Genomics Data 2 (2014) 219–225

Contents lists available at ScienceDirect

Genomics Data

journal homepage: http://www.journals.elsevier.com/genomics-data/

Genome comparison of *Listeria monocytogenes* serotype 4a strain HCC23 with selected lineage I and lineage II *L. monocytogenes* strains and other *Listeria* strains

Debarati Paul [a], Chelsea Steele [b], Janet R. Donaldson [c], Michelle M. Banes [b], Ranjit Kumar [b], Susan M. Bridges [d], Mark Arick II [e], Mark L. Lawrence [b,*]

[a] Amity Institute of Biotechnology, Amity University, Noida, India
[b] College of Veterinary Medicine, Mississippi State University, Mississippi State, MS, USA
[c] Department of Biological Sciences, Mississippi State University, Mississippi State, MS, USA
[d] Department of Computer Sciences and Engineering, Mississippi State University, Mississippi State, MS, USA
[e] Institute for Genomics, Biocomputing, and Biotechnology, Mississippi State University, Mississippi State, MS, USA

ARTICLE INFO

Article history:
Received 20 January 2014
Received in revised form 30 April 2014
Accepted 11 June 2014
Available online 7 August 2014

Keywords:
Listeria monocytogenes
Comparative genomics
Genetic lineage
Virulence

ABSTRACT

More than 98% of reported human listeriosis cases are caused by specific serotypes within genetic lineages I and II. The genome sequence of *Listeria monocytogenes* lineage III strain HCC23 (serotype 4a) enables whole genomic comparisons across all three *L. monocytogenes* lineages. Protein cluster analysis indicated that strain HCC23 has the most unique protein pairs with nonpathogenic species *Listeria innocua*. Orthology analysis of the genome sequences of representative strains from the three *L. monocytogenes* genetic lineages and *L. innocua* (CLIP11262) identified 319 proteins unique to nonpathogenic strains HCC23 and CLIP11262 and 58 proteins unique to pathogenic strains F2365 and EGD-e. BLAST comparison of these proteins with all the sequenced *L. monocytogenes* and *L. innocua* revealed 126 proteins unique to serotype 4a and/or *L. innocua*; 14 proteins were only found in pathogenic serotypes. Some of the 58 proteins unique to pathogenic strains F2365 and EGD-e were previously published and are already known to contribute to listerial virulence.

Introduction

Listeria monocytogenes is an opportunistic, intracellular pathogen that causes foodborne infections in animals and humans. Immunocompromised individuals, such as elderly, pregnant, and neonates are at particular risk for listeriosis, whose clinical manifestations include meningitis, meningoencephalitis, septicemia, abortion, perinatal infections, and gastroenteritis [1]. *L. monocytogenes* has the ability to cross tight host barriers, including the intestine, blood–brain barrier, and fetoplacental barrier [2,3]. *L. monocytogenes* also has the ability to adapt to a wide range of conditions such as refrigeration (2–4 °C), low pH, high sodium salt concentrations, and the host immune system (including inside professional phagocytes) [4]. As a result, *L. monocytogenes* causes a variety of clinical syndromes, and it has been isolated from an array of both raw and processed foods, including dairy products, meat products, fresh produce, and fish products.

Based on multigene phylogenetic analyses, *L. monocytogenes* consists of at least three primary genetic lineages: lineages I, II, and III [5,6]. Each of these lineages is primarily comprised of specific serotypes, with lineage I containing serotypes 4b, 3b, 1/2b, lineage II containing serotypes 1/2a, 1/2c, 3a, and lineage III containing serotypes 4a and 4c. These serologic and genetic subtypes are clinically significant; more than 98% of reported human listeriosis cases are caused by serotypes within lineages I and II (1/2a, 1/2c, 1/2b, and 4b). Serotypes within lineage III (4a and 4c) are usually not related to disease outbreaks even though they are commonly isolated from various environmental and food specimens [2].

Whole genome comparisons between lineages I and II using *L. monocytogenes* strains F2365 (4b) and EGD-e (1/2a) revealed a large number of single nucleotide polymorphisms between *L. monocytogenes* strains and some gene additions/deletions [7]. The comparison also showed a high degree of synteny between genetic lineages. Comparison of *L. monocytogenes* strain EGD-e with nonpathogenic *Listeria innocua* strain CLIP1182 also identified potential genetic differences responsible for pathogenicity [8]. Comparative genomic analyses of *Listeria* species revealed species-specific adaptations [9]. Analysis of the listerial "pangenome" allowed identification of lineage-specific *L. monocytogenes* genes, particularly in carbohydrate utilization and stress resistance [10]. A comparison of representative lineage I, lineage II, and lineage III strains was previously conducted that focused on gene sequence

* Corresponding author at: College of Veterinary Medicine, 240 Wise Center Drive, Mississippi State, MS 39762, USA. Fax: +1 662 325 1193.
E-mail address: lawrence@cvm.msstate.edu (M.L. Lawrence).

http://dx.doi.org/10.1016/j.gdata.2014.06.010

comparisons between the lineages and did not include a thorough orthology analysis [11].

L. monocytogenes strain HCC23 (serotype 4a lineage III) isolated from a healthy channel catfish [12] is nonpathogenic in mice, even when given by injection [13]. Lineage III strains are underrepresented in both food contamination and human cases of listeriosis even though the *prfA* virulence gene cluster is present [2,10]. We previously reported the genome sequence of *L. monocytogenes* strain HCC23, which is the first strain to be sequenced from lineage III [14]. Although strain HCC23 has been used by other groups in comparative genomics [10,11], a detailed description and comparison of the strain HCC23 genome has not been published. In our analysis, we focus on comparison with representative strains from other listerial lineages, allowing us to focus on unique features of this particular strain. We also emphasized orthology analysis in our comparison, which was not done in a previous comparison of representative strains from the three *L. monocytogenes* lineages [11]. The orthology analysis was then extended to all the sequenced *L. monocytogenes* and *L. innocua* strains in GenBank by BLAST analysis.

Materials and methods

Sequencing, assembly, and annotation

The complete genome sequence of *L. monocytogenes* serotype 4a strain HCC23 was determined as described [14]. The completed genome has been deposited in DDBJ/EMBL/GenBank under accession no. CP001175.

Synteny analysis

Alignment of the complete genome of strain HCC23 with 4b strain F2365 and 1/2a strain EGD-e was conducted using Mage (Genoscope) web interface. The genome synteny visualization was generated by a custom C++ program (parallelCoord.exte) utilizing OpenGL. The program visualizes the correspondence of the locations of homologous genes on the genomes of different strains. Input consists of a single file containing the following information about each genome, one item per line: genome name, genome length, specification of colors for labels, and one line for each gene with the gene locations. The program provides capabilities for labeling the genomes, highlighting genes, and placing additional labels anywhere in the visualization. The genome color and size of the genome labels is also customizable.

Clustering algorithm

The algorithm used for clustering protein coding genes was a modification of the method described by Hiller et al. [15]. A custom BioPython script was developed to interface with SQLite and NCBI stand-alone BLAST. All predicted proteins were searched against all genomes (*L. monocytogenes* strain HCC23, EGD-e, F2365, CLIP81459, *L. innocua* strain CLIP11262, and *Listeria welshimeri* strain SLCC5334) translated in six reading frames using BLAST. We used single link clustering to assign genes to clusters. For every gene in a cluster, it was required to share at least 70% identity over 70% of its length with one or more other genes in the cluster. At least one sequence in each cluster contained at least 120 residues. Clusters were stored in a SQLite database and then classified as core, unique, or distributed. Core clusters contained at least one representative from each genome. Unique clusters contained genes from only a single genome, and distributed clusters contained genes from more than one genome but not from all. Note that genes from the same genome can be assigned to the same cluster.

Orthology analysis

L. monocytogenes strains F2365, EGD-e, and HCC23 and *L. innocua* strain CLIP11262 were compared using Inparanoid [37] to detect orthologs across strains based on best reciprocal BLAST hits. All possible pairwise combinations were run for each protein coding gene. To be considered an orthologous group, all possible combinations of best reciprocal BLAST hits had to match in all pairwise comparisons within the group. Additional comparisons were done for regions of interest using BLAST and MegAlign (Lasergene). Orthology trends identified based on Inparanoid analysis of these four listerial strains were then confirmed by conducting BLAST analysis of each protein against all of the other sequenced strains in the species *L. monocytogenes* and *L. innocua* (Table 4). To be considered an ortholog in this BLAST comparison, the protein match had to have an e value of $<1 \times 10^{-5}$.

Results and discussion

Genome features

Genome features of strain HCC23 are summarized in Steele et al. [14]. A comparison of genome features from *L. monocytogenes* strains HCC23, F2365, EGD-e, and *L. innocua* strain CLIP11262 is presented in Table 1. The six ribosomal RNA operons in serotype 4a strain HCC23 are arranged almost identically to those of serotype 4b strain F2365. Two of the ribosomal RNA operons from both of these *L. monocytogenes* strains are located in tandem. However, each genome has a unique intergenic region of less than 300 bp separating the tandem ribosomal RNA operons. COG group classifications of predicted strain HCC23 proteins and protein BLAST results with strain EGD-e are shown in Fig. 1. Protein BLAST results revealed 11 regions unique in strain HCC23 relative to strain EGD-e.

Synteny analysis

Synteny of the six genomes compared was well conserved. In particular, the three genomes of strains HCC23, F2365, and EGD-e aligned with each other perfectly except at a few loci (Fig. 2); examples are listed in Table S1. Some of these genes are potentially related to pathogenesis of *L. monocytogenes*, including some encoding predicted surface proteins and some encoding regulatory proteins. Synteny of major virulence determinants, such as internalins InlA and InlB (internalization), listeriolysin and phospholipases PlcA and PlcB (escape from the host vacuole), ActA (movement within the host cell cytoplasm), or the master virulence regulator PrfA, is conserved in *L. monocytogenes*.

Cluster analysis

We used ortholog cluster analysis [15] to analyze the listerial "supragenome." In the analysis, we included strain HCC23, *L. monocytogenes* serotype 1/2a strain EGD-e, serotype 4b strains F2365 and CLIP81459, *L. innocua* strain CLIP11262, and *L. welshimeri* strain SLCC5334. Almost 80% of the 17,149 predicted proteins in these six strains were in core orthologous clusters (those containing an orthologous protein in all six genomes) (Table 2). These core orthologous clusters made up more than half of all the clusters. In addition, *Listeria* had a sizeable repertoire (18.6%) of distributed protein coding genes that are present in at least two, but not all, strains. About a quarter of the protein clusters were unique, meaning the cluster contained a predicted protein from only one strain; however, these

Table 1
General summaries of three *L. monocytogenes* genomes and a *L. innocua* genome.

Strain	HCC23	EGD-e	F2365	CLIP11262
Serotype	4a	1/2a	4b	*L. innocua*
Chromosome size	2,976,212	2,944,528	2,905,187	3,093,113
G + C content (%)	38.2	37.98	38.04	37.38
No. of CDSs	2974	2846	2821	2968
No. of rRNA genes	18	18	18	18
No. of tRNA genes	67	67	67	66

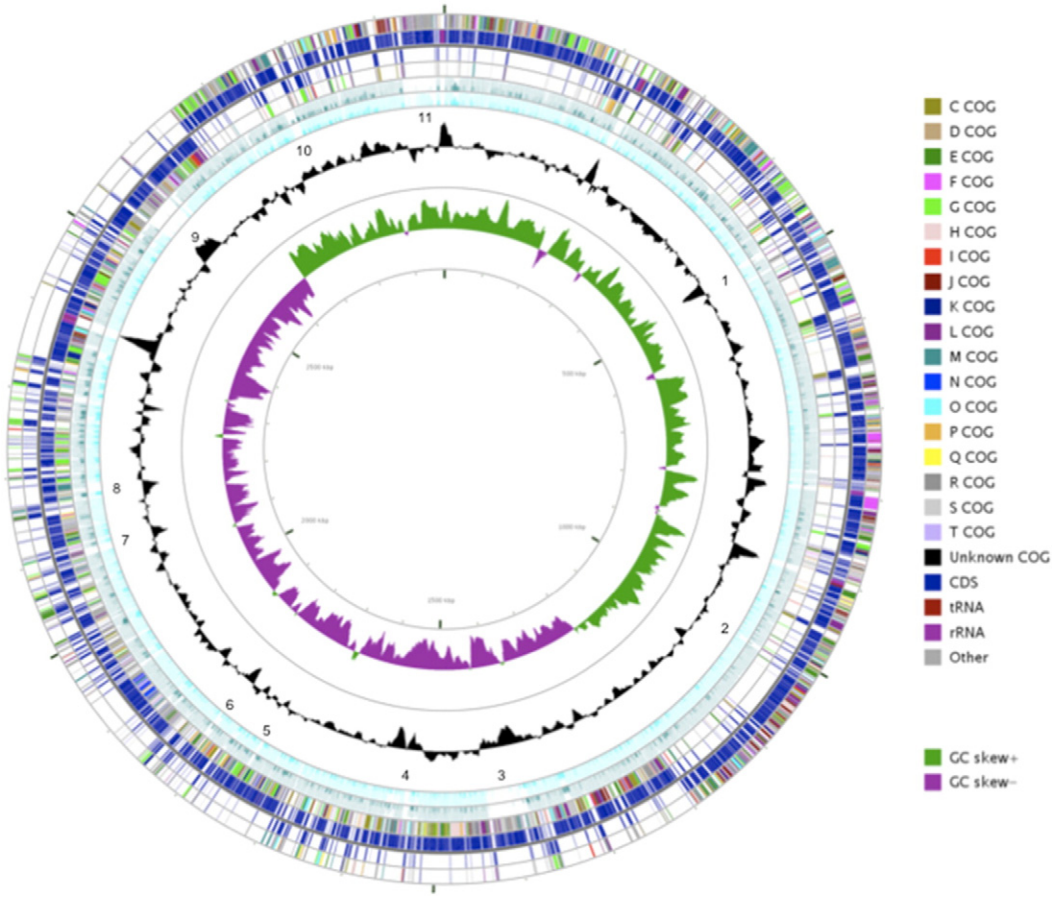

Fig. 1. Circular map of the *Listeria monocytogenes* serotype 4a strain HCC23 genome. From outside to inside, the circles represent: 1) COG group classification of proteins encoded on the positive strand, 2) CDS, tRNA, and rRNA on the positive strand, 3) CDS, tRNA and rRNA on the negative strand, 4) COG group classifications of proteins encoded on the negative strand, 5) BLAST results with *L. monocytogenes* serotype 1/2a strain EGD-e, 6) BLAST results with *L. monocytogenes* serotype 4b strain F2365, 7) G + C content, and 8) GC skew. BLAST results identified 11 clusters of strain HCC23-specific genes (labeled 1–11).
COG group classifications:

C	Energy production and conversion
D	Cell cycle control mitosis and meiosis
E	Amino acid transport and metabolism
F	Nucleotide transport and metabolism
G	Carbohydrate metabolism and transport
H	Coenzyme transport and metabolism
I	Lipid transport and metabolism
J	Translation
K	Transcription
L	DNA replication, recombination, and repair
M	Cell wall/membrane biogenesis
N	Cell motility
O	Posttranslational modification, protein turnover
P	Inorganic ion transport and metabolism
Q	Secondary metabolism biosynthesis, transport and catabolism
S	Function unknown
T	Signal transduction mechanisms
U	Intracellular trafficking and secretion
V	Defense mechanisms.

only accounted for 6% of the proteins. Overall, the results support previously reported strong conservation of core listerial functions with considerable evidence for a distributed genome among *Listeria* that could be the result of a combination of gene deletion and horizontal gene exchange.

Strain HCC23 and *L. innocua* strain CLIP11262 had similar numbers of unique protein clusters and distributed protein clusters (Table 3). Both strains had higher numbers of unique proteins than the other three *L. monocytogenes* strains and lower numbers of distributed protein clusters. Strains F2365 and CLIP81459 had the lowest number of unique clusters because these two strains are in the same genetic lineage. In pairwise comparisons, strain HCC23 had the highest number of unique protein pairs with *L. innocua* strain CLIP11262 (Fig. 3). In fact, strain HCC23 and *L. innocua* strain CLIP11262 had the highest number of unique protein pairs in all the pairwise comparisons, including the number of protein pairs unique to the two 4b strains F2365 and CLIP81459. Not surprisingly, *L. welshimeri* strain SLCC5334 had the highest percentage of proteins unique to that strain (Table 3). It also had the lowest percentage of proteins in distributed clusters. Therefore, although this strain retains core listerial functions, it has the most unique protein adaptations and has less evidence for horizontal gene transfer with the other listerial strains. *L. welshimeri* strain

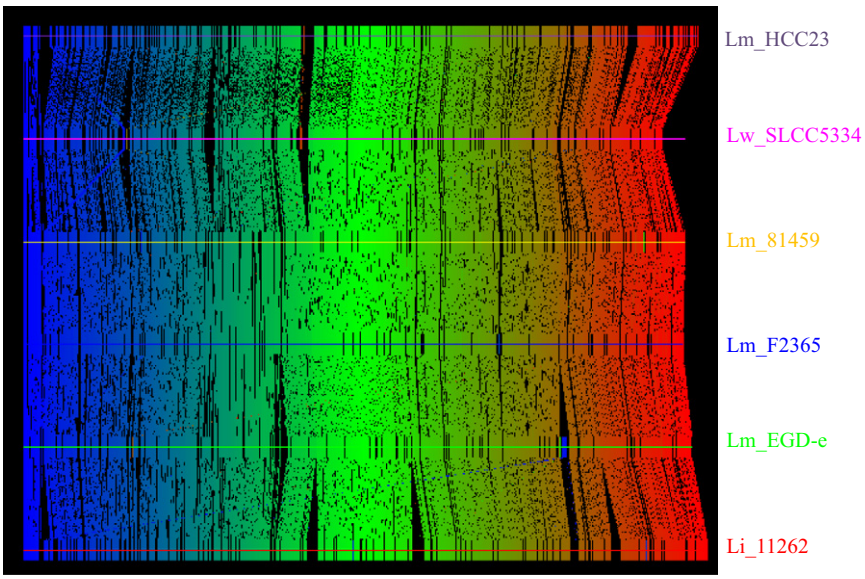

Fig. 2. Visualization of synteny analysis of six *Listeria* strains. Parallel axes with lengths corresponding to the genome sizes are used to represent each genome, and a ribbon corresponding to the width of each gene is drawn to connect corresponding genes on the parallel axes. The use of a color spectrum across the genome is used to facilitate visualization of inversions, deletions, and insertions.

SLCC5334 had the most differences and fewest shared proteins with strain HCC23.

Orthology analysis

To determine which proteins are unique to different strains and strain combinations, we used InParanoid analysis to focus on four strains: *L. monocytogenes* strains HCC23, EGD-e, and F2365, and *L. innocua* strain CLIP11262. Orthology analysis used more relaxed parameters than the cluster analysis, allowing identification of individual proteins unique to strains. Between these four strains, there were 2422 "core" proteins (orthologs present in all strains). There were 31 proteins that did not have the same matching best reciprocal BLAST hits in all pairwise comparisons; these proteins were not further considered.

Based on orthology comparison of these four listerial strains, 319 proteins that are unique to nonpathogenic strains HCC23 and *L. innocua* CLIP11262 were identified. Strain HCC23 had 253 unique proteins, and 66 proteins were only present in strains HCC23 and CLIP11262. BLAST analysis of these 319 proteins against all of the sequenced listerial strains in GenBank was conducted to determine which of these proteins are actually unique to serotype 4a and *L. innocua*. Sixty of the 319 proteins were found to be unique to serotype 4a, and 46 proteins of the proteins were present in only serotype 4a and *L. innocua* (Table S2). The BLAST analysis also revealed that an additional 26 proteins were only present in serotype 4a strains, *L. innocua* strains, and the other lineage III serotype 4c strain; these proteins are also included in Table S2. Therefore, a total of 132 of the 319 proteins were found to be unique to *L. monocytogenes* lineage III (serotypes 4a and 4c) and *L. innocua* (Table S2). Of these, 24 were conserved domain/conserved hypothetical proteins, and 76 were hypothetical proteins. A restriction endonuclease protein (LMHCC_2321) was found to be unique to serotype 4a and 4c strains. A transcriptional regulator (LMHCC_0466) was identified that is only in serotype 4a strains. Two proteins in the SMI1/KNR4 family that are unique to serotype 4a strains were identified (LMHCC_2100 and LMHCC_2748); proteins in this family are possibly primary bacterial immunity proteins that function as toxins to discriminate "self" from "non-self" strains [16].

Because strain HCC23 is avirulent, perhaps the most interesting orthology comparison utilizing this strain is what it does *not* have. In total, 58 proteins were identified that are present in strains EGD-e and F2365 and missing in strains HCC23 and CLIP11262 (Table S3). Several of the 58 proteins not found in strains HCC23 and CLIP11262 are known virulence factors. Two of these proteins catalyze the final two steps in the non-mevalonate pathway for isoprenoid biosynthesis, GcpE (4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase) and LytB (4-hydroxy-3-methylbut-2-enyl diphosphate reductase). The genes encoding these enzymes are required for full virulence and are expressed in vivo [17]. InlC, InlH (InlC2), and InlJ are also not found in strains HCC23 or CLIP11262. All three are known virulence factors [18–20] and InlC is a potential marker for virulence [21]. InlH appears to modulate host inflammation [22]. InlJ is expressed in vivo and functions as an adhesin [23].

Table 2
Number of CDS and orthologous clusters present in six *Listeria* strains.

Gene category	Number of orthologous clusters (% of total)	Number of CDS (% of proteins)
Core*	2,231 (56.5)	13,473 (78.8)
Distributed†	734 (18.6)	2,641 (15.4)
Unique‡	984 (24.9)	984 (5.8)
Total	3,949 (100)	17,098 (100)

* Core clusters: clusters that contain an orthologous protein from all six strains.
† Distributed clusters: clusters that contain orthologous proteins from more than one but not all strains.
‡ Unique clusters: clusters containing only one protein from a single strain.

Table 3
Numbers of orthologous clusters for individual *Listeria* strains.

Strain name	Number of orthologous clusters*	Percent unique clusters†	Percent distributed clusters
Li_11262	2666	8.8	14.2
Lm_81459	2713	0.9	17.1
Lm_EGD-e	2712	4.1	16.9
Lw_SLCC5334	2461	11.6	7.4
Lm_HCC23	2687	9.3	14.7
Lm_F2365	2715	2.8	17.2

‡Percentage of distributed orthologous clusters relative to the total number of orthologous clusters in each strain.
* Total number of orthologous clusters present in each strain.
† Percentage of unique orthologous clusters relative to the total number of orthologous clusters in each strain.

Accession#	Lm_EGD-e	Lm_F2365	Lm_81459	Lw_SLCC5334	Lm_HCC23	Category
Li_11262	2523	2478	2504	2348	2541	Shared
	678	736	629	950	756	Different
	9	3	7	31	55	Pair Unique
	Lm_EGD-e	2587	2596	2343	2555	Shared
		442	369	884	652	Different
		5	1	14	20	Pair Unique
		Lm_F2365	2655	2348	2525	Shared
			219	842	680	Different
			41	13	4	Pair Unique
			Lm_81459	2339	2536	Shared
				805	603	Different
				1	2	Pair Unique
				Lw_SLCC5334	2336	Shared
					1012	Different
					13	Pair Unique

Fig. 3. Protein cluster analysis of six *Listeria* strains. Shared clusters are the number of protein clusters shared by both strains. Different clusters are the number of protein clusters present in only one of the strains. Pair unique clusters are the number of protein clusters present in these two strains and no other strains.

Interestingly, several of the proteins unique to pathogenic strains EGD-e and F2365 were encoded by genes located in clusters containing two or more genes specific to these strains. One of these loci has six genes, including genes encoding a putrescine carbamoyltransferase, an amino acid permease, a peptidyl-arginine deiminase-like protein, and carbamate kinase (*lmo0084–lmo0089*). These four genes are organized similar to the agmatine catabolic locus in *Enterococcus faecalis* [24]. The gene encoding peptidyl-arginine deiminase-like protein was previously associated with lineage I and II *L. monocytogenes* strains [25], and this gene cluster contributes to virulence in mice [26]. Another cluster of seven genes unique to strains EGD-e and F2365 (*lmo1030–lmo1036*) contains one gene encoding a putative glycerol kinase ($GlpK_2$), which was previously identified as *L. monocytogenes*-specific [27]. Our sequencing results show that *lmo1034* is actually not present in some *L. monocytogenes* serotype 4a strains. *lmo1034* is not strongly induced by extracellular glycerol, but loss of this gene affects the ability of *L. monocytogenes* to replicate in Caco-2 cells [27]. Other gene clusters unique to strains EGD-e and F2365 include one containing four genes that encode putative PTS system components and a hypothetical protein (*lmo630–lmo634*) and another that encodes six hypothetical proteins (*lmo0069–lmo0074*).

When each of these 58 proteins was compared against all the other sequenced *L. monocytogenes* and *L. innocua* strains (Table 4) by BLAST, 14 proteins were identified that are not present in any *L. monocytogenes* serotype 4a strain or *L. innocua* strain (Table S3). Twelve of the 14 proteins missing in all serotype 4a and *L. innocua* strains encode hypothetical proteins. One encodes a cell wall surface anchor protein, and one encodes a metallo-beta-lactamase (SepA).

Metabolic comparison

Table S4 shows that proteins involved in energy metabolism and motility in strains HCC23, F2365, and EGD-e are almost identical. As noted above, some proteins involved in nitrogen metabolism (including carbamate kinase) are missing in strain HCC23 compared to the other listerial strains. Carbohydrate metabolism in these three strains appears conserved (not shown).

Prophages and CRISPRs

Strain HCC23 provides an interesting model of evolutionary recombination with its mosaic of lysogenic bacteriophages. It has three prophage regions, each sharing significant identity with prophages found in the genomes of other *Listeria* strains. Putative prophage HCC23.1 is most similar to *comK*-specific prophages in previously sequenced *L. monocytogenes* strains. It is inserted into tRNA-Lys4 and has only 27.9% nucleotide identity to phage A118, which is a well-conserved phage among *L. monocytogenes*. Six of nine prophages in previously sequenced *L. monocytogenes* strains show significant homology to phage A118 in ≥11 ORFs [7]. Interestingly, *L. innocua* strain CLIP11262 is the

Table 4
List of *L. monocytogenes* and *L. innocua* strains used for BLAST analysis.

1/2a	1/2b	1/2c	3a	3b	3c	4a	4b	4c	4e	7	*L. innocua*
EGD-e	SLCC2755	FSLR2-561	Finland 1998	SLCC2540	SLCC2479	HCC23	07PF0776	SLCC2376	SLCC2378	SLCC2482	CLIP 11262
La111	FSL J1-175	SLCC2372	SLCC7179			L99	ATCC 19117	FSL J2-071			FSL J1-023
08-5578	FSL J1-194	LO28				M7	J1-220				FSL S4-378
08-5923	FSL J2-064					FSL J1-208	J1816				ATCC 33091
10403S	FSL R2-503					FSL F2-208	L312				
J0161							CLIP 80459				
N53-1							F2365				
SLCC5850							LL195				
F6900							FSL N1-017				
FSL F2-515							HPB2262				
FSL J2-003							H7858				
FSL N3-165											
J2818											
F6854											
NCCP No. 15743											

only other sequenced *Listeria* that has a prophage inserted into the tRNA-Lys4 site. Prophage HCC23.2 has an unknown target, and it is very similar to prophage A118, having 67.2% identity at the nucleotide level and a consensus length of 21,690 bp. ΦHCC23.2 is 39,118 bp in length compared to phage A118 (NC_003216.1), which is 40,834 bp. Prophage HCC23.3 is a PSA-like phage that is 43,265 bp in length (bacteriophage PSA is 37,618 bp). It has 47.2% identity with PSA at the nucleotide level, and the two phages share a consensus sequence of 25,093 bp.

Clustered regularly interspaced palindromic repeats (CRISPRs) have been previously identified in *L. monocytogenes* serotype 1/2a (found at three loci) and *L. innocua* (found at one locus), but they are not present in *L. monocytogenes* serotype 4b [7]. CRISPR loci and *cas* genes in serotype 4a and 1/2a strains have been previously reviewed [11].

Virulence proteins

Proteins encoded within the PrfA virulence locus are well conserved in strain HCC23 compared to F2365 and EGD-e (Table 5). Listeriolysin, phosphatidylcholine phospholipase C (PC-PLC), phosphatidylinositol phospholipase C (PI-PLC), and ActA in strain HCC23 have a higher degree of identity at the amino acid level to strain F2365 compared to EGD-e. PrfA is nearly 100% identical in strains HCC23 and EGD-e with two amino acid substitutions (T:A at position 165 and K:N at position 197). These two amino acid changes surround the helix-turn-helix motif, which is possibly an important location for binding target DNA [28].

Presence of internalin genes is one of the distinguishing features between pathogenic and nonpathogenic *L. monocytogenes* strains [9]. InlA and InlB are critical for invasion of intestinal epithelial and hepatic cells, respectively [29], and InlC is specifically required for cell-to-cell spread [30]. InlA and InlB in strains HCC23, F2365, and EGD-e have a high degree of identity. InlC is not encoded in the strain HCC23 and *L. innocua* strain CLIP11262 genomes as expected from previous reports [21]. InlJ is a protein of the LPXTG-internalin family and is required for virulence [20], and it is also not encoded in strain HCC23 and *L. innocua* strain CLIP11262.

Cell wall and teichoic acid modification proteins

All three of the *L. monocytogenes* strains encode amidases and autolysins that have high identity (Table 5). These include a gene encoding *N*-acetylmuramoyl-L-alanine amidase (family 3) (LMHCC_1048) and a gene encoding a family 4 amidase (LMHCC_1437). A GW repeat domain autolysin (LMHCC_0041) (*ami*) is also encoded; this is the only listerial amidase showing autolysin activity [31]. The three *L. monocytogenes* strains also encode autolysin lyt-G (Exo-beta-*N*-acetylglucosaminidase) (LMHCC_1438), which hydrolyzes the sugar backbone of peptidoglycan between MurNAc and GlcNAc residues of the glycan chain. The three *L. monocytogenes* strains also encode two well conserved penicillin binding proteins (PBPs): PBP3 (PSPB20) and PBP5. In other bacteria, these proteins play roles in β-lactam resistance. PBP3 is central to β-lactam resistance in *Listeria* [32]. PBP5 is now known as PBPD1 and catalyzes the removal of the C-terminal D-alanine residue from peptidoglycan pentapeptides.

Teichoic acids (TA) are electronegative polymers of ribitol-phosphates or glycerol-phosphates with D-Ala residues and sugar residues that vary depending on the serotype. Serotype 4 *L. monocytogenes* strains have GlcNAc in their teichoic acid chains [33], and serotype 4b strains are unique in bearing both galactose and glucose substituents on the GlcNAc of TA. Genes *gtcA* and *gltAB* encode enzymes that are essential for serotype-specific glycosylation of the teichoic acid of *L. monocytogenes* serotype 4b with glucose and galactose [34,35]. Gene *gtc*A encodes an enzyme that catalyzes addition of galactose and glucose to TA of serotype 4b; the gene is present in strain HCC23 as well and is nearly identical in strains HCC23 and F2365 (Table 5). Originally, *gtcA* was thought to be unique to serogroup 4 strains [36]; however, a divergent homolog of this gene is present in EGD-e (80% nucleotide identity and 82% amino acid identity to F2365) in a genomically equivalent location [8,34]. Genes *glt*A and *gltB* encode enzymes required for expression of teichoic acid-associated surface antigens in serotype 4b; specifically, they mediate attachment of glucose substituents in TA [35]. The *gltA–gltB* cassette is found only in strains of the serotype 4b complex (serotype 4b and the highly similar serotypes 4d and 4e). As expected, serotype 4a strain HCC23 does not have the *gltAB* cassette.

Table 5
Sequence identity of some virulence related and surface proteins in *L. monocytogenes* strains EGD-e and F2365 compared to HCC23.

Gene name	EGD-e	F2365
PC-PLC	96.9*	98.9
PlcA	97.5	97.2
ActA	92.2	96.2
PrfA	99.2	99.2
InlA	97.6	99.1
InlB	92.4	96.7
Ami	98.6	98.1
Lyt-G	90.3	90.7
Pbp-3	98.4	98.7
Pbp-5	95.3	95.7
GtcA	84.9	98.3
GltC	100	99.7

* Percent identity compared to strain HCC23 ortholog based on Clustal W alignment using MegAlign (Lasergene).

Conclusions

In the current study, we compared the genome sequences of representative strains from the three *L. monocytogenes* genetic lineages (F2365, serotype 4b and lineage I; EGD-e, and serotype 1/2a lineage II; HCC23, serotype 4a and lineage III) and *L. innocua* (CLIP11262). Orthology analysis identified a core genome of 2422 orthologous proteins between these four strains. Strain HCC23 is nonpathogenic in the mouse model for listeriosis, and it had the most unique protein pairs with nonpathogenic species *L. innocua*. However, many of the proteins unique to these two nonpathogenic strains have poorly characterized functions. Metabolic pathways are well conserved between the three strains, but there is more variation between them in the surface modification proteins they encode, as expected for distinct serotypes. Some well-characterized virulence factors such as the PrfA locus and internalins are well conserved across *L. monocytogenes* lineages. However, orthology analysis allowed identification of 58 proteins unique to pathogenic strains F2365 and EGD-e that are missing in strain HCC23 and CLIP11262. Two of these are characterized and are known to contribute to listerial virulence, but the function of many is still not known. The uncharacterized proteins are interesting candidates to consider as putative virulence determinants.

Acknowledgments

This project was supported by USDA NIFA grant number 2007-35201-17732 and USDA Agricultural Research Service (Agreement No. 58-6202-5-083).

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gdata.2014.06.010.

References

[1] A. Vasquez, J.M. Ramos, E. Pacho, A. Rodriguez-Perez, M. Cuenca-Estrella, J. Esteban, Cutaneous listeriosis in a patient infected with the human immunodeficiency virus. Clin. Infect. Dis. 19 (1994) 988–989.

[2] D. Liu, M.L. Lawrence, M. Wiedmann, L. Gorski, R.E. Mandrell, A.J. Ainsworth, F.W. Austin, *Listeria monocytogenes* subgroups IIIA, IIIB, and IIIC delineate genetically distinct populations with varied pathogenic potential. J. Clin. Microbiol. 44 (2006) 4229–4233.
[3] M.A. Hamon, E. Batsche, B. Regnault, T.N. Tham, S. Seveau, C. Muchardt, P. Cossart, Histone modifications induced by a family of bacterial toxins. Proc. Natl. Acad. Sci. U. S. A. 104 (2007) 13467–13472.
[4] M. Gandhi, M.L. Chikindas, *Listeria*: a foodborne pathogen that knows how to survive. Int. J. Food Microbiol. 113 (2007) 1–15.
[5] C.A. Nadon, D.L. Woodward, C. Young, F.G. Rodgers, M. Wiedmann, Correlations between molecular subtyping and serotyping of *Listeria monocytogenes*. J. Clin. Microbiol. 39 (2001) 2704–2707.
[6] O.F. Rasmussen, P. Skouboe, L. Dons, L. Rossen, J.E. Olsen, *Listeria monocytogenes* exists in at least three evolutionary lines: evidence from flagellin, invasive associated protein and listeriolysin O genes. Microbiology 141 (Pt 9) (1995) 2053–2061.
[7] K.E. Nelson, D.E. Fouts, E.F. Mongodin, J. Ravel, R.T. DeBoy, J.F. Kolonay, D.A. Rasko, S. V. Angiuoli, S.R. Gill, I.T. Paulsen, J. Peterson, O. White, W.C. Nelson, W. Nierman, M.J. Beanan, L.M. Brinkac, S.C. Daugherty, R.J. Dodson, A.S. Durkin, R. Madupu, D.H. Haft, J. Selengut, S. Van Aken, H. Khouri, N. Fedorova, H. Forberger, B. Tran, S. Kathariou, L. D. Wonderling, G.A. Uhlich, D.O. Bayles, J.B. Luchansky, C.M. Fraser, Whole genome comparisons of serotype 4b and 1/2a strains of the food-borne pathogen *Listeria monocytogenes* reveal new insights into the core genome components of this species. Nucleic Acids Res. 32 (2004) 2386–2395.
[8] P. Glaser, L. Frangeul, C. Buchrieser, C. Rusniok, A. Amend, F. Baquero, P. Berche, H. Bloecker, P. Brandt, T. Chakraborty, A. Charbit, F. Chetouani, E. Couve, A. de Daruvar, P. Dehoux, E. Domann, G. Dominguez-Bernal, E. Duchaud, L. Durant, O. Dussurget, K.D. Entian, H. Fsihi, F. Garcia-del Portillo, P. Garrido, L. Gautier, W. Goebel, N. Gomez-Lopez, T. Hain, J. Hauf, D. Jackson, L.M. Jones, U. Kaerst, J. Kreft, M. Kuhn, F. Kunst, G. Kurapkat, E. Madueno, A. Maitournam, J.M. Vicente, E. Ng, H. Nedjari, G. Nordsiek, S. Novella, B. de Pablos, J.C. Perez-Diaz, R. Purcell, B. Remmel, M. Rose, T. Schlueter, N. Simoes, A. Tierrez, J.A. Vazquez-Boland, H. Voss, J. Wehland, P. Cossart, Comparative genomics of *Listeria* species. Science 294 (2001) 849–852.
[9] H.C. den Bakker, C.A. Cummings, V. Ferreira, P. Vatta, R.H. Orsi, L. Degoricija, M. Barker, O. Petrauskene, M.R. Furtado, M. Wiedmann, Comparative genomics of the bacterial genus *Listeria*: genome evolution is characterized by limited gene acquisition and limited gene loss. BMC Genomics 11 (2010) 688.
[10] X. Deng, A.M. Phillippy, Z. Li, S.L. Salzberg, W. Zhang, Probing the pan-genome of *Listeria monocytogenes*: new insights into intraspecific niche expansion and genomic diversification. BMC Genomics 11 (2010) 500.
[11] T. Hain, R. Ghai, A. Billion, C.T. Kuenne, C. Steinweg, B. Izar, W. Mohamed, M.A. Mraheil, E. Domann, S. Schaffrath, U. Karst, A. Goesmann, S. Oehm, A. Puhler, R. Merkl, S. Vorwerk, P. Glaser, P. Garrido, C. Rusniok, C. Buchrieser, W. Goebel, T. Chakraborty, Comparative genomics and transcriptomics of lineages I, II, and III strains of *Listeria monocytogenes*. BMC Genomics 13 (2012) 144.
[12] S. Erdenlig, A.J. Ainsworth, F.W. Austin, Production of monoclonal antibodies to *Listeria monocytogenes* and their application to determine the virulence of isolates from channel catfish. Appl. Environ. Microbiol. 65 (1999) 2827–2832.
[13] D. Liu, A.J. Ainsworth, F.W. Austin, M.L. Lawrence, Use of PCR primers derived from a putative transcriptional regulator gene for species-specific determination of *Listeria monocytogenes*. Int. J. Food Microbiol. 91 (2004) 297–304.
[14] C.L. Steele, J.R. Donaldson, D. Paul, M.M. Banes, T. Arick, S.M. Bridges, M.L. Lawrence, Genome sequence of lineage III *Listeria monocytogenes* strain HCC23. J. Bacteriol. 193 (2011) 3679–3680.
[15] N.L. Hiller, B. Janto, J.S. Hogg, R. Boissy, S. Yu, E. Powell, R. Keefe, N.E. Ehrlich, K. Shen, J. Hayes, K. Barbadora, W. Klimke, D. Dernovoy, T. Tatusova, J. Parkhill, S.D. Bentley, J.C. Post, G.D. Ehrlich, F.Z. Hu, Comparative genomic analyses of seventeen *Streptococcus pneumoniae* strains : insights into the pneumococcal supragenome. J. Bacteriol. 189 (2007) 8186–8195.
[16] D. Zhang, L.M. Iyer, L. Aravind, A novel immunity system for bacterial nucleic acid degrading toxins and its recruitment in various eukaryotic and DNA viral systems. Nucleic Acids Res. 39 (2011) 4532–4552.
[17] M. Begley, P.A. Bron, S. Heuston, P.G. Casey, N. Englert, J. Wiesner, H. Jomaa, C.G. Gahan, C. Hill, Analysis of the isoprenoid biosynthesis pathways in *Listeria monocytogenes* reveals a role for the alternative 2-C-methyl-D-erythritol 4-phosphate pathway in murine infection. Infect. Immun. 76 (2008) 5392–5401.
[18] F. Engelbrecht, C. Dickneite, R. Lampidis, M. Gotz, U. DasGupta, W. Goebel, Sequence comparison of the chromosomal regions encompassing the internalin C genes (*inlC*) of *Listeria monocytogenes* and *L. ivanovii*. Mol. Gen. Genet. 257 (1998) 186–197.
[19] D. Raffelsbauer, A. Bubert, F. Engelbrecht, J. Scheinpflug, A. Simm, J. Hess, S.H. Kaufmann, W. Goebel, The gene cluster *inlC2DE* of *Listeria monocytogenes* contains additional new internalin genes and is important for virulence in mice. Mol. Gen. Genet. 260 (1998) 144–158.
[20] C. Sabet, M. Lecuit, D. Cabanes, P. Cossart, H. Bierne, LPXTG protein InlJ, a newly identified internalin involved in *Listeria monocytogenes* virulence. Infect. Immun. 73 (2005) 6912–6922.
[21] D. Liu, M.L. Lawrence, F.W. Austin, A.J. Ainsworth, A multiplex PCR for species- and virulence-specific determination of *Listeria monocytogenes*. J. Microbiol. Methods 71 (2007) 133–140.
[22] N. Personnic, S. Bruck, M.A. Nahori, A. Toledo-Arana, G. Nikitas, M. Lecuit, O. Dussurget, P. Cossart, H. Bierne, The stress-induced virulence protein InlH controls interleukin-6 production during murine listeriosis, Infect. Immun. 78 1979-1989.
[23] C. Sabet, A. Toledo-Arana, N. Personnic, M. Lecuit, S. Dubrac, O. Poupel, E. Gouin, M.A. Nahori, P. Cossart, H. Bierne, The *Listeria monocytogenes* virulence factor InlJ is specifically expressed in vivo and behaves as an adhesin. Infect. Immun. 76 (2008) 1368–1378.
[24] J.L. Llacer, L.M. Polo, S. Tavarez, B. Alarcon, R. Hilario, V. Rubio, The gene cluster for agmatine catabolism of *Enterococcus faecalis*: study of recombinant putrescine transcarbamylase and agmatine deiminase and a snapshot of agmatine deiminase catalyzing its reaction. J. Bacteriol. 189 (2007) 1254–1265.
[25] J. Chen, L. Jiang, Q. Chen, H. Zhao, X. Luo, X. Chen, W. Fang, lmo0038 is involved in acid and heat stress responses and specific for *Listeria monocytogenes* lineages I and II, and Listeria ivanovii. Foodborne Pathog. Dis. 6 (2009) 365–376.
[26] S. Ryan, M. Begley, C.G. Gahan, C. Hill, Molecular characterization of the arginine deiminase system in *Listeria monocytogenes*: regulation and role in acid tolerance. Environ. Microbiol. 11 (2009) 432–445.
[27] B. Joseph, S. Mertins, R. Stoll, J. Schar, K.R. Umesha, Q. Luo, S. Muller-Altrock, W. Goebel, Glycerol metabolism and PrfA activity in *Listeria monocytogenes*. J. Bacteriol. 190 (2008) 5412–5430.
[28] M. Herler, A. Bubert, M. Goetz, Y. Vega, J.A. Vazquez-Boland, W. Goebel, Positive selection of mutations leading to loss or reduction of transcriptional activity of PrfA, the central regulator of *Listeria monocytogenes* virulence. J. Bacteriol. 183 (2001) 5562–5570.
[29] L. Braun, S. Dramsi, P. Dehoux, H. Bierne, G. Lindahl, P. Cossart, InlB: an invasion protein of *Listeria monocytogenes* with a novel type of surface association. Mol. Microbiol. 25 (1997) 285–294.
[30] T. Rajabian, B. Gavicherla, M. Heisig, S. Muller-Altrock, W. Goebel, S.D. Gray-Owen, K. Ireton, The bacterial virulence factor InlC perturbs apical cell junctions and promotes cell-to-cell spread of Listeria. Nat. Cell Biol. 11 (2009) 1212–1218.
[31] K.E. Beauregard, K.D. Lee, R.J. Collier, J.A. Swanson, pH-dependent perforation of macrophage phagosomes by listeriolysin O from *Listeria monocytogenes*. J. Exp. Med. 186 (1997) 1159–1163.
[32] L.M. Brunt, D.A. Portnoy, E.R. Unanue, Presentation of *Listeria monocytogenes* to CD8+ T cells requires secretion of hemolysin and intracellular bacterial growth. J. Immunol. 145 (1990) 3540–3546.
[33] H. Bierne, P. Cossart, *Listeria monocytogenes* surface proteins: from genome predictions to function. Microbiol. Mol. Biol. Rev. 71 (2007) 377–397.
[34] Y. Cheng, N. Promadej, J.W. Kim, S. Kathariou, Teichoic acid glycosylation mediated by *gtcA* is required for phage adsorption and susceptibility of *Listeria monocytogenes* serotype 4b. Appl. Environ. Microbiol. 74 (2008) 1653–1655.
[35] X.H. Lei, F. Fiedler, Z. Lan, S. Kathariou, A novel serotype-specific gene cassette (*gltA-gltB*) is required for expression of teichoic acid-associated surface antigens in *Listeria monocytogenes* of serotype 4b. J. Bacteriol. 183 (2001) 1133–1139.
[36] N. Promadej, F. Fiedler, P. Cossart, S. Dramsi, S. Kathariou, Cell wall teichoic acid glycosylation in *Listeria monocytogenes* serotype 4b requires gtcA, a novel, serogroup-specific gene. J. Bacteriol. 181 (1999) 418–425.
[37] K.P. O'Brien, M. Remm, E.L. Sonnhammer, Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res. 33 (2005) D476–D480.