek et al., 2001; Polley & Conway, 2001). While this is consistent with the observation that both of the positively selected sites that could be mapped to the InlA crystal structure were totally or partially exposed on the surface of InlA, further studies are necessary to determine whether these sites are indeed recognized by the host immune system.

A single nucleotide deletion in a lineage II-specific 5$^9$ homopolymeric tract in inlA leads to complete loss of InlA or to production of truncated InlA, which is associated with reduced invasiveness

Among the 40 isolates characterized here, we identified four food isolates that carried mutations in inlA that led either to a premature stop codon after aa 699 (one isolate) or to a premature stop codon after aa 8 (three isolates). The inlA allelic type encoding a premature stop codon after aa 699 has previously been reported by Nightingale et al. (2005b) and leads to expression of a truncated and secreted form of InlA that lacks the cell-wall-binding LPXTG motif; this premature stop codon appears to be common among food isolates in the USA (Nightingale et al., 2005b). However, the frameshift mutation that leads to an early premature stop codon (after aa 8) has, to our knowledge, not previously been described. Interestingly, the mutations leading to this premature stop codon represents a single nucleotide deletion in a homopolymeric tract of seven adenines, which is found in 14/21 lineage II inlA sequences. Frameshift mutations due to single nucleotide deletions on homopolymeric tracts have previously been reported to occur naturally in a number of organisms (e.g. Kearns et al., 2004; Segura et al., 2004; Theiss & Wise, 1997) and these types of deletion events appear to be reversible, allowing organisms to undergo phase variation. For example, phase variation involving a frameshift mutation in a run of A:T base pairs has been reported in a swarming gene in Bacillus subtilis (Kearns et al., 2004). Future experiments will thus focus on evaluating the reversibility of the inlA frameshift mutation identified here. As we have identified identical inlA sequences with and without the single adenine deletion that leads to a premature stop codon, we hypothesize that this mutation may be reversible, possibly allowing for InlA phase variation. Interestingly, while most lineage II strains, which are overrepresented among food and environmental isolates and underrepresented among human isolates (Sauders et al., 2006), have a homopolymeric tract of seven adenines, this homopolymeric tract was found here to be interrupted by a guanine in strains that group into lineage I, a lineage which appears to be overrepresented among human isolates (Sauders et al., 2006). We hypothesize that potentially environmentally adapted L. monocytogenes (i.e. many lineage II strains) may have evolved to allow for InlA phase variation, while potentially human-adapted strains may have evolved to stably express InlA. Interestingly, our recombination analyses showed that at least one of the four lineage II isolates with a homopolymeric tract of seven adenines at the 59 end of inlA that is interrupted by a guanine acquired this inlA
fragment by horizontal gene transfer (event 9) from a lineage I strain, indicating that some lineage II strains may evolve towards stable expression of full-length InlA.

Our data also provide further evidence that mutations leading to premature stop codons in inlA occur commonly and at different locations in the genes, leading to either expression of a secreted InlA or to expression of truncated proteins, which are unlikely to be functional. Specifically, characterization of L. monocytogenes isolates from France (Jonquieres et al., 1998; Olier et al., 2002, 2003; Rousseaux et al., 2004) and the USA (this study, Nightingale et al., 2005b) so far have identified at least 10 distinct mutations leading to premature stop codons in inlA. A number of the mutations that occur in different sites in the 59 end of inlA have previously been shown to be associated with a decreased ability of L. monocytogenes to invade epithelial cells (Olier et al., 2003; Nightingale et al., 2005b). It thus seems that there is a strong selective force driving the loss of a functional InlA in some L. monocytogenes isolates, particularly those found in foods, as supported by the fact that a considerable number of independent mutations leading to a loss of functional, cell-wall-bound InlA have been found and appear to have been fixed in the populations studied. Thus, lack of a cell-wall-bound InlA or complete abolition of InlA expression may provide a selective advantage for certain L. monocytogenes subtypes in certain environments. The specific selection pressure responsible for this remains to be determined though, which is likely to prove challenging considering that L. monocytogenes is found in a diversity of environments and can infect a number of mammalian and non-mammalian host species (Mansfield et al., 2003; Sauders et al., 2006; Vazquez-Boland et al., 2001).

ACKNOWLEDGEMENTS
This work was supported by a USDA Special Research Grant (200534459-15625). We thank Qi Sun for helpful discussions. This research was conducted in part by using the resources of the Cornell Theory Center, which receives funding from Cornell University, New York State, federal agencies, foundations, and corporate partners.

REFERENCES


Table 1. Recombination events in inlA allelic types

A detailed table with all fragments identified by GENECONV or visual evaluation is included as Supplementary Table S3 with the online version of this paper.

<table>
<thead>
<tr>
<th>Event</th>
<th>Approach*</th>
<th>Recipient allelic type (lineage)†</th>
<th>Donor allelic type (lineage)†</th>
<th>Break points‡</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>AT 13 (II)</td>
<td>AT 9 (I)</td>
<td>56-1013</td>
</tr>
<tr>
<td>1</td>
<td>G1, G2</td>
<td>AT 11 (II)</td>
<td>AT 2, 4, 5, 6, 7, 8 (I); AT 23 (III)</td>
<td>1955</td>
</tr>
<tr>
<td>2</td>
<td>G1, G2, G3</td>
<td>AT 23 (III)</td>
<td>AT 10, 17 (II)</td>
<td>1427</td>
</tr>
<tr>
<td>3</td>
<td>G2, G3, NS</td>
<td>AT 18 (II)</td>
<td>AT 23 (III)</td>
<td>238-331</td>
</tr>
<tr>
<td>4</td>
<td>G2, G3, NS</td>
<td>AT 12, 20 (II)</td>
<td>AT 9 (I); AT 20, 21, 22 (II)</td>
<td>1624-1633</td>
</tr>
<tr>
<td>5</td>
<td>G1, G2, G3, NS</td>
<td>AT 20, 21 (II)</td>
<td>AT 13, 18, 19 (II)</td>
<td>1943-1955</td>
</tr>
<tr>
<td>6</td>
<td>G1, G2, G3, NS</td>
<td>AT 16 (II)</td>
<td>AT 10, 14, 15, 16, 17, 18, 19 (II)</td>
<td>64</td>
</tr>
<tr>
<td>7</td>
<td>NS</td>
<td>AT 16 (II)</td>
<td>AT 1, 2, 3, 4, 6, 7, 8 (I)</td>
<td>1666</td>
</tr>
<tr>
<td>8</td>
<td>NS</td>
<td>AT 10 (II)</td>
<td>AT 1, 3, 4, 6, 8 (I)</td>
<td>&lt;1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| Events not identified by GENECONV
| 10    | VS        | AT 10 (II)                        | Lineage I                     | 255 | 352   |
| 11    | VS        | AT 13 (II)                        | Lineage I                     | 1672 | 1831  |
| 12    | VS        | AT 19 (II)                        | Lineage I                     | 255  | 684   |
| 13    | VS        | AT 19 (II)                        | Lineage I                     | 987  | 1221  |

*This column indicates which approach detected a given recombination event; approaches vary in the level and type of mismatches allowed in the recombinant fragment. G1, G2, and G3 represent different G scales in GENECONV; mismatch penalties increase from G1 to G3; NS (no singleton) indicates that a recombination event was detected using an alignment that had all singletons removed and model G0 in GENECONV (which allows for no mismatches); VS (visualization) indicates that recombinant fragments were identified by visual evaluation of polymorphic site alignments.

†inlA allelic types (AT) were visually identified as recipient or donor using polymorphic site alignments.

‡Breakpoints shown are from GENECONV outputs (for G1, G2, G3 and NS) or inferred by visualization (for fragments identified by VS); a range of breakpoints is given if GENECONV identified multiple overlapping fragments with different breakpoints (a detailed table with all fragments is provided as Supplementary Table S3).

Table 2. Summary of positive selection analyses for full-length inlA sequence

| Model* | P† | £‡ | Estimates of parameters§ | Positively selected sites|| |
|--------|----|----|--------------------------|-----------------------|
| M0     | 1  | 5109.03 | \(\omega=0.131\) | None |
| M1a    | 2  | 5019.55 | \(\omega=0.095\); \(\theta=1\) | Not allowed |
| M2a    | 3  | 5013.18 | \(\omega=0.938\); \(\theta=1\) | 187⁷⁷, 594⁸⁸ |
| M3     | 5  | 5013.18 | \(\omega=0.003\); \(\theta=0.915\); \(\omega=2.063\) | 3⁷⁸, 19⁷⁸, 32, 39, 51⁷⁸, 94⁸⁸, 118⁸⁸, 142⁷⁸, 157⁷⁸, 187⁷⁸, 416, 420⁸⁸, 426⁷⁸, 454⁷⁸, 472⁸⁸, 474, 476, 500⁸⁸, 530⁷⁸, 538⁸⁸, 539, 544, 546⁸⁸, 557, 598⁸⁸, 564, 572⁸⁸, 594⁸⁸, 644⁸⁸, 648⁸⁸, 549, 652⁸⁸, 664⁸⁸, 671, 683⁸⁸, 707⁸⁸, 728, 735, 738⁸⁸, 764⁸⁸, 765, 774, 781, 790 |
| M7     | 2  | 5019.62 | \(p=0.005\); \(q=0.049\) | Not allowed |
| M8     | 4  | 5013.18 | \(p=0.837\); \(q=99.000\); \(\omega=2.063\) | 94⁸⁸, 187⁷⁸, 594⁸⁸, 764⁸⁸ |

*Model (e.g. M0, model 0); models are detailed in Yang et al. (2000).
†Number of parameters estimated.
‡Likelihood score.
§For M1, M2a and M3, \(p\) indicates the proportion of sites that fall into a given \(\omega\) category; for M7 and M8, \(p\) and \(q\) are the parameters of the beta distribution; for M8, \(p\) is the proportion of sites that falls into the unconstrained \(\omega\) category. All \(\omega\) values >1 are in bold.

⁷⁷, ⁸⁸ and * indicate amino acid sites with posterior probabilities >0.95 and >0.99, respectively, of being under positive selection; NEB posterior probabilities were calculated for M3 while BEB posterior probabilities were calculated for M2a and M8.
Fig. 1. Single nucleotide deletion at 5’ of *inlA* that leads to a premature stop codon after aa 8. Isolate FSL S4-497 represents a lineage II isolate with a homopolymeric tract of seven adenines between nt 6 and 12; FSL F2-640 represents a lineage II isolate with deletion of one adenine in the homopolymeric tract; FSL E1-119 represents a lineage I isolate with a guanine after the second adenine in the homopolymeric tract (the same sequence was present in all lineage I isolates and the one lineage III isolate characterized).
Fig. 2. inIA phylogeny based on 2400 bp nucleotide sequence data for 40 isolates. The tree was constructed using a maximum-likelihood approach and the HKY + G + I model without a molecular clock. Isolate FSL F2-695 (lineage III) was set as outgroup. Bootstrap values greater than 50 are shown in the respective branches. Isolates are shown to indicate isolate number (see Supplementary Table S1 for details) and isolate sources, which are shown in parentheses with a one-letter code indicating isolates that were obtained from human clinical cases (H), animals (A), foods (F) or natural environments (E). Arrows indicate sequences in which fragments were introduced by horizontal gene transfer (as determined by GENECONV analysis and/or visually); event numbers are the same as those used in Table 1 and Fig. 3. The symbols α and β indicate branches used to test the positive selection for lineage I and lineage II, respectively.

Table 3. L. monocytogenes strains and isolates used for invasion assays (IA) and Western blot (WB)

<table>
<thead>
<tr>
<th>Strain/isolate</th>
<th>Relevant strain/isolate characteristics</th>
<th>Experiment</th>
<th>Source or reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>104035</td>
<td>Laboratory control strain</td>
<td>IA &amp; WB</td>
<td>Bishop &amp; Hinrichs (1987)</td>
</tr>
<tr>
<td>FSL K4-006</td>
<td>104035 ΔinIA (ΔI4005)</td>
<td>IA &amp; WB</td>
<td>Bakardjieva et al. (2004)</td>
</tr>
<tr>
<td>FSL 4-2-007</td>
<td>Environmental isolate, ribotype DUP-1039G, full-length inIA</td>
<td>IA</td>
<td>This article</td>
</tr>
<tr>
<td>FSL F2-723</td>
<td>Food isolate, ribotype DUP-1039C, inIA frameshift mutation resulting in a stop codon after as 8</td>
<td>IA &amp; WB</td>
<td>This article</td>
</tr>
<tr>
<td>FSL F2-640</td>
<td>Food isolate, ribotype DUP-1039C, inIA frameshift mutation resulting in a stop codon after as 8</td>
<td>IA</td>
<td>This article</td>
</tr>
</tbody>
</table>
Fig. 3. Schematic representation of recombination events 5, 6 and 7. The grey sequence is a lineage I sequence that is most closely related to the iniA sequence for isolates FSL F2-637 and FSL F2-672 (iniA nq AT 9). The white sequence represents a lineage II iniA sequence likely representing the ancestor of the iniA sequences in lineage II isolates FSL N4-292, FSL S6-072, FSL S4-766 and FSL S4-887. The hatched sequence represents a lineage II iniA sequence most closely related to the iniA sequences for isolate FSL S4-296. The black sequence represents a lineage II iniA sequence most closely related to the iniA sequences for isolates FSL S4-304. Bold isolate names identify iniA sequences for isolates that were sequenced as part of this study. Numbers above sequences indicate break points; the 3’ break point for event 6 is located 3’ of the sequenced fragment and is thus not indicated.

Fig. 4. Recombination event 8, which is responsible for generation of the iniA sequence found in isolates FSL F2-590, FSL F2-634 and FSL F2-639. (a) Alignment of the polymorphic sites of the iniA sequences identified by GENECONV as being involved in this event. FSL S4-497 (labelled as s4497) and FSL F2-683 (labelled as f2683) are lineage II isolates; iniA sequences for these isolates are more closely related to the iniA sequence for FSL F2-590 (a lineage II strain, as determined by ribotyping) than the iniA sequences for the seven lineage I isolates included (labelled as e1124, e1126, f2897, f2899, f2667, f2669, f2671, f2698, f2699, f2601 and f2602). Shaded areas are nucleotides that are identical to the F2-590 sequence. Numbers above the alignment represent the positions of the polymorphic sites in the alignment. (b) Schematic representation of recombination event 8; nucleotide sites above the sequence represent the location of the 5’ and 3’ breakpoints.
Fig. 5. Schematic representation of InIA with recombination events and positively selected sites. Recombinant fragments (1–13; numbers are identical to those used in Table 1) are shown above the structure. The functional regions are shown below the structure: SP, signal peptide; α, α-helix domain; LRR, leucine-rich repeat; IR, inter-region; ID, interdomain; PGS, proline/glycine-rich segment; TM, transmembrane domain. Positively selected sites identified by M8 are represented as vertical lines (the codon for aa 187 was identified by M2a and M8 performed on the full-length inIA sequences as well as by M8 performed on the LRR regions only). The table lists the ω (dω/dH) value for each region (as determined by M0) as well as the number of polymorphic amino acids in a given region (pol. aa). The results of comparisons between M0 and M3 (M0 vs M3) and M7 and M8 (M7 vs M8) are also shown; the model that best describes the substitution patterns for a given region is indicated (e.g. M8 indicates that M8 was accepted). ‘ω (M8)’ indicates the ω value for sites classified by M8 into the site category with ω>1; ‘frequency of ω (M8)’ indicates the frequency of sites classified by M8 in the site category with ω>1. NC indicates regions for which values were not calculated.

Fig. 6. Partial structure of InIA (grey) in contact with the human receptor E-cadherin (orange). The two sites with posterior probabilities >95% of being under positive selection (i.e. sites 94 and 187) are shown in red; the two other sites with posterior probabilities >95% of being under positive selection (i.e. sites 594 and 784) are located outside the region encompassed by this structure. Polymorphic sites that are not under positive selection are shown in light blue.
**Fig. 7.** InIA Western blot to determine InIA presence in a *L. monocytogenes* isolate with premature stop codons after InIA aa 8. Western blot was performed with isolate FSL F2-723 (which has a premature InIA stop codon) and strain 10403S, which encodes a full-length InIA (Table 3). Western blot analysis was performed on cytoplasmic protein fractions (C) and on proteins precipitated from culture supernatants (S). The expected molecular mass of InIA is approximately 80 kDa.