R-036: RAPID, SHORT-READ SEQUENCING AND COMPARISON OF 8 DIFFERENT *LISTERIA* STRAINS

Henk C. den Bakker[1], Paolo Vatta[2], Lovorka Degoricija[2], Melissa Barker[2], Craig Cummings[2], Olga Petrauskene[2], Manohar R Furtado[2], Martin Wiedmann[1]

[1]Department of Food Science, Cornell University, Ithaca NY, 14853; [2]Applied Biosystems, Foster City, CA 94404

Abstract

Detection in the food supply of pathogenic bacteria of the genus *Listeria*, such as *Listeria monocytogenes*, has become a public health priority. These organisms are responsible for a relatively rare infection and have the capability of invading host cells and cell-to-cell movement. Due to the ubiquitous presence of *Listeria* in the environment and the high mortality rate of about 25%, listeriosis is an important foodborne illness. Up to now complete genome sequences were only publicly available for *Listeria monocytogenes*, *Listeria innocua* and *Listeria welshimeri*. The rapid increase in sequencing capabilities with the advent of next generation sequencing systems has permitted us to undertake the task of generating the complete genome sequence of multiple genomes of different *Listeria* species, most of which have not been previously sequenced, and to do complete genome comparisons of the obtained sequences. Here we report the genome sequencing and genomic comparison of 8 different *Listeria* strains representing 5 species; two *Listeria monocytogenes* strains, two *Listeria innocua* strains (one hemolytic and one non-hemolytic), two *Listeria seeligeri* strains (one hemolytic and one non-hemolytic), one *Listeria ivanovii* strain (subspecies *londoniensis*) and one strain of a newly discovered *Listeria*-species. The benefit of generating and comparing the genome sequences of all known *Listeria* species is two-fold. First, the elucidation of genomic differences in the various *Listeria* species will aid in the development of rapid molecular detection systems such as genus-specific and species-specific real-time PCR assays. Second, access to genome sequences for the different species will give insight into the mechanisms behind the gain or loss of pathogenicity in the diverse *Listeria* species.

Introduction

The genus *Listeria* is composed of seven species; *L. monocytogenes*, *L. innocua*, *L. welshimeri*, *L. seeligeri*, *L. ivanovii*, *L. grayi* and a recently discovered new species, *L. marthii* nom. prov.. Two species, *L. monocytogenes* and *Listeria ivanovii* are known pathogens and cause listeriosis in human and warm-blooded hosts. These pathogenic species are closely related to non-pathogenic species: *L. monocytogenes* is closely related to *L. innocua* and *L. marthii*, and *L. ivanovii* is closely related to *L. seeligeri*. Two main groups of genomic elements are associated with the pathogenic potential of *Listeria* strains; (1) the presence of the *prfA* virulence cluster, also known as the *Listeria* pathogenicity island (LiPI) and (2) the members of the internalin gene family. The *prfA* cluster contains genes that are necessary for inter-and intracellular survival and motility when in the host, while internalins are essential for host invasion through internalization.

The advent of next generation sequencing systems has made it possible to sequence multiple bacterial genomes to extremely high coverage within a short amount of time. Here we present data obtained with the Applied Biosystems SOLiD™ system, a high throughput sequencing system capable of producing over 400 million short reads per run. When sequencing eight *Listeria* genomes, each with an average genome size of 3 Mbp, in a single run with 25 bp reads, this throughput translates into a projected coverage of over 200X per genome. Here we report the genome sequencing and genomic comparison of 8 *Listeria* strains. These strains were selected to represent species that have currently no publicly available genome or represent atypical strains of previously sequenced species (e.g. hemolytic strains of normally non-hemolytic species or vice versa). The objective of this effort was twofold: (1) to obtain sequence data to elucidate the evolution of pathogenicity in *Listeria* and (2) to use this sequencing effort as a test case for the 'de novo' assembly of bacterial genomes with short reads.

Materials and Methods

Isolate/Strains sequenced:

Control strain:*L. monocytogenes* 'lineage I' F2365 FSL R2-574: This is the same strain that has been sequenced under the name F2365 at TIGR. This isolate originates from food that was involved in the Mexican Style Cheese outbreak in Los Angeles in 1985. This strain is sequenced to serve as a control for sequence and assembly quality.

***1.L. marthii* nom. prov. isolate FSL S4-120**: This isolate has been designated as the type strain of *L. marthii*. *L. marthii* has recently been discovered in several natural areas in the Finger Lake region in NY. It is non-pathogenic and sequence data show that the pathogenicity island is completely missing in this species.

2. *L. innocua* (hemolytic) isolate FSL J1-023: *L. innocua* is a non-pathogenic species and the fast majority of strains lack genes that are involved in pathogenicity. This strain is exceptional in that it contains the pathogenicity island (Johnson et al., 2004) and a homologue of inlA (Volokov et al., 2007). The genome sequence of this strain will help to understand the role of horizontal gene transfer and recombination in the evolution of pathogenicity in *Listeria*.

3. *L. innocua* (non-hemolytic) isolate FSL S4-378: Preliminary analyses of MLST data for *L. innocua* suggest that this species shows a high frequency of intra- and interspecific recombination. This strain is distantly related to the already published *L. innocua* genome and the genome sequence of this strain will help us understand the overall importance of homologues recombination in the evolution of *L. innocua* and *L. monocytogenes*.

4. ***L. seeligeri* (hemolytic) isolate FSL N1-067:** Though *L. seeligeri* is not considered a pathogen, it does contain the pathogenicity island and homologues of other genes involved in pathogenicity in *L. monocytogenes*. Genome sequence data of this species may provide insight why this species is not a pathogen but does have all the genes involved in pathogenicity.

5. *L. seeligeri* (non-hemolytic) isolate FSL S4-171: This isolate is very closely related to FSL N1-067, however it lacks some of the pathogenicity genes like hemolysin. Genome sequence data will help to understand the mechanism behind the loss of these pathogenicity genes.

6. *L. ivanovii* subsp. *londoniensis* ATCC 49954: *Listeria ivanovii* is a pathogen of ruminants and is closely related to *L. seeligeri*. This species seems to be more host-adapted than *L. monocytogenes* as it is only reported in a small number of human listeriosis cases. Two subspecies are recognized within this species; subsp. *ivanovii* and subsp. *londoniensis*. This genome sequence may help us to understand why this species is relatively host specific as compared to *L. monocytogenes*.

***7.Listeria monocytogenes* 'lineage IIIC' FSL F2-208:** *Listeria monocytogenes* can be subdivided into several evolutionary lineages; lineage I, lineage II, lineage IIIA, lineage IIIB and lineage IIIC. Lineage III isolates form a distinct clade from the lineage I and II isolates. Sequence analyses have shown that lineage III isolates are frequently involved in intra and inter specific recombination.

Genome sequencing and assembly

Genomes were sequenced using the SOLiD™ system (Applied Biosystems, Foster City). Mate-paired libraries with approximately 1.5 kb inserts were constructed and deposited on one quarter of a flowcell. Then, 25 bp reads were obtained from each of the F3 and R3 tags. Between 27 million and 57 million reads were obtained for each of the genomes. Referenced assembly was performed using the Applied Biosystems corona_lite package. After correcting errors in colorspace reads using a modified version of the spectral alignment tools from the EULER-USR package (Chaisson, et al., 2009), de novo assembly was performed using the SOLiD™ de novo pipeline, which employs the Velvet assembly engine (Zerbino & Birney, 2008).

Draft genome sequences were created using the 'move contigs' tool in Mauve (Darling et al., 2004). This algorithm uses a reference sequence to determine the most likely order of contigs based on a reference sequence. This approach is especially helpful in *Listeria* since the gene order within the core genome of *Listeria* seems to be extremely conserved. To assess the influence of the reference sequence on the order of the contigs of the draft sequences the move contig procedure was repeated with different reference sequences.

Whole genome alignments

Whole genome alignments and a Neighbor Joining tree based on 'gene content' were created using Mauve (Darling et al., 2004). The alignments contained previously published reference sequences (*L. monocytogenes* EGDe, *L. monocytogenes* F2365, *L. innocua* CLIP 11262) and the newly obtained draft genomes. Presence or absence of the *prfA* virulence cluster and internalins was assessed by comparison of reference sequences with the draft genomes.

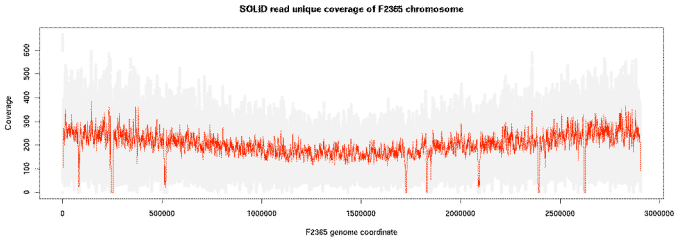

Figure 1. Unique coverage of *L. monocytogenes* F2365 chromosome with reads obtained FSL R2-574. The median coverage is 200X. Five of the uncovered gaps correspond to the six rRNA operons (A and B are adjacent, accounting for the wider gap around 250 kb). The other three regions with lower coverage presumably represent non-unique sequence in the genome. The "smiling" coverage profile (higher on the ends and lower in the middle) is due to the fact that bacteria fire multiple rounds of bidirectional replication from the origin (near position 1), leading to a gradient of copy number.

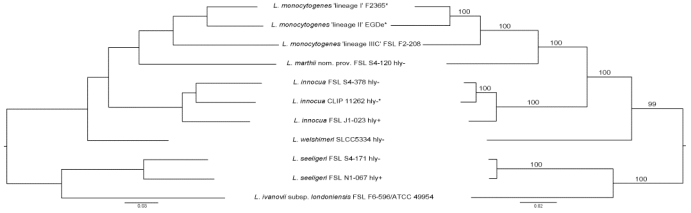

Figure 2. Neighbor Joining tree based on an estimate of the shared gene content of each pair of the input genomes (left). This tree is remarkably similar to our current understanding of the phylogeny of *Listeria* (ML tree on the right) and suggests shared gene content is phylogenetically informative within the genus *Listeria*.

Figure 3. A. Part of a whole genome alignment of *L. monocytogenes* F2365, *L. innocua* CLIP 11262, *L. innocua* FSL S4-378 and L. innocua FSL J1-023. The blue region is the prfA-cluster and is present in *L. monocytogenes* F2365 and the hemolytic *L. innocua* strain FSL J1-023. The non-hemolytic *L. innocua* strains lack the complete *prfA* cluster and only display the genes adjacent to the *prfA* cluster (*prs* and a hypothetical lipoprotein (ORF Z), marked with red arrows). **B.** Part of a whole genome alignment of the non-hemolytic *L. welshimeri* SLCC 5334, the hemolytic *L. ivanovii* subsp. *londoniensis* (FSL F6-596)and an hemolytic *L. seeligeri* strain (FSL N1-067) and a non-hemolytic *L. seeligeri* strain (FSL S4-171). In *L. ivanovii* the prfA-cluster is flanked by *prs* and ORF P (marked with red arrows). Comparison of the hemolytic versus non-hemolytic L. seeligeri strains shows that they differ in the presence or absence of the complete prfA cluster. In *L. seeligeri* the prfA cluster (blue arrows) is flanked by several ORFs of unknown function (purple blocks).

Results

Assembly

Assembly of the short reads with the *de novo* assembly pipeline resulted in 786 to 2551 contigs per genome; the sum of the length of the contigs is between 2.8 and 3.1 Mb, which is comparable to genome sizes of previously sequenced *Listeria* genomes. A summary of the *de novo* assembly can be found in **table 1.** A reference based assembly of FSL R2-578 to F2365 (figure 1) indicates a very high coverage (200X median) and 26 putative SNPs.

Shared Gene content based Neighbor Joining tree

The Neighbor joining tree based on the shared gene content of the reference genomes and the genomes sequenced is in agreement with the phylogeny of *Listeria*. The genomes of *L. seeligeri* and *L. ivanovii* are more similar to each other than to the rest of the *Listeria* species. Within the *L. welshimeri*/*L. innocua*/*L. marthii* cluster the *L. welshimeri* genome has the least shared gene content with the other species. *L. marthii* is found to be intermediate between *L. innocua* and *L. monocytogenes*.

Presence/Absence virulence associated genes

Comparison of hemolytic and non-hemolytic strains within the same species (*L. innocua* and *L. seeligeri*) reveal a remarkably similar pattern: The difference between these strains does not only involve the presence or absence of the hemolysin gene, but seems to involve the presence or absence of the complete *prfA* cluster. This may mean the *prfA* cluster is either a mobile element (a true pathogenicity island) or there is rapidly acting selection for the deletion of the rest of the genes found in the *prfA* cluster once one gene has been deleted.

Table 1. Summary contig and scaffold statistics of *Listeria de novo* assembly

		L. monocytogenes Lineage I	*L. monocytogenes* Lineage IIIC	*L. marthii*	*L. innocua* hly-	*L. innocua* hly+	*L. seeligeri* hly+	*L. seeligeri* hly-	*L. ivanovii*
		R2-574	**F2-208**	**S4-120**	**S4-378**	**J1-023**	**N1-067**	**S4-171**	**F6-596**
Contigs	**Sum length**	3,062,757	3,123,116	2,851,450	3,084,342	2,879,083	3,067,075	2,872,655	3,088,916
	Number	2,046	2,551	926	1,872	791	786	1,110	1,479
	Mean length	1,496	1,224	3,079	1,647	3,639	3,902	2,587	2,088
	Median length	563	555	1,061	629	1,434	1,096	1,218	798
	N50	3,977	2,659	7,848	4,257	9,049	10,848	5,760	5,128
	Max length	28,237	14,531	43,048	30,679	40,158	59,905	23,923	24,710
Scaffolds	**Sum length**	3,121,891	3,240,636	2,863,694	3,134,111	2,897,709	3,067,355	2,898,380	3,112,749
	Number	1,144	1,437	404	896	324	343	216	601
	Mean length	2,728	2,255	7,088	3,497	8,943	8,942	13,418	5,179
	Median length	338	371	376	326	453	258	165	351
	N50	137,174	49,992	257,992	102,515	247,625	282,765	226,677	95,455
	Max length	397,508	387,782	751,787	474,773	769,104	499,416	744,917	289,375

Conclusions

1. **Our analyses show that high quality draft genomes can be obtained through *de novo* assembly of short read sequences.**
2. **Shared gene content of *Listeria* species is phylogenetically informative.**
3. **The difference in hemolytic versus non-hemolytic strains within *L. seeligeri* and *L. innocua* can be attributed to the presence/absence of the complete *prfA* cluster.**

Acknowledgements

We would like to thank Barbara Bowen and Esther Fortes for their help with the preparation of the isolates for sequencing. This work was supported, in part, by USDA CSREES Special Research Grants 2006-34459-16952 and 2008-34459-19043 (to M. Wiedmann).

References

Chaisson MJ, Brinza D, Pevzner PA. 2009. De novo fragment assembly with short mate-paired reads: Does the read length matter? Genome Res vol. 19 (2) pp. 336-46.

Darling ACE, Mau B, Blattner FR, Perna NT. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res vol. 14 (7) pp. 1394-40.

Volokhov DV, Duperrier S, Neverov AA, George J, Buchrieser C, Hitchins AD. 2007. The presence of the internalin gene in natural atypically hemolytic *Listeria innocua* strains suggests descent from *L. monocytogenes*. Appl Environ Microbiol vol. 73 (6) pp. 1928-39.

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res vol. 18 (5) pp. 821-9.