Genetic Separation of *Listeria monocytogenes* Causing Central Nervous System Infections in Animals

Lisandra Aguilar-Bultet [1,2,3], Pamela Nicholson [1], Lorenz Rychener [1], Margaux Dreyer [2,4], Bulent Gözel [2,4], Francesco C. Origgi [1,5], Anna Oevermann [4], Joachim Frey [1]* and Laurent Falquet [3,6]*

[1] Institute of Veterinary Bacteriology, Vetsuisse Faculty, University of Bern, Bern, Switzerland, [2] Graduate School for Cellular and Biomedical Sciences, University of Bern, Bern, Switzerland, [3] BUGFri Group, Swiss Institute of Bioinformatics, Fribourg, Switzerland, [4] Division of Neurological Sciences, Department of Clinical Research and Veterinary Public Health (DCR-VPH), Vetsuisse Faculty, University of Bern, Bern, Switzerland, [5] Centre for Fish and Wildlife Health (FIWI), Vetsuisse Faculty, University of Bern, Bern, Switzerland, [6] Division of Biochemistry, Department of Biology, University of Fribourg, Fribourg, Switzerland

Listeria monocytogenes is a foodborne pathogen that causes abortion, septicemia, gastroenteritis and central nervous system (CNS) infections in ruminants and humans. *L. monocytogenes* strains mainly belong to two distinct phylogenetic groups, named lineages I and II. In general, clinical cases in humans and animals, in particular CNS infections, are caused by lineage I strains, while most of the environmental and food strains belong to lineage II. Little is known about why lineage I is more virulent than lineage II, even though various molecular factors and mechanisms associated with pathogenesis are known. In this study, we have used a variety of whole genome sequence analyses and comparative genomic tools in order to find characteristics that distinguish lineage I from lineage II strains and CNS infection strains from non-CNS strains. We analyzed 225 strains and identified single nucleotide variants between lineages I and II, as well as differences in the gene content. Using a novel approach based on Reads Per Kilobase per Million Mapped (RPKM), we identified 167 genes predominantly absent in lineage II but present in lineage I. These genes are mostly encoding for membrane-associated proteins. Additionally, we found 77 genes that are largely absent in the non-CNS associated strains, while 39 genes are especially lacking in our defined "non-clinical" group. Based on the RPKM analysis and the metadata linked to the *L. monocytogenes* strains, we identified 6 genes potentially associated with CNS cases, which include a transcriptional regulator, an ABC transporter and a non-coding RNA. Although there is not a clear separation between pathogenic and non-pathogenic strains based on phylogenetic lineages, the presence of the genes identified in our study reveals potential pathogenesis traits in ruminant *L. monocytogenes* strains. Ultimately, the differences that we have found in our study will help steer future studies in understanding the virulence mechanisms of the most pathogenic *L. monocytogenes* strains.

Keywords: *Listeria monocytogenes*, comparative genomics, virulence, lineages I and II, Reads Per Kilobase per Million Mapped, Single Nucleotide Variants, central nervous system infections, listeriosis

INTRODUCTION

Listeria monocytogenes is a rod-shaped Gram-positive bacterium that is an opportunistic food-borne pathogen (Farber and Peterkin, 1991; Vazquez-Boland et al., 2001b; Conly and Johnston, 2008). It is the etiological agent of listeriosis in humans and ruminants leading to abortion, septicemia, gastroenteritis and central nervous system (CNS) infections (Low and Donachie, 1997; Oevermann et al., 2010). Based on classical molecular subtyping methods, such as ribotyping, pulse field gel electrophoresis (PFGE) and multilocus sequence typing (MLST), *L. monocytogenes* strains are grouped into four distinct phylogenetic lineages called I, II, III and IV (Ward et al., 2008; Den Bakker et al., 2010; Orsi et al., 2011). Strains belonging to lineages I and II are the most representative in terms of number of strains isolated worldwide and impact on the disease (Chenal-Francisque et al., 2011; Orsi et al., 2011). Studies based on multilocus variable-number tandem-repeat analysis (MLVA) have revealed that lineage I strains are linked to CNS infections whereas most of the food and environmental strains belong to lineage II (Balandyte et al., 2011). Likewise, it was shown in recent studies based on MLST, that lineage I is mostly composed of clinical strains, from both ruminants (Dreyer et al., 2016) and humans (Maury et al., 2016), while lineage II typically clusters food and environmental strains. Lineage III and IV strains are very rare and mostly come from animals (Kuenne et al., 2013).

Listeria monocytogenes is a facultative anaerobic, non-spore forming, motile intracellular pathogen that can survive and reproduce under extreme conditions. It can persist ubiquitously in the environment, such as in soil, manure and grass. It is able to proliferate inside host organisms as well as in refrigerators and food processing factories (Doumith et al., 2004). Various *L. monocytogenes* virulence factors have been identified and the most important one to date is the *Listeria* pathogenicity island number 1 (LIPI-1). This is a 9 kb long region composed of six genes encoding proteins that are required for crucial steps in the intracellular life cycle of *L. monocytogenes* (Portnoy et al., 1992; Chakraborty et al., 2000; Kreft and Vazquez-Boland, 2001; Vazquez-Boland et al., 2001a). These six genes produce important virulence factors called listeriolysin O (encoded by the *hly* gene), phospholipases A and B (encoded by *plcA* and *plcB*, respectively), zinc metalloproteinase (encoded by *mpl*), actin assembly-inducing protein (encoded by *actA)* and the transcriptional activator PrfA (encoded by *prfA)*. PrfA is a 27 kDa site-specific DNA binding protein that regulates the transcription of all LIPI-1 genes (Leimeister-Wachter et al., 1990; Vazquez-Boland et al., 2001a; Scortti et al., 2007).

PrfA also regulates virulence genes not located on LIPI-1, such as the internalin genes *inlA*, *inlB*, and *inlC* (Freitag et al., 2009). Internalins are a group of surface proteins that are important for the pathogenesis of *L. monocytogenes*. A large family of internalins was identified in the EGD-e strain (Bierne et al., 2007). Internalins A and B are directly associated with the invasion of the host (Gaillard et al., 1991; Dramsi et al., 1995), while internalin C is important for cell-to-cell spread (Engelbrecht et al., 1996).

Many studies concerning the molecular mechanisms of virulence in *L. monocytogenes* have been conducted using the lineage II EGD-e strain (Glaser et al., 2001), but much less research has been performed using lineage I strains, despite the fact that they typically constitute more virulent strains. To this end, we fully sequenced and assembled a lineage I strain (JF5203) that we defined as the reference of CNS infections in ruminants in order to identify virulence genes by Whole Genome Sequencing (WGS) and a comparative genomics analysis. This strain is a rhombencephalitis isolate from sequence type 1 (ST1), has been extensively characterized (Henke et al., 2015; Dreyer et al., 2016; Rupp et al., 2017). Due to its capacity to infect bovine brain cell cultures and to spread by intra-axonal migration, we have chosen this strain as it is assumed to have the molecular factors needed for neuroinvasion.

In addition to the above mentioned multilocus-based studies, large scale WGS studies recently confirmed the distribution in the four phylogenetic lineages and the clonal population structure of *L. monocytogenes* (Kwong et al., 2016; Maury et al., 2016; Moura et al., 2016). WGS approaches have also become an important tool in the epidemiological surveillance of *L. monocytogenes* (Bergholz et al., 2015; Jackson et al., 2016; Kwong et al., 2016; Maury et al., 2016). Moreover, studies analyzing large amounts of *L. monocytogenes* genomic data has led to the identification of hypervirulent and hypovirulent groups (Dreyer et al., 2016; Maury et al., 2016) and various putative virulence factors (Maury et al., 2016). Recently, a cluster of six genes called LIPI-4 and annotated as a cellobiose-family phosphotransferase system was described in clonal complex 4 (CC4) *L. monocytogenes* strains (Maury et al., 2016). LIPI-4 revealed to have CNS invasion capacity in humanized mice. However, ST1 (CC1) is devoid of LIPI-4 (Maury et al., 2016) and constitutes a predominant group in ruminant rhombencephalitis cases (Dreyer et al., 2016). Therefore, to better understand neurovirulence in *L. monocytogenes*, the aim of our study was to elucidate the characteristics that distinguish lineage I from lineage II strains in ruminants, as well as to differentiate between clinical strains (in particular CNS infection strains) from non-disease related strains, using WGS analyses and comparative genomics tools.

MATERIALS AND METHODS

Bacterial Strains

A total of 121 strains from lineage I and 104 from lineage II were included in the study. These strains come from our internal collection and have been phenotypically characterized and used in previous studies (Balandyte et al., 2011; Rupp et al., 2015; Dreyer et al., 2016). The strains were isolated by enrichment in Oxoid Novel Enrichment Broth at 30°C and subsequently growth on Brilliance Listeria agar (Oxoid, Ltd., Basingstoke, United Kingdom) at 37°C for 24 h. Single colonies suggestive for *L. monocytogenes*, were then transferred to Tryptic Soy Agar (TSA) containing 5% (v/v) sheep blood (BD, Becton Dickinson and Company, Sparks, U.S.A.) and incubated at 37°C for another 24 h. For a few strains, colonies presenting haemolysis on the TSA were then applied to the VITEK Compact 2 phenotypic

analysis identification system, using Gram-positive identification cards (Biomerieux, Geneva, Switzerland) for the phenotypic identification of the species *Listeria monocytogenes*. For the rest of the strains, the species and lineage were defined according to Matrix-Assisted Laser Desorption Ionization-Time Of Flight Mass Spectrometry (MALDI-TOF MS) (Dreyer et al., 2016), after confirming the equivalence between the two methods. Two strains from lineage I (JF5203 and JF5861) and two strains from lineage II (JF4839 and LMNC088) were selected as internal reference strains (**Table 1**). The JF5203 strain belongs to sequence type (ST) 1 and was isolated from a rhombencephalitis case in cattle. The JF5861 strain belongs to ST4 and originated from a human CNS infection. The LMNC088 strain belongs to ST412 and came from the farm environment. The JF4839 strain belongs to ST9 and originated from food not related to any listeriosis outbreak.

Genomic DNA Extraction

Listeria monocytogenes strains that were re-sequenced in this study were grown overnight at 37°C on TSA supplemented with 5% (v/v) sheep blood (Becton Dickinson GmbH, BD™Trypticase™, PA-254053.07). Colonies were picked and directly treated with lysozyme (at a final concentration of 0.4 µg/µL). Thereafter, the bacterial cells were lysed in guanidium buffer (60% w/v) (Pitcher et al., 1989) and genomic DNA (gDNA) was extracted according to a previously published phenol-chlorofom-isoamyl alcohol protocol (Wilson, 1987).

Whole Genome Sequencing

Ninety-one of the total number of strains used in this study were sequenced in a previous study (Accession numbers PRJEB15123 and PRJEB15195; Dreyer et al., 2016). Some of them were re-sequenced to improve the data quality and coverage. The remaining 134 strains were sequenced specifically for this study (Table S1).

All strains were sequenced using the Illumina® technology (https://www.illumina.com/), either on MiSeq (300 bp paired-end reads) or HiSeq 2500/3000/4000 (95–150 bp paired-end reads) platforms, according to the manufacturer's protocols. Genome coverage varied from 19 x to more than 1000 x (Table S1). The four internal reference strains (JF5203, JF5861, JF4839, and LMNC088) were also sequenced using the Pacific Biosciences® (PacBio) technology (http://www.pacb.com/) in order to combine the high level of accuracy from the short-reads

sequencing generated by Illumina technology with the long fragments from PacBio sequencing technology.

Genome Assembly and Annotation

De novo assembly of PacBio data was done using HGAP v3.0 (Chin et al., 2013) from the SMRT® Analysis package v2.3.0. Quality control of the assembly was performed by mapping the Illumina reads to the obtained contigs and then performing an analysis with Qualimap v.2.2.1 (Okonechnikov et al., 2016). Circularization of the single contigs was carried out using the AMOS package v.3.1.0 (Treangen et al., 2011). Genomes were compared using the BRIG application v.0.95 (Alikhan et al., 2011) and ANI server (Rodriguez-R and Konstantinidis, 2016). Annotation of the whole genomes was made using Prokka v1.12 (Seemann, 2014) and MicroScope (Vallenet et al., 2013).

Variant Detection

To detect variations between the two lineages, Illumina reads of all the sequenced strains were mapped using BWA v0.7.13 (Li and Durbin, 2009) to the whole genome sequence of JF4839 (internal reference strain of lineage II). This was done to obtain the same position in all genomes relative to the same position in the reference genome. Reads with quality values below 20 in Sanger scale (Phred+33) were excluded from the analysis using sickle (https://github.com/najoshi/sickle). The total number of genomic variants including short insertions/deletions (INDELs) and single nucleotides variants (SNVs) were identified per strain using SAMtools v0.1.19 (Li et al., 2009). For variant calling files (vcf) filtering and manipulation SAMtools and vcflib (https://github.com/vcflib/vcflib) were used. Variants with mapping and assertion quality values lower than 30 in Phred-scale and with less than 20 reads supporting the alternate allele were filtered out from the vcf files in individual genomes. A Mann-Whitney-Wilcoxon test ("stats" R-package) was performed to identify differences between the numbers of SNVs per lineage (Neuhäuser, 2011). The vcf files per strain were then combined into a single merged vcf file using VCFtools v0.1.14 (Danecek et al., 2011). Manipulation and annotation of the merged vcf file was done with VCFtools, SnpEff v4.3i, (Cingolani et al., 2012) and *in house* bash scripts.

A phylogenetic tree based on the multiple alignment of the SNVs found (ignoring heterozygous sites) was built using RAxML v8.2.9 (Stamatakis, 2014), with a generalized time reversible (GTR) substitution model and using a strain from lineage III as an outgroup. Bootstrap scores (350 replicates) were calculated. The tree was re-rooted to the outgroup genome using FigTree v.1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/), edited and displayed adding metadata information with CLC Genomics Workbench v.9.5.2 (https://www.qiagenbioinformatics.com/products/clc-genomics-workbench).

The list of variants present in lineage I with respect to lineage II was filtered by excluding the SNVs with low impact according to SnpEff (http://snpeff.sourceforge.net/SnpEff_manual.html). The same approach as detailed above for read mapping and SNVs filtering was repeated with JF5203 (lineage I, CC1 and ST1) as the reference to identify SNVs only present in lineage II strains with respect to lineage I strains.

TABLE 1 | Information about the four strains selected as internal reference strains in our study.

Lineage	Strain	ST	CC	Source	Year
Lineage I	JF5203	1	1	Cattle brain	2007
	JF5861	4	4	Human brain	2006
Lineage II	JF4839	9	9	Food (cheese)	2006
	LMNC088	412	412	Environment	2014

Using an *in house* python script the variants per gene were counted in the CNS-related strains, in order to look for genes with multiple SNVs. The BED file of the annotated JF4839 genome and a filtered merged vcf file containing the variants private to CNS cases (according to SnpSift) were used as input files.

Core- Pan- Genome Analyses Based on a Reduced Set of Strains

Thirty-six published genomes with sufficient information about their lineages (18 belonging to lineage I and 18 to lineage II), along with the genomes of the four internal reference strains underwent pan-genome analyses (Table S2). Using the MicroScope platform (Vallenet et al., 2013), the core-genome of lineage I excluding the pan-genome of lineage II was calculated (at the protein level) for 80% sequence identity and 80% length coverage. As a result, the protein-coding genes shared by the 20 strains of lineage I but absent in any of the 20 strains of lineage II were predicted. A putative function was assigned to proteins with no described function according to InterPro (Apweiler et al., 2000) and BLASTp search (Altschul et al., 1990) against UniProtKB (Bairoch et al., 2005). The list was further filtered by taking into account the presence of certain amino acid motifs and domains potentially related to surface proteins and virulence factors (Bierne and Cossart, 2007).

Reads Per Kilobase Per Million Analysis

To check for the presence of the previously (section Core- pan-genome analyses based on a reduced set of strains) selected genes present in all lineage I strains of our set but absent in all lineage II strains, we used the information recorded in the bam files (all Illumina reads of each strain mapped to the JF5203 genome) to calculate the Reads Per Kilobase per Million mapped (RPKM) values. RPKM analysis is an established method for RNA sequencing (RNA-seq) data examination (Mortazavi et al., 2008; Deng et al., 2012; Tonner et al., 2012). It allows the RNA-seq gene expression quantification by normalizing for total read length and the number of sequencing reads. RPKM values were obtained according to the following equation:

$$RPKM = \frac{numReads}{\frac{geneLength}{1000} * \frac{totalNumReads}{1000000}}$$

numReads is the number of reads mapped to a gene sequence, *geneLength* is the length of the gene sequence (in bases) and *totalNumReads* is the total number of reads mapped to the genome.

Two housekeeping genes, *dnaA* and *gyrB* genes were used as controls. The Mann-Whitney-Wilcoxon test ("stats" R-package) was used to check for differences between the RPKM values for each of the previously identified genes in the two lineages (Neuhäuser, 2011).

To corroborate that there were differences between the RPKM values only in the selected genes with respect to the controls, a pairwise comparison between elements of different lineages by calculating the RPKM-difference values was performed (subtracting each RPKM value of lineage II to each RPKM value of lineage I). The post hoc Dunn's test for the Kruskal-Wallis multiple comparison test was performed with "dunn.test" R-package v.1.3.4 to assess the significance among the groups (Dinno, 2017).

RPKM Analysis at Whole Genome Level

RPKM values were calculated in the 121 sequenced lineage I genomes and in the 104 sequenced lineage II genomes using the 2981 genes of the annotated JF5203 genome as references. Thereafter a similar procedure as described in section Reads Per Kilobase per Million analysis was performed to calculate the RPKM-difference values amongst the different lineages for all genes. The median of the RPKM-difference per gene was calculated and genes with a median greater than or equal to 233.77 (2*standard deviation) were retained to obtain a list of genes predominantly absent in lineage II. A matrix was built based on the RPKM values of the selected genes. A heatmap was calculated and plotted in R ("gplots" R-package) (Warnes et al., 2016) using hierarchical clustering algorithms (average linkage clustering) based on Euclidean distance.

The same RPKM approach described above for the 2981 genes in the genome was used for comparing the CNS related strains with the non-CNS infection associated ones. Genes with a median greater than or equal to 103.38 were kept. These genes are predominantly absent in food, environmental and non-neurolisteriosis strains.

Finally, the method was applied to compare strains from the clinical group (AD) with the ones present in the non-clinical group BC (Section RPKM analyses at whole genome level and PCA analysis). Genes with a median greater than or equal to 154.01 were kept.

After every comparative analysis, the selected genes underwent a Gene Ontology (GO) enrichment step by Blast2GO (Conesa et al., 2005) and Interproscan v.5.2 (Jones et al., 2014).

Principal Component Analysis

A principal component analysis (PCA) of the RPKM values for the 2981 genes in the 225 strains was performed ("stats" R-package). A permutation multivariate analysis of variance (PERMANOVA) test (Anderson, 2001) was used to identify significant differences between the different clusters ("vegan" R-package) (Oksanen et al., 2017).

Statistical Analyses

All the statistical analyses were done in R v3.3.2 (R Development Core Team, 2016) and all *p*-values < 0.0001 were considered as significant.

Accession Numbers

The four reference genomes with their annotations and the sequencing data for all the strains used in this study were submitted to the European Nucleotide Archive (ENA) under the Project number PRJEB22706 (See Tables S1, S3 for details).

RESULTS

Obtaining the Full Genome Sequence of the Internal Reference Strains

Our first aim was to use WGS and comparative genomics to find characteristics distinguishing *L. monocytogenes* lineage I strains from lineage II strains. To this end, we selected two strains belonging to lineage I (JF5203 and JF5861) and two strains of lineage II (JF4839 and LMNC088) (**Table 1**) and obtained their entire genome sequences using PacBio and Illumina sequencing technologies. The one-contig assembly for each of the chosen reference strains was obtained with PacBio data and the quality control of the assemblies was performed using the short reads from Illumina. Only very few bases (less than 30) needed to be corrected. The sequences were further circularized to generate whole non-fragmented circular chromosomes. Each reference genome had a chromosome size of approximately 2.9 Mb and GC content of 38% which is in line with previously published *Listeria* genomes (Hain et al., 2006).

Additionally, episomal sequences were generated for strains JF5203 and JF4839. In strain JF5203, three low coverage contigs, named 1, 2, and 3, belonging to phages were sequenced concomitantly with the bacterial genome. In the samples sequenced by PacBio the coverage of these regions was 2.4–4.7 times below the genome coverage while, they were barely detectable in the samples processed by Illumina sequencing (Image S1). The phages identified by PHASTER (Arndt et al., 2016) are the following: in contig_1 and contig_2, two intact prophages LP-030-3 (GenBank accession number NC_024384.1) and vB_LmoS_293 (GenBank accession number NC_028929.1), respectively, while contig_3 contained an incomplete prophage. They are bacteriophages of the *Siphoviridae* family Orthocluster IV which have already been described as *L. monocytogenes* phages (Denes et al., 2014; Casey et al., 2015). Phages from this cluster are typically between 38 and 41 kb long and have GC contents of 35.5–36.6%. Indeed, the phages identified here are approximately 36, 33, and 3.5 kb in contig_1, contig_2, and contig_3, respectively and each have a GC content of between 35.3 and 37.1%. Most likely contig_2 and contig_3 are part of the same phage according to a Mauve comparison (Darling et al., 2004; data not shown). Therefore, the phages identified in JF5203 have approximately the same size as those previously reported for this cluster, as well as a similar GC content. Additional examinations to verify the contiguity of contigs_2 and 3 were not performed because it was not possible to re-isolate or re-identify the phages. Additionally, an incomplete prophage is also integrated into the chromosome of JF5203.

A 74 kbp plasmid was found in the lineage II strain JF4839, isolated from cheese unrelated to a listeriosis outbreak (Filiousis et al., 2009). The plasmid contains genes associated with metal transport and resistance to cadmium and camphor (Image S2). Cadmium is an important environmental pollutant and a potent toxicant to bacteria (Trevors et al., 1986). The metal transport and resistance genes are common in environmental strains to allow them to better adapt to the different environmental conditions. This plasmid shows 99% identity at the DNA level to the *L. monocytogenes* strain N1-011A plasmid (GenBank accession number NC_022045.1), representing approximately 79% of its length.

Genome Comparisons and Annotations

When using the ANI server (http://enve-omics.ce.gatech.edu/ani/) and Nucmer (Kurtz et al., 2004) to compare the reference strains from lineage I (JF5203 and JF5861) to the previously published strain F2365 from a listeriosis outbreak in California (Mascola et al., 1988), we found more than 99.6% identity at the DNA level (Image S3). On the other hand, a comparison between the reference strains from lineage II (JF4839 and LMNC088) to the EGD-e strain, showed more variation, resulting in 99% sequence identity (Image S4). The differences between the reference strains of lineage I and II were much larger representing 5.7% (Image S5).

The numeric summary of results of the annotation step using Prokka, as well as basic metrics of the genomes obtained are detailed in the **Table 2**. A similar number of internalin-like proteins were identified in all four reference sequences. Likewise,

TABLE 2 | Genes specific to lineage I selected after the RPKM comparison between the two lineages.

Locus tag	Blast2GO Description	Length	Additional annotation
LMJF5203_00387	GNAT family acetyltraansferase	136	
LMJF5203_00388	internalin	589	
LMJF5203_00428	cell surface	407	
LMJF5203_00429	family transcriptional regulator	216	
LMJF5203_00430	macrolide transporter subunit	208	
LMJF5203_00431	macrolide ABC transporter ATP-binding	224	
LMJF5203_00432	ABC transporter permease	392	
LMJF5203_00688	Uncharacterized	223	integral component of membrane
LMJF5203_00689	Uncharacterized	61	integral component of membrane
LMJF5203_00713	cell surface	824	
LMJF5203_00714	DNA-binding	217	
LMJF5203_00715	cell surface	538	
LMJF5203_01290	Uncharacterized	120	Immunity protein 51
LMJF5203_01291	cell surface	1229	
LMJF5203_01730	family transcriptional regulator	55	
LMJF5203_01731	permease	482	
LMJF5203_01732	N-acetylmuramic acid 6-phosphate etherase	296	
LMJF5203_01733	family transcriptional regulator	283	
LMJF5203_02058	cell surface	2003	
LMJF5203_02147	family transcriptional regulator	197	
LMJF5203_02312	cell surface	1529	
LMJF5203_02537	Leucine Rich repeats (2 copies)	353	
LMJF5203_02767	cell wall anchor	489	

we examined the integrity and synteny of the LIPI-1 island in the four reference genomes and found a preserved co-localization and order of all the genes in this region (Image S6).

Variant Calling at Whole Genome Level and Phylogenetic Relationship Determination

The Illumina sequencing data was used for the detection of variants that can distinguish between lineages I and II. Using the genome sequence of lineage II strain JF4839 as a reference to map all the strains in the study, the average number of SNVs in lineage II was 26,826 while in lineage I this value was 129,632 SNVs. The distribution of all SNVs in each lineage was represented in a kernel density plot (**Figure 1**, Table S4), showing more heterogeneity within lineage II (see the wider x-axis range). Significant differences with a *p*-value < 0.0001 were obtained between the two groups (Mann-Whitney-Wilcoxon test). The distribution of the total genomic variants was very similar to the distribution of the SNVs (results not shown).

A tree based on the number of SNVs at the whole genome level using JF4839 (lineage II) as reference is shown in **Figure 2**. A more distant strain, LMNC318 from lineage III, was used as an outgroup. In the resulting tree, three main branches are observed, corresponding to the different lineages. The topology of the tree confirms the clustering based on CC and ST classifications; interestingly a single branch of two strains corresponding to ST91 (CC14) are not grouped with other CC14 strains (see asterisks in **Figure 2**).

In lineage I, there are more clinical cases associated with CNS infections (75 compared to 30 cases in lineage II). One septicemia cases is also present in this group. In contrast, in lineage II, environmental strains are more common along with two strains of food origin. Other clinical manifestations such as mastitis, gastroenteritis and abortion are present in both groups. It is also relevant that lineage I has far less SNVs among them than lineage II, showing a closer distribution, while lineage II displays more diversity.

The total list of conserved variants in lineage I (present in ≥80% of the strains taking the JF4839 annotated genome of lineage II as a reference) allows the differentiation between the two lineages and all are documented in Supplementary Material DS1.

Addressing our next aim, the SNVs private to CNS infection cases were analyzed separately in order to find a pattern specific for strains of CNS origin. We looked for genes with high number of SNVs in this fraction of strains. However, we did not identify specific genes correlating to more variations in the neurolisteriosis strains. The maximum percentage of strains having common SNVs was approximately 20% (Supplementary Material DS2). Given this low percentage we think that the SNVs identified in our study are not related to neurovirulence in ruminants.

The SNVs analysis (calculation of number of SNVs per strain) was done in both directions because both lineages contain elements or genes not present in the other one. Thus, taking the JF5203 annotated genome (lineage I) as a reference, the SNVs were identified for lineage II strains. A list of variants of lineage II present in ≥80% of the strains with respect to lineage I was also created (Supplementary Material DS3).

Differences in the Thermosensor Region between Lineages I and II

Our whole genome variant analysis revealed variations in the *prfA* gene between lineage I and II. The PrfA protein is a master regulator essential for the activation of the transcription of many bacterial virulence factors within infected host cells. Specifically, we identified the presence of two substitutions of cytosine (C) in lineage II to thymine (T) in lineage I at positions 10 and 13 in the 5′UTR of the *prfA* gene in all of the 121 lineage I strains analyzed (Table S5). This specific untranslated region acts like a thermosensor in *L. monocytogenes* (Johansson et al., 2002).

Another pair of variants between the two lineages was found in the S-adenosylmethionine (SAM) riboswitch SreA. An Adenine (A) at position 83 in lineage II is substituted by a guanine (G) in lineage I, and a G at position 88 in lineage II is changed to an A in lineage I (Table S5). This SAM riboswitch participates in the negative regulation of PrfA translation, since it can bind and make and hybrid structure with the *prfA* transcript (Loh et al., 2009).
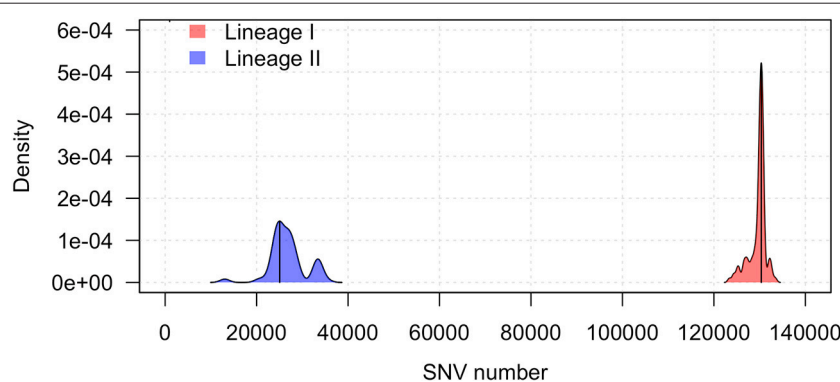
FIGURE 1 | Kernel density plot of the SNV distribution per lineage taking the JF4839 strain from lineage II as a reference. Black lines represent the mode of the data. Differences between lineages were significant (*p*-value < 0.0001, Mann-Whitney-Wilcoxon test).
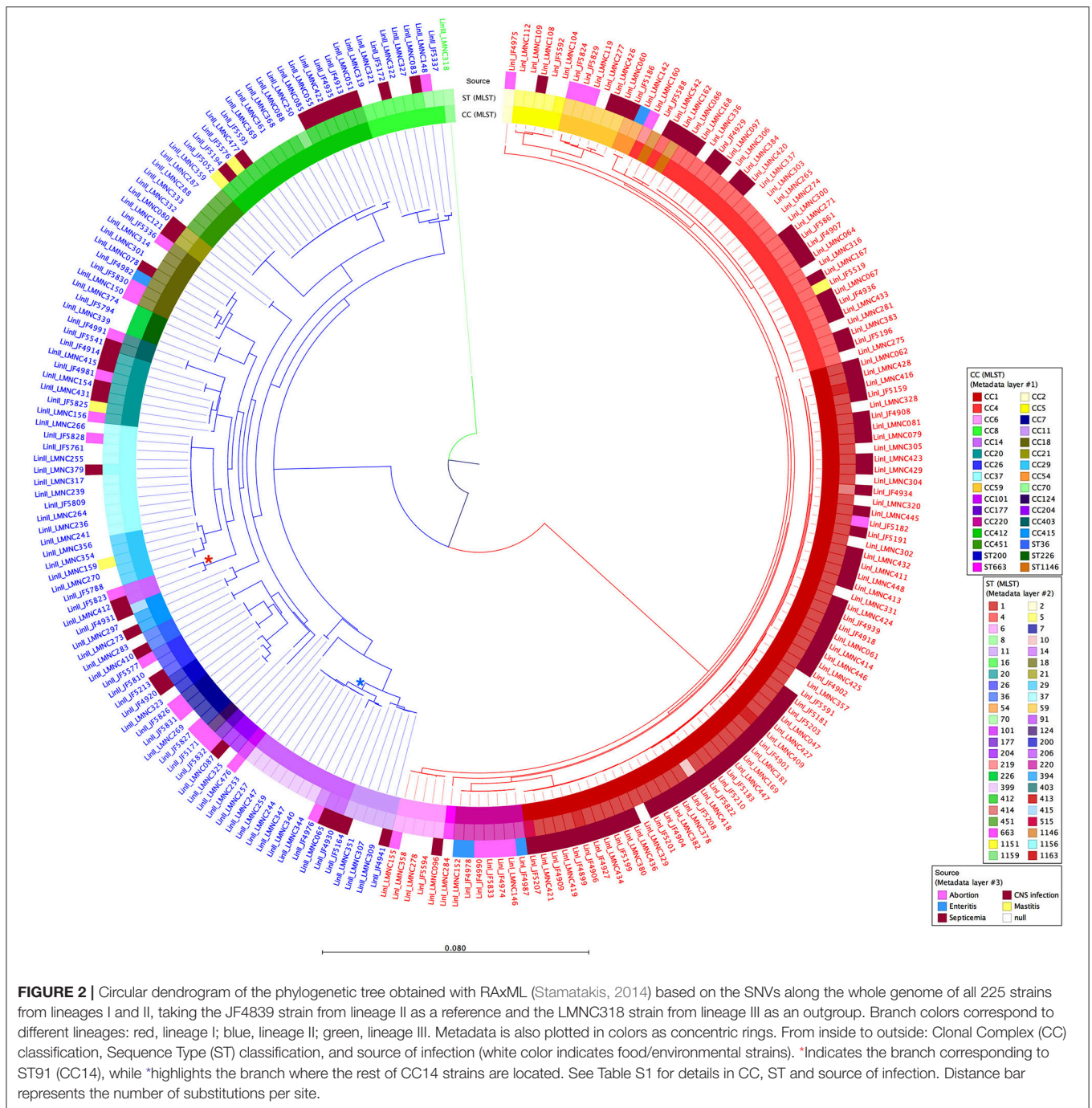
FIGURE 2 | Circular dendrogram of the phylogenetic tree obtained with RAxML (Stamatakis, 2014) based on the SNVs along the whole genome of all 225 strains from lineages I and II, taking the JF4839 strain from lineage II as a reference and the LMNC318 strain from lineage III as an outgroup. Branch colors correspond to different lineages: red, lineage I; blue, lineage II; green, lineage III. Metadata is also plotted in colors as concentric rings. From inside to outside: Clonal Complex (CC) classification, Sequence Type (ST) classification, and source of infection (white color indicates food/environmental strains). *Indicates the branch corresponding to ST91 (CC14), while *highlights the branch where the rest of CC14 strains are located. See Table S1 for details in CC, ST and source of infection. Distance bar represents the number of substitutions per site.

These two pairs of variants are given as interesting examples, however there are thousands of other potentially interesting variants to look at.

Differential Core-Genome Analysis between Lineage I and Lineage II

Genetic differences between the lineages I and II are not only due to single nucleotide differences, but also due a different gene composition. Therefore, in order to identify a group of genes shared by lineage I strains but not present in lineage

II, *L. monocytogenes* genomes (20 of each lineage including our references strains) were analyzed in the MicroScope platform (http://www.genoscope.cns.fr/agc/microscope/home/). The lineage I differential pan-genome comprised a total of 5,730 genes (2,838 families according to MicroScope MICFAM parameters 80% amino acid identity and 80% alignment coverage) including the lineage I differential variable-genome of 5,290 genes (2,816 families) and the lineage I differential core-genome of 440 genes. This core-genome corresponds to 22 gene families specific to lineage I (Image S7). From these

22, a further filtering was done, taking into account the gene length (longer than 90 bp, according to Prodigal Hyatt et al., 2010) and the presence of certain domains/motifs (internalin-like domains, cell wall anchor protein domains, adhesion domains such as LPXTG). According to these criteria, we identified a reduced group of six genes specific for lineage I strains (LMOF 2365_RS01905, LMOF2365_RS06250, LMOF2365_RS12245, LM OF2365_RS11140, LMOF2365_RS03470, LMOF2365_RS13380), which have the potential to be putative virulence attributes of *L. monocytogenes* (**Figure 3**).

In order to confirm the absence of these six selected genes in all the strains of lineage II and their exclusive presence in lineage I, we examined the remaining 221 sequenced strains. Specifically, RPKM values were calculated and compared between lineages (**Figure 4A**). Significant differences (p-value < 0.0001) were found between the lineages (Mann-Whitney-Wilcoxon test). All six genes were absent in lineage II and present in all lineage I strains analyzed, with the exception of the LMNC284 strain, where the gene LMOF2365_RS06250 is not present after verification with the Integrative Genomics Viewer (IGV) (Robinson et al., 2011; Thorvaldsdottir et al., 2013) (Image S8).
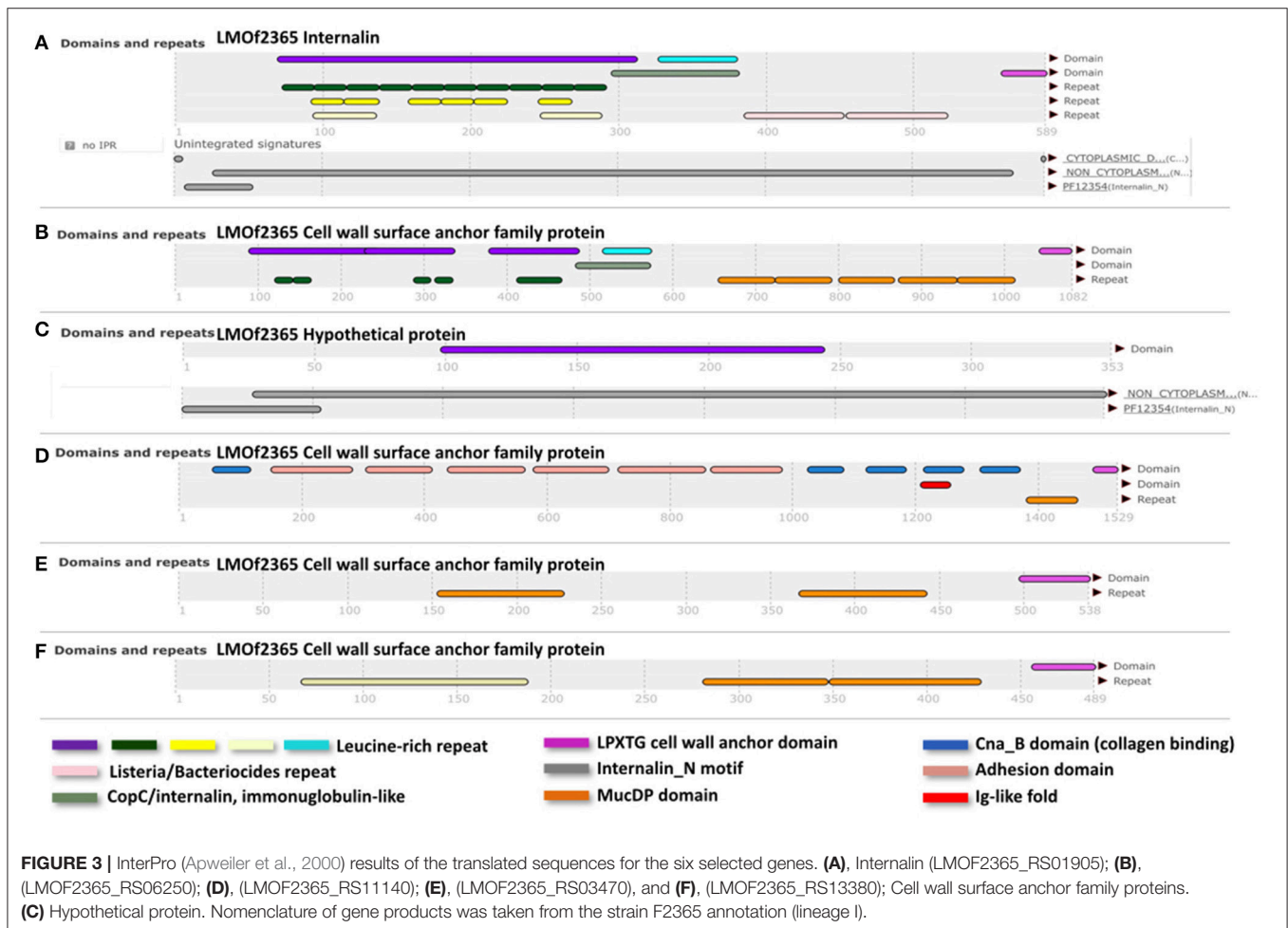
After RPKM pairwise analysis, box plots of the difference between lineage I and II gene by gene were generated

(**Figure 4B**). For the control genes (*dnaA* and *gyrB*), the difference between the two lineages is close to zero whereas for the six selected genes, this difference increases to approximately 400, showing similar behavior for all the genes. A Dunn's test revealed significant differences (p-value < 0.0001) between the control group and the six selected genes.
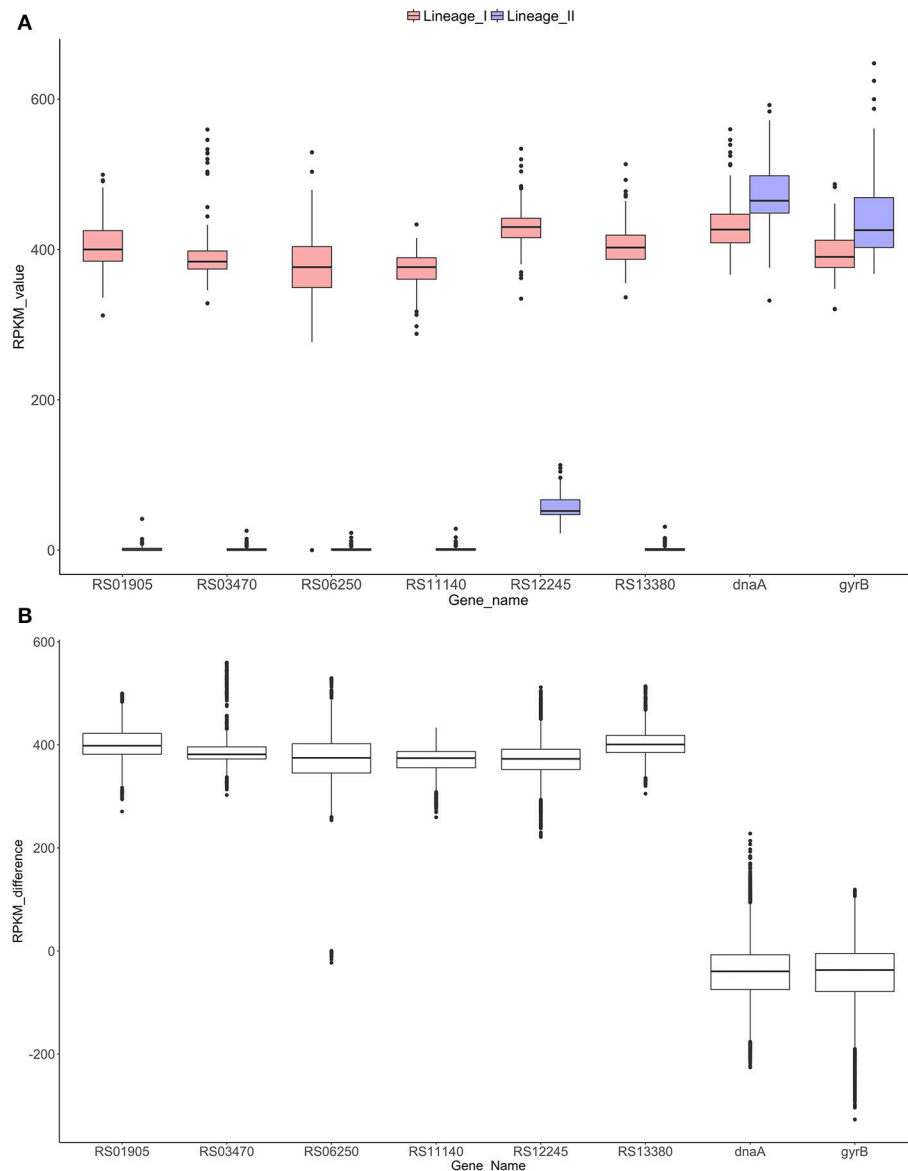
RPKM Analyses at Whole Genome Level and PCA Analysis

The RPKM values-based analysis was extended to a whole genome level in order to explore the presence of the 2981 genes from the reference genome JF5203 in the 225 genomes of this study.

A heatmap showing the degree of presence/absence of the 167 genes predominantly absent in lineage II was created for the 225 strains (**Figure 5**, Table S6). For the gene filtering, a cut-off of two times the standard deviation of the RPKM-difference values was selected (Section RPKM analysis at whole genome level). In **Figure 5**, the heatmap shows that the two lineages are perfectly separated based on the genes selected and that the strains generally grouped according to the CC classification, except for the same two strains of ST91 (CC14) already mentioned in section Variant calling at whole genome level and phylogenetic


FIGURE 3 | InterPro (Apweiler et al., 2000) results of the translated sequences for the six selected genes. **(A)**, Internalin (LMOF2365_RS01905); **(B)**, (LMOF2365_RS06250); **(D)**, (LMOF2365_RS11140); **(E)**, (LMOF2365_RS03470), and **(F)**, (LMOF2365_RS13380); Cell wall surface anchor family proteins. **(C)** Hypothetical protein. Nomenclature of gene products was taken from the strain F2365 annotation (lineage I).

FIGURE 4 | RPKM analyses to check the absence/presence of the selected genes in both lineages **(A)**. Distribution of the RPKM values in both lineages for the six selected genes and the control genes *dnaA* and *gyrB*. All the differences were significant for a *p*-value of 0.0001 (Mann-Whitney-Wilcoxon test). Gene RS12245 is affected by an artifact when counting reads in lineage II due to border effect (data not shown). **(B)** Distribution of the RPKM-differences between lineages I and II for each gene. Significant differences were found among the control groups and the six genes (*p*-value < 0.0001, Dunn's test).

relationship determination (Data not shown in the **Figure 5**; refer to Table S1 for CC/ST classification). Furthermore, it was possible to distinguish 28 genes that are only present in lineage I (**Table 2**). After Blast2GO analysis, only 5 proteins remain uncharacterized, while the remaining 23 are proteins of interest because of their classification (internalin-like proteins, cell wall anchor proteins, transcriptional regulators, ABC transporters). Notably, the majority of them are bacterial surface proteins.

Figure 6 shows the results of a PCA analysis with the whole RPKM matrix (RPKM values of all 2981 genes in all

225 strains). PCA 1 and 2 explains the 55% of the variance. Four groups are clearly defined. On one side the lineages are perfectly separated with significant differences (*p*-value < 0.0001; PERMANOVA test). On the other side, two other groups can be distinguished, one with the majority of the clinical strains (either from lineage I or II) and the other one with the majority of the environmental strains. Significant differences were also detected with the PERMANOVA test (*p*-value < 0.0001). Based on these results, the groups were defined as follows: A-lineage I clinical, B-lineage I non-clinical, C-lineage II non-clinical, and D-lineage II clinical. Only one strain (JF5593) was not classified as either
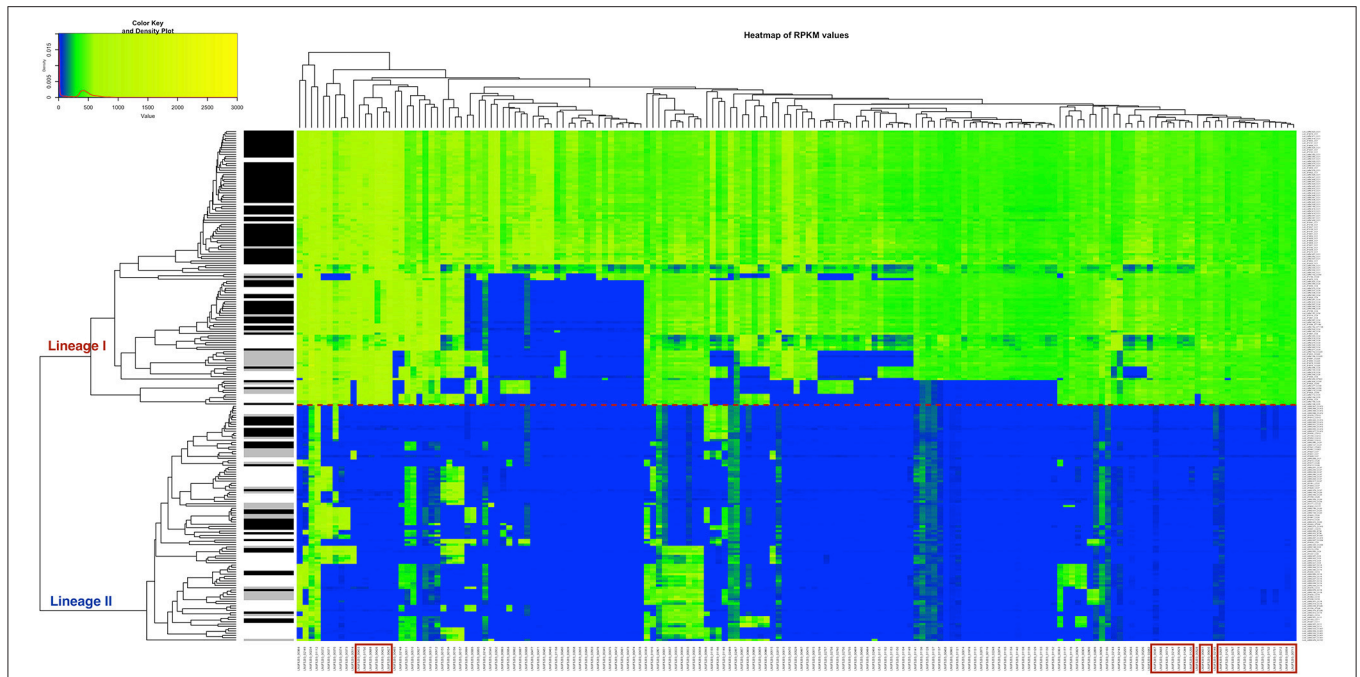
FIGURE 5 | Heatmap of the RPKM values per gene for the 167 selected genes in the 225 sequenced strains after the comparison between the two lineages. Red line: separation between the lineages I and II. Gray scale color bar corresponds with the source of infection: black, CNS infection strains and one septicemia case; gray, other clinical manifestations; white, environmental and two food strains. Red boxes denote genes that are specific to lineage I.
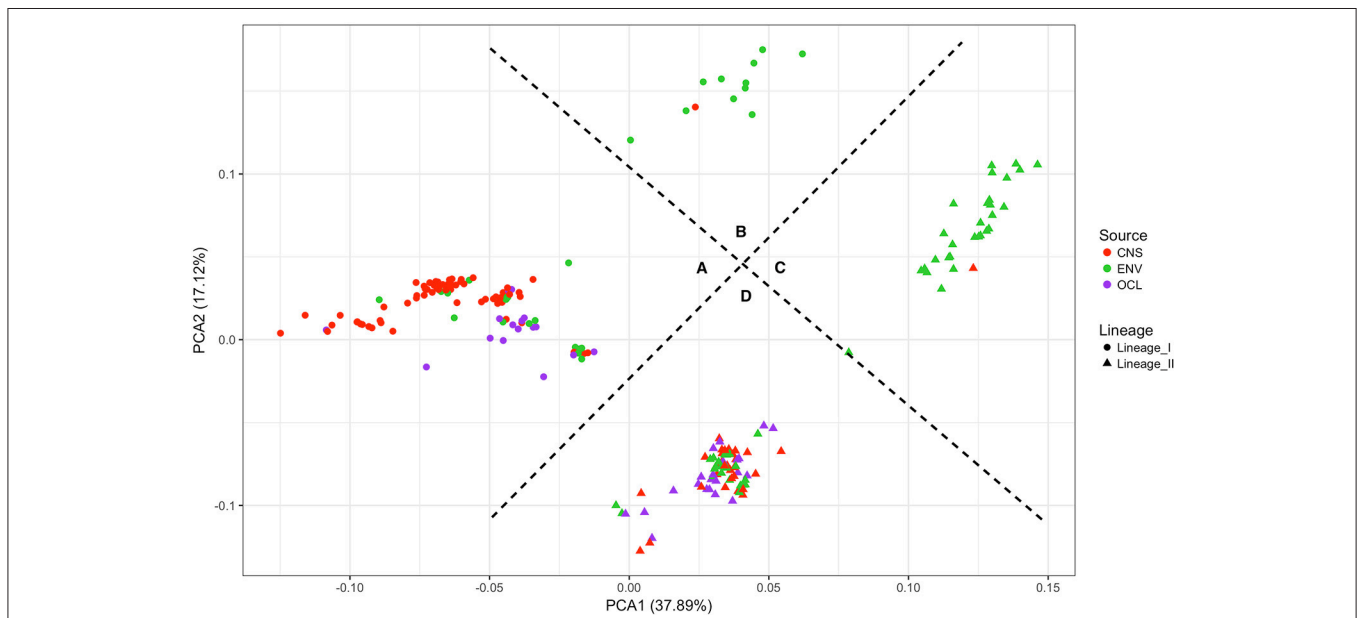
FIGURE 6 | Principal component analysis of the RPKM matrix. Four groups are defined based on the clustering. **(A)** lineage I clinical, **(B)** lineage I non-clinical, **(C)** lineage II non-clinical and **(D)** lineage II clinical.

clinical or non-clinical, since is located at the middle of the two groups.

In order to apply a clinically relevant filter to look for genes that could possibly be related to CNS infections, the RPKM method was applied, but this time to compute the differences of RPKM values between the CNS infection-related and non-related strains. We found that 77 genes are predominantly absent in the non-CNS group of the strains. According to our data,

a single gene cannot perfectly separate the two groups and be assigned as CNS-infection causative, but the combination of various genes could be a signature of neurolisteriosis (**Figure 7**, Table S7).

There are 65 mutual genes from the 167 predominantly absent in the less virulent lineage II and the 77 genes predominantly absent in the non-CNS group. Of these 65 genes, 24 encode for membrane proteins, 5 for transcriptional regulators and the rest have other functions or are hypothetical proteins (asterisks in Table S6). These genes would be of further interest since are differentially present in strains that are pathogenic and have been associated to a CNS-infections.

We decided to do a comparison between the clinical AD group vs. the non-clinical BC group because of the PCA results and considering the fact that strains originating in the environment or food does not exclude the possibility of potential pathogenicity. In this case, the combination of the resulting 39 genes seem to be essential to visibly separate clinical from the non-clinical strains (**Figure 8**, Table S8).

Combining all the methods together and looking for common genes among the three comparisons, we have compiled a list of six genes that may represent interesting targets for future *in vitro* and *in vivo* studies to evaluate their neuroinvasion potential. These genes are: LMJF5203_02482 (transcriptional regulator), LMJF5203_01155 (ncRNA rli38) (Toledo-Arana et al., 2009), LMJF5203_00294 (ABC transporter ATPase), LMJF5203_00370 (uncharacterized conserved protein) and two short membrane proteins LMJF5203_00013 and LMJF5203_00470.
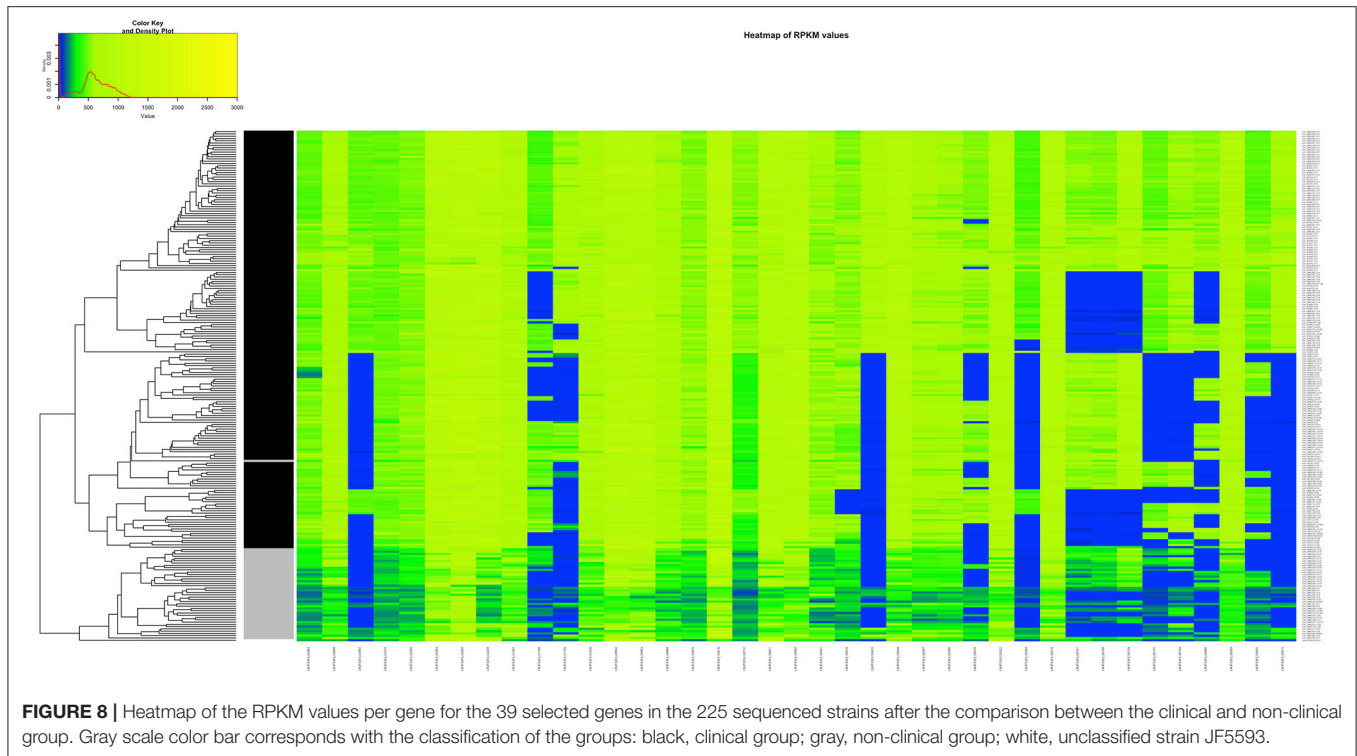
DISCUSSION

Recent studies using many *L. monocytogenes* strains have demonstrated that strains from lineage I are more frequently associated with clinical cases of listeriosis, both in animals or humans, while lineage II strains are more often found in food and environmental samples (Balandyte et al., 2011; Orsi et al., 2011; Dreyer et al., 2016; Maury et al., 2016). To date, there is little information concerning the molecular factors associated with CNS invasion. Therefore, the main aim of our study was to elucidate the characteristics distinguishing lineage I strains from lineage II and clinical strains from non-disease related strains, with a particular emphasis on CNS infection strains. To do this, we selected four internal references strains JF5203, JF5861, JF4839, and LMNC088 (**Table 1**). We used both Illumina and PacBio sequencing technologies in order to obtain complete chromosome and any additional episomal sequences.

Thereafter, we analyzed the differences between the two lineages with the goal of identifying variants and/or genes that can distinguish lineages or other phenotypic differences. The localization of these SNVs could have a potential relevance for pathogenesis. Since not all of the genes involved in virulence have been found for *L. monocytogenes*, we decided to report the list of SNVs in all genes of our reference strains: JF5203 (rhombencephalitis isolate from lineage I) and JF4839 (food isolate not related to an outbreak from lineage II). We reported the SNVs in intergenic regions and those in coding regions that have a moderate or high impact, according to SnpEff
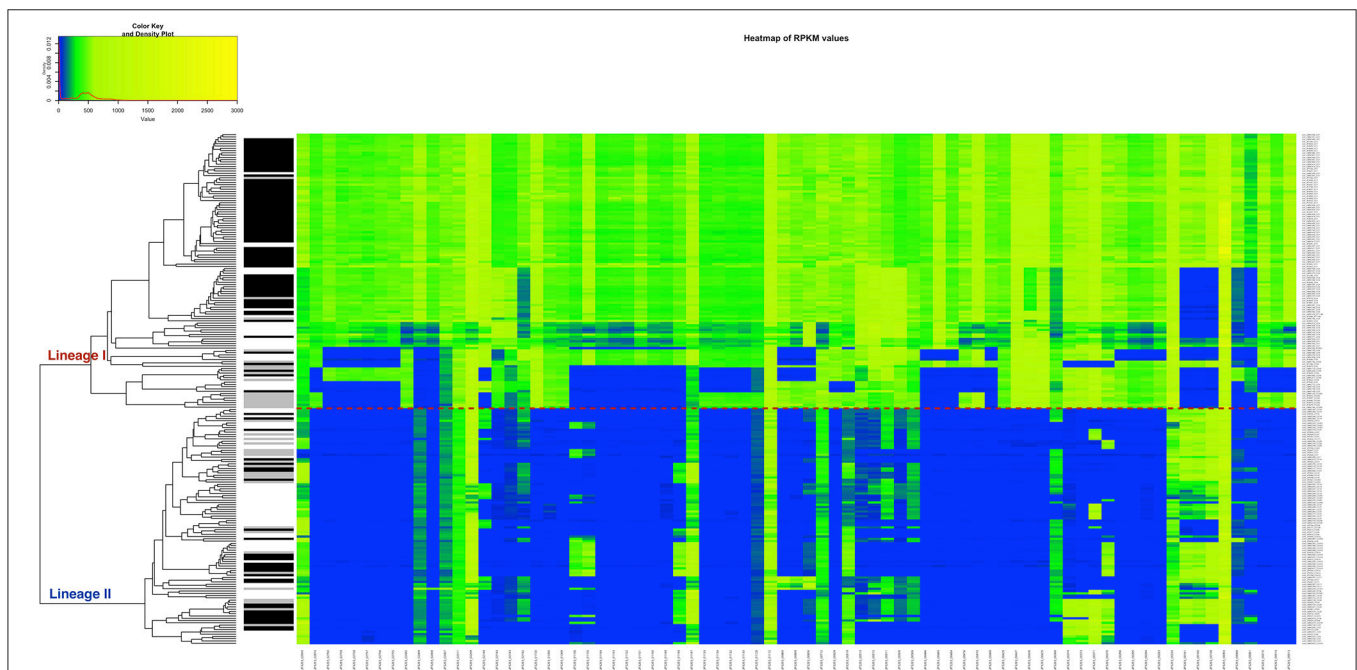
FIGURE 7 | Heatmap of the RPKM values per gene for the 77 selected genes in the 225 sequenced strains after the comparison between the CNS-infection strains and non-CNS infection strains. Red line: separation between the lineages I and II. Gray scale color bar corresponds with the source of infection: black, CNS infection isolates and one septicemia case; gray, other clinical manifestations; white, environmental and two food strains.

FIGURE 8 | Heatmap of the RPKM values per gene for the 39 selected genes in the 225 sequenced strains after the comparison between the clinical and non-clinical group. Gray scale color bar corresponds with the classification of the groups: black, clinical group; gray, non-clinical group; white, unclassified strain JF5593.

classification (Supplementary Materials DS1, DS3). Reporting the SNVs with respect to JF5203 gives a putative list of nucleotide changes that could explain pathogenesis in JF5203 (e.g., SNVs in environmental strains and not in pathogenic strains), while reporting the SNVs with respect to JF4839 offers a list of variations presumably related to loss or lack of pathogenesis in JF4839 (e.g., SNvs in pathogenic strain and not in environmental strains).

The phylogenetic tree based on multiple alignment of the SNVs along the whole genome of the reference, clearly shows three main branches corresponding to the different lineages (**Figure 2**). It also confirms that lineage I strains are more conserved while lineage II strains are more diverse. Overall, the clustering of the strains corresponds to the CC and ST scheme. This is also evident in the dendrogram showing the clustering based on RPKM values (**Figure 5**, Table S1). However, we consider that the distribution based on whole genome analysis (SNV and RPKM methods), offers more information than standard MLST classification, which is based on only 7 house-keeping genes. The possible inconsistencies between the whole genome based methods and MLST-based method may also be due to the high level of similarity of the strains, which is also indicated by the relatively low bootstrap values at many tips of the tree.

The pathogenicity island LIPI-1 is considered to be the main virulence attribute of *L. monocytogenes* and *prfA* gene is the master regulator of this island and other virulence genes (Wernars et al., 1992; Bohne et al., 1994; Scortti et al., 2007; De Las Heras et al., 2011). Temperature is important in the activation of the PrfA mRNA translation. Specifically, the 5′-untranslated

region (5′UTR) of the PrfA mRNA acts like a thermosensor. At low temperatures (30°C), it forms a secondary structure in form of a hairpin that masks the ribosome binding site (RBS) hindering its translation. While, the structure melts and opens at higher temperatures (37°). This permits access to the RBS and translation of the PrfA mRNA, which is crucial for the activation of other virulence genes (Johansson et al., 2002). In the 225 strains, two characteristic differences that separates all strains between lineage I and lineage II were found, suggesting a possible role in the regulation of the virulence genes under the control of the PrfA protein potentially via thermodynamic stability. Future work, should examine if the thermodynamic structural differences of the 5′ UTR of *prfA* between lineage I and II make lineage I strains intrinsically better prepared for a change from low (environmental) to higher temperatures (host), by activating the transcription of the downstream virulence genes, faster than lineage II strains.

The differences between the two lineages are not only due to the single point variants. Differential gene composition is another important criterion. Genes contained in lineage I but missing in lineage II are of particular interest because they represent a list of potential genes or regulatory elements that might facilitate the invasion of the bacteria ultimately producing CNS disorders. Several of these genes encode proteins with particular domains or motifs related to surface location, cell adhesion and internalin-like features.

Internalins are a family of surface proteins typically from *L. monocytogenes*. They are known virulence factors involved in the bacterial colonization and cell-to-cell spread in the host (Bierne et al., 2007). Internalins usually have different

repeats or motifs, such as internalin_N (PF12354), Leucine-rich-repeats (LRR) domains, LRR adjacent domains, Ig-like fold regions, Mucin-Binding Protein (MucBP) repeats and LPXTG motifs. A combination of some of these elements was present amongst the six selected proteins exclusive to lineage I (**Figure 3**). According to their annotation in the F2365 genome, one of the shortlisted proteins corresponds to a hypothetical protein (LMOF2365_RS12245). In this hypothetical protein, we identified an internalin_N motif and a LRR domain. Among the remaining five proteins, one is classified as an internalin (LMOF2365_RS01905) with no further details, and the other four are cell wall surface anchor proteins (LMOF2365_RS03470, LMOF2365_RS06250, LMOF2365_RS11140, and LMOF2365_RS13380). Cell wall surface anchor proteins have been reported to be important in bacterial adherence, motility and survival within the host. They are also involved in the controlled synthesis and turnover of peptidoglycan (Navarre and Schneewind, 1999). Three of these proteins (LMOF2365_RS01905, LMOF2365_RS06250 and LMOF2365_RS12245) were identified before as being specific to lineage I (Bierne et al., 2007), but taking only into account three strains of lineage I (F2365, H7858, and Clip80459) and two strains of lineage II (EGD-e and F6854).

Upon analysis of the read mapping to the six gene sequences in the 225 strains, we could show that they are absent in lineage II and present in lineage I. In general, surface proteins are very important for the pathogen, as they constitute the first point of contact of the bacteria with the host, and in many cases, an effective infection process depends on them. Hence, the six genes found in lineage I but not in lineage II strains are potential candidates to play a role in pathogenicity.

For the comparative genomics study, we successfully used the Microscope platform in a first step with 36 published genomes and our 4 references strains. MicroScope constitutes a user-friendly web-based framework with several integrated tools for analyzing individual or groups of genomes. However, this system requires the assembled genomes to be previously uploaded in the platform. The submission process can take from 4 to 10 weeks per batch and only ten genomes can be compiled per batch (http://www.genoscope.cns.fr/agc/microscope/home/).

All widely used core-pan genome methods require a preliminary assembly of the reads and allele calling (e.g., PGAP Zhao et al., 2012, Roary Page et al., 2015, BPGA Chaudhari et al., 2016, panX Ding et al., 2018, etc.). For the comparative genomic analysis of the 225 *L. monocytogenes* strains against one reference strain, we propose to use directly the raw sequencing data in a novel targeted approach based on RPKM values calculation This method is less strict than the pan- core-genome performed with MicroScope or other similar methods. The mapped reads to the selected reference can be directly used to solve questions. The generation of high quality assemblies, in some cases, can require long calculation times and is prone to errors. The method developed here is faster because it only requires remapping the reads and calculating the RPKM values. These values provide a gradual quantification of the presence

of the genes the different groups compared. We are aware of the limitations of this method; for example, we do not take into account all genes absent in our reference genome, which may include additional virulence factors potentially related to *L.monocytogenes* pathogenicity and neuroinvasion. In addition, we cannot address genes of which their absence could lead to pathogenicity.

In the RPKM analysis at the whole genome level, an array of presence/absence genes for all the strains was established with a heatmap. A set of 167 genes predominantly absent in lineage II was identified and from this list 28 genes are exclusive to lineage I (**Figure 5**). This list also includes the above-mentioned six genes identified by the core- pan-genome analysis, which constitutes a way of evaluating the effectiveness of our method.

The performance of this new approach was also evaluated by a PCA analysis, in which all the RPKM values for 2981 genes of the 225 strains were plotted (**Figure 6**). The graph evidenced that based on our data, the strains can be grouped following their lineage and also their clinical designation. Based on this graph, a new classification was assigned to the strains giving them a putative implication in pathogenesis: clinical group or non-clinical group. The disease-associated strains (clinical group) are significantly grouped together, with the exception of two small ruminant rhombencephalitis strains in the non-clinical group.

We examined some of the environmental strains that clustered together within the clinical group, and some of them are outbreak-related. Namely, the human brain strain, LMNC108, the two environmental strains LMNC104/109 and one food strain LMNC112 are most probably related to a local outbreak in Switzerland in 2005.

We have other cases in which the environmental strains were isolated in the same farm where outbreaks took place some years ago. This is the case for strain JF5591 that was found in the same farm as the clinical strain LMNC382 which was isolated one year previously. Furthermore, LMNC328/329 and 331 were isolated in a farm with reported outbreaks 5 years before, and LMNC302/304 and 305 were found in the same farm where the clinical LMNC378 strain was isolated two and a half year prior. All of this indicates the persistence of outbreak strains in ruminant farms.

Since it was not possible to clearly separate the pathogenic form the non-pathogenic strains by lineage classification, two other comparisons were performed: CNS-infection strains against non-CNS infection strains and strains belonging to the clinical group with strains form the non-clinical group (**Figures 7**, **8**). Our results suggest that not one single gene, but a gene signature might be implicated in the increased virulence of specific *L. monocytogenes* strains. This characteristic has been described for other bacteria. For example, in the different pathotypes of *Escherichia coli*, such as enteropathogenic *E. coli* (EPEC), enterotoxigenic *E. coli* (ETEC), enterohaemorrhagic *E. coli* (EHEC), enteroaggregative *E. coli* (EAEC) and enteroinvasive *E. coli* (EIEC), it has been reported that depending on the pathotype, a group of genes are responsible of the particular pathogenesis (Kaper et al.,

2004; Chapman et al., 2006). The six genes identified by RPKM in our study, that may play a role in neurolisteriosis, represent our targets for future *in vitro* and *in vivo* studies.

CONCLUSIONS

In this study, we sequenced and fully assembled two lineage I and two lineage II *L. monocytogenes* strains and used them as reference genomes. Subsequently, we sequenced 221 additional strains of both lineages and performed whole genome comparative analyses using a variety of different approaches. We identified a list of private SNVs exclusive to each lineage. While, we could not observe a typical pattern for CNS infection associated strains based on SNVs, we did find two interesting variants in the important 5′UTR of the *prfA* gene. Future studies should examine if these variants provide an adaptive advantage for the pathogenic strains of lineage I compared to the food and environmental strains of lineage II. Based on a core- pan-genome analysis of published *L. monocytogenes* genomes together with our reference genomes, we identified a set of putative virulence proteins that are present exclusively in lineage I. Subsequently, a new method based on RPKM-difference values was developed for performing a rapid comparative genomic analysis of our hundreds of strains. After comparing all lineage I strains vs. lineage II, CNS-related strains against non-CNS strains and strains belonging to the clinical group to the ones present in the non-clinical group, a common fraction of 6 genes seems to be relevant for the increased virulence of the CNS disease-related strains. In addition, our study indicates that although there is not a well-defined separation between pathogenic and non-pathogenic strains according to their phylogenetic lineages, the existence of the genes identified suggests a better indication of pathogenesis in the ruminant *L. monocytogenes* strains analyzed. This work provides an excellent basis for future studies aiming to elucidate *L. monocytogenes* virulence mechanisms.

AUTHOR CONTRIBUTIONS

LA-B performed all the bioinformatics and statistical analysis, contributed to the design of the study and wrote the article. PN participated in the design of the study and wrote the manuscript. LR assisted with the development of the scripts. MD contributed to the sample collection and gDNA purification. FO, BG and AO collaborated in the design of the study. JF and LF supervised and conceived the whole study, and wrote the manuscript. All the authors contributed to the article writing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcimb.2018.00020/full#supplementary-material

REFERENCES

Alikhan, N. F., Petty, N. K., Ben Zakour, N. L., and Beatson, S. A. (2011). BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* 12:402. doi: 10.1186/1471-2164-12-402

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/s0022-2836(05)80360-2

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Aus. Ecol.* 26, 32–46. doi: 10.1111/j.1442-9993.2001.01070.pp.x

Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., et al. (2000). InterPro–an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics* 16, 1145–1150. doi: 10.1093/bioinformatics/16.12.1145

Arndt, D., Grant, J. R., Marcu, A., Sajed, T., Pon, A., Liang, Y., et al. (2016). PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res.* 44, W16–W21. doi: 10.1093/nar/gkw387

Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., et al. (2005). The Universal Protein Resource (UniProt). *Nucleic Acids Res* 33, D154–D159. doi: 10.1093/nar/gki070

Balandyte, L., Brodard, I., Frey, J., Oevermann, A., and Abril, C. (2011). Ruminant rhombencephalitis-associated *Listeria monocytogenes* alleles linked to a multilocus variable-number tandem-repeat analysis complex. *Appl. Environ. Microbiol.* 77, 8325–8335. doi: 10.1128/aem.06507-11

Bergholz, T. M., Den Bakker, H. C., Katz, L. S., Silk, B. J., Jackson, K. A., Kucerova, Z., et al. (2015). Determination of Evolutionary Relationships of Outbreak-Associated *Listeria monocytogenes* Strains of Serotypes 1/2a and 1/2b by Whole-Genome Sequencing. *Appl. Environ. Microbiol.* 82, 928–938. doi: 10.1128/aem.02440-15

Bierne, H., and Cossart, P. (2007). Listeria monocytogenes surface proteins: from genome predictions to function. *Microbiol. Mol. Biol. Rev.* 71, 377–397. doi: 10.1128/mmbr.00039-06

Bierne, H., Sabet, C., Personnic, N., and Cossart, P. (2007). Internalins: a complex family of leucine-rich repeat-containing proteins in Listeria monocytogenes. *Microb. Infect.* 9, 1156–1166. doi: 10.1016/j.micinf.2007.05.003

Bohne, J., Sokolovic, Z., and Goebel, W. (1994). Transcriptional regulation of prfA and PrfA-regulated virulence genes in Listeria monocytogenes. *Mol. Microbiol.* 11, 1141–1150.

Casey, A., Jordan, K., Coffey, A., and Mcauliffe, O. (2015). Complete Genome Sequences of vB_LmoS_188 and vB_LmoS_293, Two Bacteriophages with Specificity for Listeria monocytogenes Strains of Serotypes 4b and 4e. *Genome Announc.* 3:e00040-15. doi: 10.1128/genomeA.00040-15

Chakraborty, T., Hain, T., and Domann, E. (2000). Genome organization and the evolution of the virulence gene locus in *Listeria* species. *Int. J. Med. Microbiol.* 290, 167–174. doi: 10.1016/s1438-4221(00)80086-7

Chapman, T. A., Wu, X. Y., Barchia, I., Bettelheim, K. A., Driesen, S., Trott, D., et al. (2006). Comparison of virulence gene profiles of *Escherichia coli* strains isolated from healthy and diarrheic swine. *Appl. Environ. Microbiol.* 72, 4782–4795. doi: 10.1128/aem.02885-05

Chaudhari, N. M., Gupta, V. K., and Dutta, C. (2016). BPGA- an ultra-fast pan-genome analysis pipeline. *Sci. Rep.* 6:24373. doi: 10.1038/srep24373

Chenal-Francisque, V., Lopez, J., Cantinelli, T., Caro, V., Tran, C., Leclercq, A., et al. (2011). Worldwide distribution of major clones of *Listeria monocytogenes*. *Emerg. Infect. Dis.* 17, 1110–1112. doi: 10.3201/eid1706.101778

Chin, C. S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., et al. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* 10, 563–569. doi: 10.1038/nmeth.2474

Cingolani, P., Platts, A., Wang Le, L., Coon, M., Nguyen, T., Wang, L., et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* 6, 80–92. doi: 10.4161/fly.19695

Conesa, A., Gotz, S., Garcia-Gomez, J. M., Terol, J., Talon, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21, 3674–3676. doi: 10.1093/bioinformatics/bti610

Conly, J., and Johnston, B. (2008). Listeria: a persistent food-borne pathogen. *Can. J. Infect. Dis. Med. Microbiol.* 19, 327–328. doi: 10.1155/2008/702565

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., Depristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330

Darling, A. C., Mau, B., Blattner, F. R., and Perna, N. T. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* 14, 1394–1403. doi: 10.1101/gr.2289704

De Las Heras, A., Cain, R. J., Bielecka, M. K., and Vazquez-Boland, J. A. (2011). Regulation of Listeria virulence: PrfA master and commander. *Curr. Opin. Microbiol.* 14, 118–127. doi: 10.1016/j.mib.2011.01.005

Den Bakker, H. C., Bundrant, B. N., Fortes, E. D., Orsi, R. H., and Wiedmann, M. (2010). A population genetics-based and phylogenetic approach to understanding the evolution of virulence in the genus *Listeria*. *Appl. Environ. Microbiol.* 76, 6085–6100. doi: 10.1128/aem.00447-10

Denes, T., Vongkamjan, K., Ackermann, H. W., Moreno Switt, A. I., Wiedmann, M., and Den Bakker, H. C. (2014). Comparative genomic and morphological analyses of Listeria phages isolated from farm environments. *Appl. Environ. Microbiol.* 80, 4616–4625. doi: 10.1128/aem.00720-14

Deng, Y., Yao, J., Wang, X., Guo, H., and Duan, D. (2012). Transcriptome sequencing and comparative analysis of Saccharina japonica (Laminariales, Phaeophyceae) under blue light induction. *PLoS ONE* 7:e39704. doi: 10.1371/journal.pone.0039704

Ding, W., Baumdicker, F., and Neher, R. A. (2018). panX: pan-genome analysis and exploration. *Nucleic Acids Res.* 46:e5. doi: 10.1093/nar/gkx977

Dinno, A. (2017). *Dunn.test: Dunn's Test of Multiple Comparisons Using Rank Sums*. R package version 1.3.4. Available online: https://CRAN-R-project.org/package=dunn.test

Doumith, M., Cazalet, C., Simoes, N., Frangeul, L., Jacquet, C., Kunst, F., et al. (2004). New aspects regarding evolution and virulence of Listeria monocytogenes revealed by comparative genomics and DNA arrays. *Infect. Immun.* 72, 1072–1083. doi: 10.1128/IAI.72.2.1072-1083.2004

Dramsi, S., Biswas, I., Maguin, E., Braun, L., Mastroeni, P., and Cossart, P. (1995). Entry of Listeria monocytogenes into hepatocytes requires expression of inIB, a surface protein of the internalin multigene family. *Mol. Microbiol.* 16, 251–261.

Dreyer, M., Aguilar-Bultet, L., Rupp, S., Guldimann, C., Stephan, R., Schock, A., et al. (2016). *Listeria monocytogenes* sequence type 1 is predominant in ruminant rhombencephalitis. *Sci. Rep.* 6:36419. doi: 10.1038/srep36419

Engelbrecht, F., Chun, S. K., Ochs, C., Hess, J., Lottspeich, F., Goebel, W., et al. (1996). A new PrfA-regulated gene of Listeria monocytogenes encoding a small, secreted protein which belongs to the family of internalins. *Mol. Microbiol.* 21, 823–837.

Farber, J. M., and Peterkin, P. I. (1991). Listeria monocytogenes, a food-borne pathogen. *Microbiol. Rev.* 55, 476–511.

Filiousis, G., Johansson, A., Frey, J., and Perreten, V. (2009). Prevalence, genetic diversity and antimicrobial susceptibility of *Listeria monocytogenes* isolated from open-air food markets in Greece. *Food Control* 20, 314–317. doi: 10.1016/j.foodcont.2008.05.018

Freitag, N. E., Port, G. C., and Miner, M. D. (2009). *Listeria monocytogenes* - from saprophyte to intracellular pathogen. *Nat. Rev. Microbiol.* 7, 623–628. doi: 10.1038/nrmicro2171

Gaillard, J. L., Berche, P., Frehel, C., Gouin, E., and Cossart, P. (1991). Entry of L. monocytogenes into cells is mediated by internalin, a repeat protein reminiscent of surface antigens from gram-positive cocci. *Cell* 65, 1127–1141.

Glaser, P., Frangeul, L., Buchrieser, C., Rusniok, C., Amend, A., Baquero, F., et al. (2001). Comparative genomics of *Listeria* species. *Science* 294, 849–852. doi: 10.1126/science.1063447

Hain, T., Steinweg, C., and Chakraborty, T. (2006). Comparative and functional genomics of *Listeria* spp. *J. Biotechnol.* 126, 37–51. doi: 10.1016/j.jbiotec.2006.03.047

Henke, D., Rupp, S., Gaschen, V., Stoffel, M. H., Frey, J., Vandevelde, M., et al. (2015). Listeria monocytogenes spreads within the brain by actin-based intra-axonal migration. *Infect. Immun.* 83, 2409–2419. doi: 10.1128/iai.00316-15

Hyatt, D., Chen, G. L., Locascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformat.* 11:119. doi: 10.1186/1471-2105-11-119

Jackson, B. R., Tarr, C., Strain, E., Jackson, K. A., Conrad, A., Carleton, H., et al. (2016). Implementation of Nationwide Real-time Whole-genome Sequencing to Enhance Listeriosis Outbreak Detection and Investigation. *Clin. Infect. Dis.* 63, 380–386. doi: 10.1093/cid/ciw242

Johansson, J., Mandin, P., Renzoni, A., Chiaruttini, C., Springer, M., and Cossart, P. (2002). An RNA thermosensor controls expression of virulence genes in Listeria monocytogenes. *Cell* 110, 551–561. doi: 10.1016/S0092-8674(02)00905-4

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., Mcanulla, C., et al. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30, 1236–1240. doi: 10.1093/bioinformatics/btu031

Kaper, J. B., Nataro, J. P., and Mobley, H. L. (2004). Pathogenic *Escherichia coli*. *Nat. Rev. Microbiol.* 2, 123–140. doi: 10.1038/nrmicro818

Kreft, J., and Vazquez-Boland, J. A. (2001). Regulation of virulence genes in Listeria. *Int. J. Med. Microbiol.* 291, 145–157. doi: 10.1078/1438-4221-00111

Kuenne, C., Billion, A., Mraheil, M. A., Strittmatter, A., Daniel, R., Goesmann, A., et al. (2013). Reassessment of the Listeria monocytogenes pan-genome reveals dynamic integration hotspots and mobile genetic elements as major components of the accessory genome. *BMC Genomics* 14:47. doi: 10.1186/1471-2164-14-47

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., et al. (2004). Versatile and open software for comparing large genomes. *Genome Biol* 5:R12. doi: 10.1186/gb-2004-5-2-r12

Kwong, J. C., Mercoulia, K., Tomita, T., Easton, M., Li, H. Y., Bulach, D. M., et al. (2016). Prospective whole-genome sequencing enhances national surveillance of *Listeria* monocytogenes. *J. Clin. Microbiol.* 54, 333–342. doi: 10.1128/jcm.02344-15

Leimeister-Wachter, M., Haffner, C., Domann, E., Goebel, W., and Chakraborty, T. (1990). Identification of a gene that positively regulates expression of listeriolysin, the major virulence factor of listeria monocytogenes. *Proc. Natl. Acad. Sci. U.S.A.* 87, 8336–8340.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Loh, E., Dussurget, O., Gripenland, J., Vaitkevicius, K., Tiensuu, T., Mandin, P., et al. (2009). A trans-acting riboswitch controls expression of the virulence regulator PrfA in Listeria monocytogenes. *Cell* 139, 770–779. doi: 10.1016/j.cell.2009.08.046

Low, J. C., and Donachie, W. (1997). A review of Listeria monocytogenes and listeriosis. *Vet. J.* 153, 9–29.

Mascola, L., Lieb, L., Chiu, J., Fannin, S. L., and Linnan, M. J. (1988). Listeriosis: an uncommon opportunistic infection in patients with acquired immunodeficiency syndrome. A report of five cases and a review of the literature. *Am. J. Med.* 84, 162–164.

Maury, M. M., Tsai, Y. H., Charlier, C., Touchon, M., Chenal-Francisque, V., Leclercq, A., et al. (2016). Uncovering *Listeria monocytogenes* hypervirulence by harnessing its biodiversity. *Nat. Genet.* 48, 308–313. doi: 10.1038/ng.3501

Mortazavi, A., Williams, B. A., Mccue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5, 621–628. doi: 10.1038/nmeth.1226

Moura, A., Criscuolo, A., Pouseele, H., Maury, M. M., Leclercq, A., Tarr, C., et al. (2016). Whole genome-based population biology and epidemiological surveillance of *Listeria monocytogenes*. *Nat. Microbiol.* 2:16185. doi: 10.1038/nmicrobiol.2016.185

Navarre, W. W., and Schneewind, O. (1999). Surface proteins of gram-positive bacteria and mechanisms of their targeting to the cell wall envelope. *Microbiol. Mol. Biol. Rev.* 63, 174–229.

Neuhäuser, M. (2011). "Wilcoxon–Mann–Whitney Test," in *International Encyclopedia of Statistical Science*, ed M. Lovric (Berlin; Heidelberg: Springer), 1656–1658.

Oevermann, A., Zurbriggen, A., and Vandevelde, M. (2010). Rhombencephalitis Caused by Listeria monocytogenes in Humans and Ruminants: a Zoonosis on the Rise? *Interdiscip. Perspect Infect. Dis.* 2010:632513. doi: 10.1155/2010/632513

Okonechnikov, K., Conesa, A., and Garcia-Alcalde, F. (2016). Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* 32, 292–294. doi: 10.1093/bioinformatics/btv566

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., Mcglinn, D., et al. (2017). *vegan: Community Ecology Package.* R package version 2.4-4. Available online at: https://CRAN.R-project.org/package=vegan

Orsi, R. H., Den Bakker, H. C., and Wiedmann, M. (2011). Listeria monocytogenes lineages: genomics, evolution, ecology, and phenotypic characteristics. *Int. J. Med. Microbiol.* 301, 79–96. doi: 10.1016/j.ijmm.2010.05.002

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T., et al. (2015). Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 31, 3691–3693. doi: 10.1093/bioinformatics/btv421

Pitcher, D. G., Saunders, N. A., and Owen, R. J. (1989). Rapid extraction of bacterial genomic DNA with guanidium thiocyanate. *Lett. Appl. Microbiol.* 8, 151–156. doi: 10.1111/j.1472-765X.1989.tb00262.x

Portnoy, D. A., Chakraborty, T., Goebel, W., and Cossart, P. (1992). Molecular determinants of Listeria monocytogenes pathogenesis. *Infect. Immun.* 60, 1263–1267.

R Development Core Team (2016). *R: A Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., et al. (2011). Integrative genomics viewer. *Nat. Biotechnol.* 29, 24–26. doi: 10.1038/nbt.1754

Rodriguez-R, L. M., and Konstantinidis, K. T. (2016). The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. *PeerJ Preprints* 4:e1900v1. doi: 10.7287/peerj.preprints.1900v1

Rupp, S., Aguilar-Bultet, L., Jagannathan, V., Guldimann, C., Drogemuller, C., Pfarrer, C., et al. (2015). A naturally occurring prfA truncation in a Listeria monocytogenes field strain contributes to reduced replication and cell-to-cell spread. *Vet. Microbiol.* 179, 91–101. doi: 10.1016/j.vetmic.2015.03.002

Rupp, S., Bartschi, M., Frey, J., and Oevermann, A. (2017). Hyperinvasiveness and increased intercellular spread of *Listeria* monocytogenes sequence type 1 are independent of listeriolysin S, internalin F and internalin J1. *J. Med. Microbiol.* 66, 1053–1062. doi: 10.1099/jmm.0.000529

Scortti, M., Monzo, H. J., Lacharme-Lora, L., Lewis, D. A., and Vazquez-Boland, J. A. (2007). The PrfA virulence regulon. *Microb. Infect.* 9, 1196–1207. doi: 10.1016/j.micinf.2007.05.007

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30, 2068–2069. doi: 10.1093/bioinformatics/btu153

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi: 10.1093/bioinformatics/btu033

Thorvaldsdottir, H., Robinson, J. T., and Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform.* 14, 178–192. doi: 10.1093/bib/bbs017

Toledo-Arana, A., Dussurget, O., Nikitas, G., Sesto, N., Guet-Revillet, H., Balestrino, D., et al. (2009). The Listeria transcriptional landscape from saprophytism to virulence. *Nature* 459, 950–956. doi: 10.1038/nature08080

Tonner, P., Srinivasasainagendra, V., Zhang, S., and Zhi, D. (2012). Detecting transcription of ribosomal protein pseudogenes in diverse human tissues from RNA-seq data. *BMC Genomics* 13:412. doi: 10.1186/1471-2164-13-412

Treangen, T. J., Sommer, D. D., Angly, F. E., Koren, S., and Pop, M. (2011). Next generation sequence assembly with AMOS. *Curr. Protoc. Bioinformat.* Chapter 11, Unit 11.18. doi: 10.1002/0471250953.bi1108s33

Trevors, J. T., Stratton, G. W., and Gadd, G. M. (1986). Cadmium transport, resistance, and toxicity in bacteria, algae, and fungi. *Can. J. Microbiol.* 32, 447–464.

Vallenet, D., Belda, E., Calteau, A., Cruveiller, S., Engelen, S., Lajus, A., et al. (2013). MicroScope–an integrated microbial resource for the curation and comparative analysis of genomic and metabolic data. *Nucleic Acids Res.* 41, D636–D647. doi: 10.1093/nar/gks1194

Vazquez-Boland, J. A., Dominguez-Bernal, G., Gonzalez-Zorn, B., Kreft, J., and Goebel, W. (2001a). Pathogenicity islands and virulence evolution in *Listeria. Microbes Infect.* 3, 571–584. doi: 10.1016/S1286-4579(01)01413-7

Vazquez-Boland, J. A., Kuhn, M., Berche, P., Chakraborty, T., Dominguez-Bernal, G., Goebel, W., et al. (2001b). Listeria pathogenesis and molecular virulence determinants. *Clin Microbiol Rev* 14, 584–640. doi: 10.1128/cmr.14.3.584-640.2001

Ward, T. J., Ducey, T. F., Usgaard, T., Dunn, K. A., and Bielawski, J. P. (2008). Multilocus genotyping assays for single nucleotide polymorphism-based subtyping of Listeria monocytogenes isolates. *Appl. Environ. Microbiol.* 74, 7629–7642. doi: 10.1128/aem.01127-08

Warnes, G. R., Bolker, B., Bonebakker, L., Gentleman, R., Liaw, W. H. A., Lumley, T., et al. (2016). *gplots: Various R Programming Tools for Plotting Data.* R package version 3.0.1. Available online at: https://CRAN.R-project.org/package=gplots

Wernars, K., Heuvelman, K., Notermans, S., Domann, E., Leimeister-Wachter, M., and Chakraborty, T. (1992). Suitability of the prfA gene, which encodes a regulator of virulence genes in Listeria monocytogenes, in the identification of pathogenic *Listeria* spp. *Appl. Environ. Microbiol.* 58, 765–768.

Wilson, K. (1987). "Preparation of genomic DNA from bacteria," in *Current Protocols in Molecular Biology,* eds R .B. F. M. Ausubel, R. E. Kingston, D. D. Moore, J. A. Smith, J. G. Seidman, and K. Struhl (New York, NY: John Wiley & Sons, Inc.), 2.4.1–2.4.5.

Zhao, Y., Wu, J., Yang, J., Sun, S., Xiao, J., and Yu, J. (2012). PGAP: pan-genomes analysis pipeline. *Bioinformatics* 28, 416–418. doi: 10.1093/bioinformatics/btr655