Comparative genomics of pathogenic lineages of *Listeria monocytogenes*

-

A bioinformatic approach

Inaugural-Dissertation (Cumulative thesis)
Submitted to the Faculty of Medicine
in fulfilment of the requirements for the degree of Dr. biol. hom.
of the Faculty of Medicine
of the Justus-Liebig-University Gießen

By Carsten Tobias Künne
of Westerburg

Gießen 2012

From the Institute of Medical Microbiology

Director: Prof. Dr. Trinad Chakraborty

Universitätsklinikum Gießen und Marburg GmbH

Standort Gießen

First supervisor and Committee Member: Prof. Dr. Trinad Chakraborty

Committee Member: Prof. Dr. Klaus T. Preissner

Date of Doctoral Defense: 18.06.2013

Table of Contents

1 List of publications

During the course of this thesis the author was involved with the following publications. The majority of work was employed on development and implementation of bioinformatic analyses of the underlined papers, which should be regarded as the main outcome of this thesis. Shared primary authorship is denoted by an asterisk (*).

1.1 Publication I - Complete genome sequence of *Listeria seeligeri*, a nonpathogenic member of the genus *Listeria*

Steinweg C*, Kuenne CT*, Billion A, Mraheil MA, Domann E, Ghai R, Barbuddhe SB, Kärst U, Goesmann A, Pühler A, Weisshaar B, Wehland J, Lampidis R, Kreft J, Goebel W, Chakraborty T, Hain T.

J Bacteriol. 2010 Mar. 192(5):1473-4.

1.2 Publication II - Comparative analysis of plasmids in the genus *Listeria*

Kuenne C, Voget S, Pischimarov J, Oehm S, Goesmann A, Daniel R, Hain T, Chakraborty T.

PLoS One. 2010 Sep. 5(9): pii: e12511.

1.3 Publication III - Genome-wide identification of small RNAs in the opportunistic pathogen *Enterococcus faecalis*

Shioya K, Michaux C, Kuenne C, Hain T, Verneuil N, Budin-Verneuil A, Hartsch T, Hartke A, Giard JC.

PLoS One. 2011 Sep. 6(9):e23948.

1.4 Publication IV - Complete Sequences of Plasmids from the Hemolytic-Uremic Syndrome-Associated *Escherichia coli* strain HUSEC41

Künne C*, Billion A*, Mshana SE, Schmiedel J, Domann E, Hossain H, Hain T, Imirzalioglu C, Chakraborty T.

J Bacteriol. 2012 Jan. 194(2):532-3.

1.5 Publication V - sRNAdb: A small non-coding RNA database for gram-positive bacteria

Pischimarov J*, Kuenne C*, Billion A, Hemberger J, Cemič F, Chakraborty T and Hain T

BMC Genomics. 2012 Jul. 13:384.

1.6 Publication VI - Comparative genomics and transcriptomics of lineages I, II, and III strains of *Listeria monocytogenes*

Hain T*, Ghai R*, Billion A*, Kuenne C*, Steinweg C, Izar B, Mohamed W, Mraheil M, Domann E, Schaffrath S, Kärst U, Goesmann A, Oehm S, Pühler A, Merkl R, Vorwerk S, Glaser P, Garrido P, Rusniok C, Buchrieser C, Goebel W and Chakraborty T

BMC Genomics. 2012 Apr. 13:144.

1.7 Publication VII – Dynamic integration hotspots and mobile genetic elements shape the genome structure of the species *Listeria monocytogenes*

Kuenne C, Billion A, Mraheil M, Strittmatter A, Daniel R, Goesmann A, Barbuddhe S, Hain T, Chakraborty T

Submitted to BMC Genomics (October 8, 2012).

2 Introduction

2.1 From Sanger to Next-Generation Sequencing (NGS)

The first sequencing of prokaryotic DNA was performed in 1977 following the development of the chain-termination method by Frederick Sanger [1]. The first completely resolved prokaryotic genome was that of phage phiX174 harbouring 5000 bp [2]. It took another 20 years of technological development in DNA sequencing to reach the next milestone which was the sequencing of the chromosome of *Haemophilus influenzae* Rd. bearing 1.8 Mb in 1995. This task involved 40 researchers and took one whole year [3]. Since 1995 the number of bacterial genomes sequenced has grown exponentially and at the end of 2011 the complete genomes of over 1700 bacterial species were available.

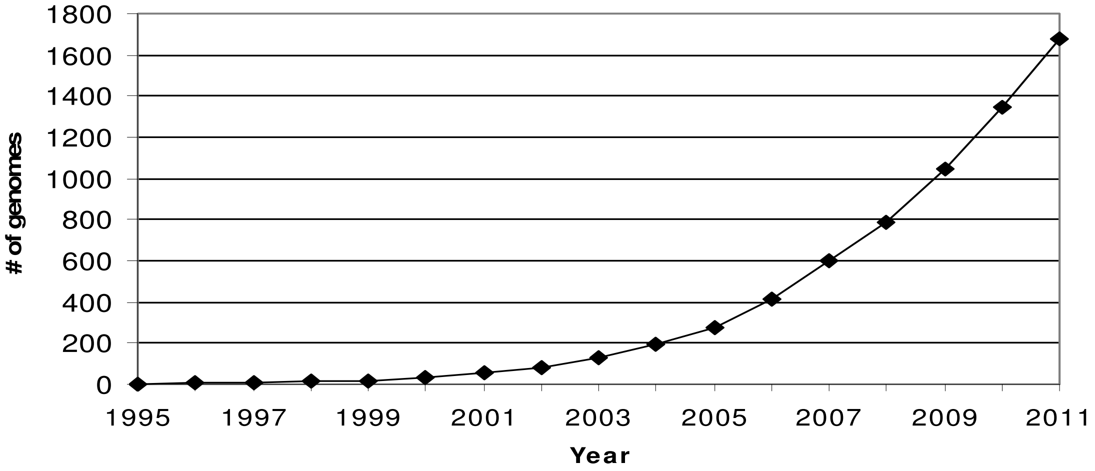


Figure 1. Completely sequenced prokaryotic genomes listed by the NCBI database (Nov. 2011).

The dramatic reduction in cost of sequencing bacterial genomes has now permitted large scale projects including those of metagenomes of the microbiota of the gut and environmental samples and includes the sequencing of 10k bacterial genomes until the end of 2012 [4]. Nevertheless, it is expected that this growth rate will even be exceeded in the coming years with the development of novel high-throughput sequencing technologies.

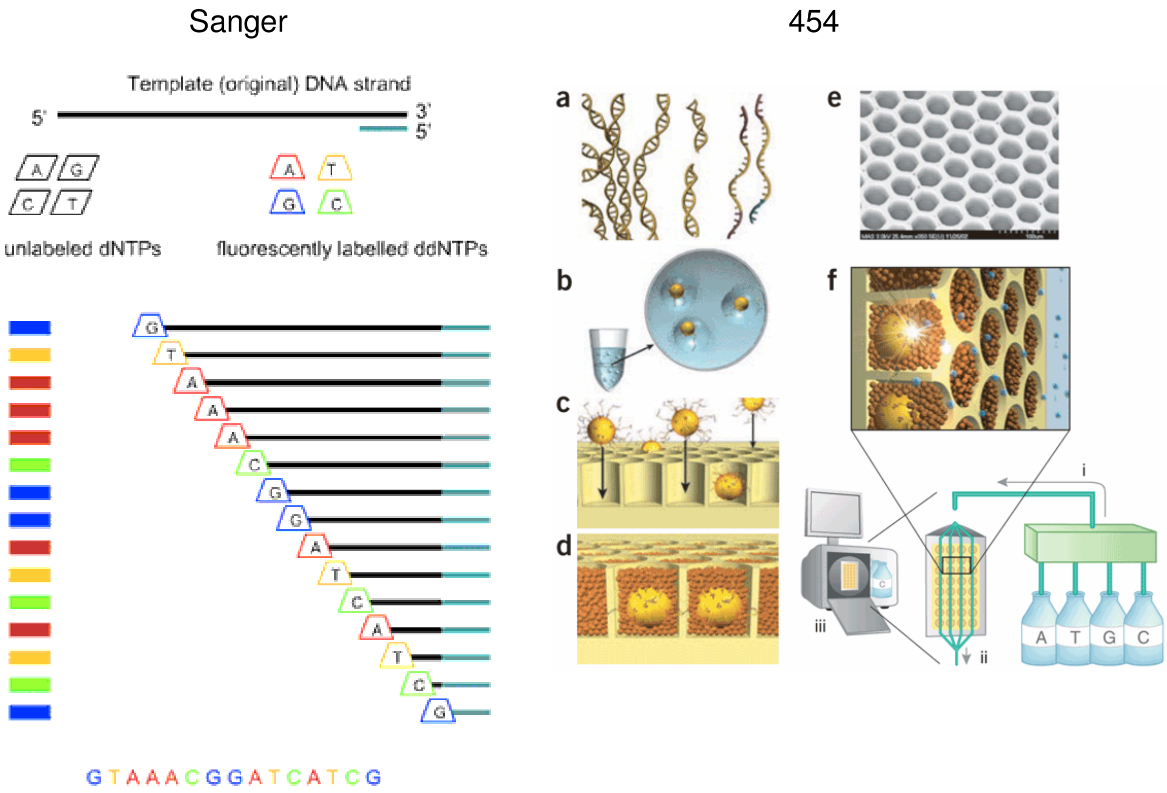


Figure 2. Sanger and 454 sequencing mechanisms [5, 6].

The shift in the pace of DNA sequencing was made possible by technical developments that allowed Sanger sequencing to be replaced by other novel technologies. Modern capillary Sanger sequencing employs a modified DNA synthesis reaction of a single-stranded template using a combination of normal nucleotides and labelled dideoxyribonucleoside triphosphates (ddNTP) with different fluorescent markers. The latter terminate further synthesis by a DNA polymerase leading to a mix of sequence fragments of different sizes which end with a labelled chain-terminating ddNTP. Fragments are separated by size via electrophoresis in gel-filled capillary tubes and the type of fluorescence is automatically recognized by a camera leading to read lengths of up to 1000 bp. The main drawback for application of the Sanger technology on whole-genome sequencing is based on the necessity of cloning sheared genomic DNA into vectors for replication [7].

This was no longer necessary with the development of the first massively parallel next generation sequencing technology by Roche/454 in 2005. Here, two different adapters are attached to each end of a single-stranded fragment of genomic DNA. One of these adapters is then linked to a bead and the sequence is amplified by polymerase chain reaction (PCR) inside a water-oil emulsion to create multiple copies per bead. Beads are separated into picolitre-sized wells on the sequencing chip and the reaction is carried out with a primer homologous to the second adapter of each immobilized DNA fragment using pyrophosphate nucleotides. These are cyclically flowed across the chip one type of nucleotide at a time, resulting in the release of pyrophosphate groups relative to the number of nucleotides incorporated. Pyrophosphate is converted to ATP by a sulfurylase, which is harvested by a luciferase to emit light that can be recognized by a camera resulting in read lengths of up to 500 bp. A known weakness of pyrosequencing stems from the elevated error rates for homolopolymeric tracts (repetitions of one nucleotide), since the signal strength is only linear for up to eight consecutive nucleotides [8].

Concurrent technologies for massively parallel sequencing have appeared quickly [9]. The idea of sequencing populations of amplified template DNA molecules was taken up by other commercial entities employing different chemistries (SOLiD, Illumina/Solexa) resulting in higher coverage but relatively short read lengths of 50-100 bp. The IonTorrent strategy does also employ an amplification step, but replaced the optical measurement of DNA synthesis with recognition of released hydrogen ions after incorporation of a nucleotide, leading to cost reductions. More recent developments try to sequence single DNA molecules to avoid the bias introduced by amplification. Pacific Bioscience sequencing (SMRT) is based on an immobilized DNA polymerase. During synthesis, fluorophores are cleaved from the incorporated dNTPs and can be continuously detected, leading to very long reads of up to 5000 bp, but also high error rates of up to 15%. Future third-generation techniques will likely include nanopore sequencing, where single DNA molecules in solution can be driven electrophoretically through a nano-scale pore, leading to highly confined spaces, which may offer advancements in cost, speed and quality [10].

When considering *de novo* sequencing of prokaryotic genomes, 454 pyrosequencing offers several advantages. The unique combination of relatively long and high quality reads permit the unambiguous integration of repeat sequences, thereby simplifying the finishing phase of current whole-genome sequencing and reducing the total cost. In contrast, sequencing technologies that have relatively short read lengths require a higher coverage, more bioinformatic analysis and a higher number of additional gap-spanning long reads to successfully complete a genome.

	Read length	Throughput	Accuracy	Cost per Gb
Sanger 96 Lane Capillary	1200	115 Kbp	99.9%	$$$
Roche-454/GS-FLX Titanium	500	500 Mbp	99.9%	$$
Illumina/HiSeq 2000	100	200 Gbp	99.5%	$$
ABI/SOLiD 5500xl	50-100	>100 Gbp	99.9%	$$
Pacific Biosciences	>1000	50 Mbp	85%	$$
IonTorrent	200	1 Gbp	99.6%	$

Table 1. General features of prevalent sequencing technologies [11,12].

Introduction of high-throughput technologies and an increase in computing power has led to an explosion of primary data deriving not only from bacterial genome sequences but also from the exploration of bacterial communities in different ecological niches. Currently, sequence data is collected by various databases including the NCBI [13] or the EBI [14] and analyzed by workflows for annotation, protein structure, etc. established at these websites. However the qualities of these first-pass annotation tools, while useful are not adequate for further in-depth comparative analysis. Therefore, specialized databases focusing on specific aspects or single organisms have been established with input from dedicated researchers in the field. In addition, there is a lot of activity in constantly generating novel bioinformatics software tools for the processing and analysis of data and their integration into pathways and networks.

2.2 From reads to replicons - Assembly and finishing strategies

Current genome sequencing projects combine a high-throughput primary sequencing with PCR Sanger reads to close gaps which remain, either due to the presence of repeats or non-unique regions (e.g. rRNA) that could not be automatically resolved by *de novo* assembly of reads.

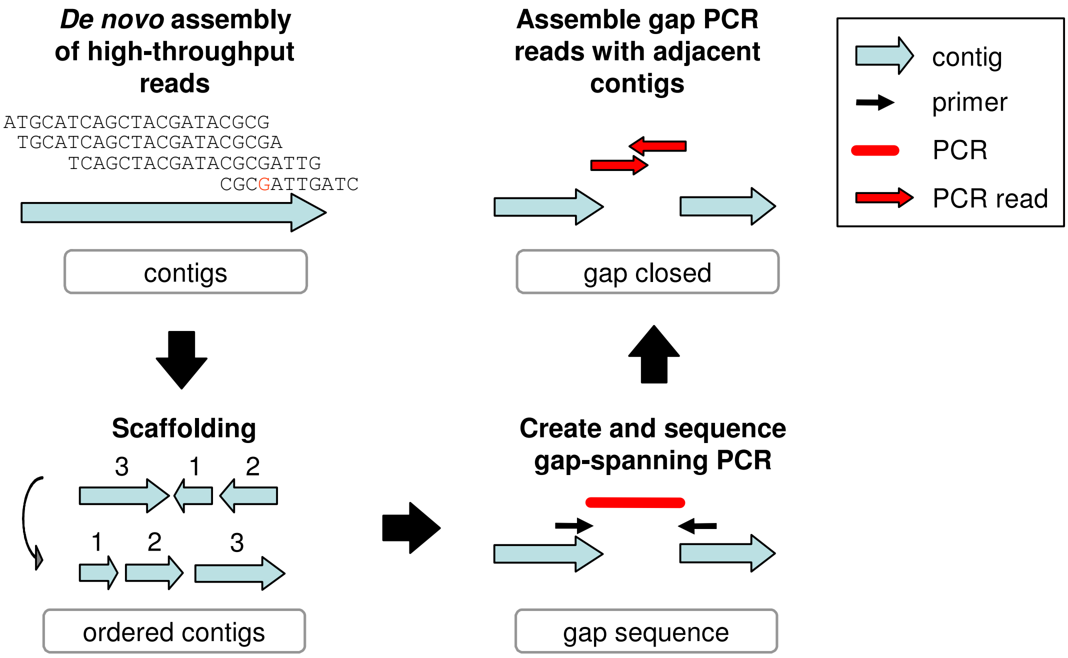


Figure 3. Gap closure strategy to completely sequence replicons.

High-throughput sequencing leads to a large number of overlapping sequence fragments of 30-1000 bp (reads) depending on the technology employed. These have to be joined to consecutive sequences (contigs) using various assembly strategies optimized for read size and amount of coverage supplied by a host of different tools [15]. The 454 Newbler assembler is arguably the dominant tool for *de novo* assembly of long reads due to its unsurpassed performance in the field [16].

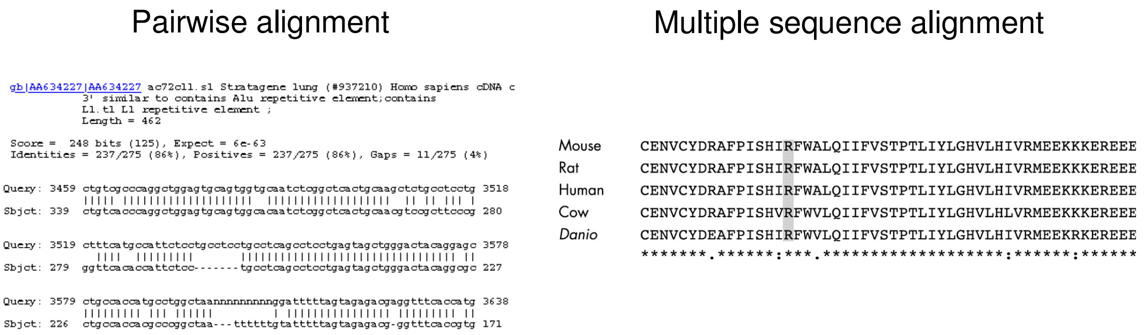
Due to the random sequencing approach, order and orientation of contigs in the sequenced replicon is unknown. Contigs can principally be scaffolded by using three methods: (i) by comparison to one or multiple related reference replicons, (ii) by mate-pair sequencing of both ends of a larger insert library which allows the determination of order and distance of two reads relative to each other, and (iii) by optical mapping based on digestion of immobilized DNA molecules and determination of size and order of fragments [17-19].

In the presence of a closely related complete replicon, reference mapping is probably the optimal solution, since little additional experimentation to close the gaps is required. In order to completely sequence replicons, gaps between contigs have to be filled by Sanger sequencing of PCRs overlapping both contigs. This necessitates the identification of primers, which specifically bind only inside the sequence bordering the gaps. This step is facilitated by software packages such as Oslay or Projector2, which are able to predict specific primers. However, fully automated layout of contigs according to a reference genome may not be optimal [20,21]. Additional commercial solutions exist that identify specific primers against a background (e.g. PrimerPremier) but these can only be retrieved gap by gap and rely on extensive manual preparation and interpretation. One of the most widespread tools to predict primer pairs without regarding specificity is the open source software Primer3, which has been widely adopted because of its rich set of features and accessibility [22].

The resulting PCR sequences need to be aligned and joined with the respective bordering contigs in order to achieve a contiguous sequence using either open source (e.g. Consed, Hawkeye) or proprietary software (e.g. Geneious, CLC, SeqMan) [23].

2.3 A primer for sequence comparison

Today, the primary bottleneck is no longer the production of sequence data but the comparison of new sequences with known sequences whose function have already been resolved. Various alignment programs offer comparison functions tailored for different situations.



Pairwise alignment

```
gb|AA634227|AA634227 ac72c11.s1 Stratagene lung (#937210) Homo sapiens cDNA c
           3' similar to contains Alu repetitive element;contains
           L1.t1 L1 repetitive element ;
           Length = 462

 Score =  248 bits (125), Expect = 6e-63
 Identities = 237/275 (86%), Positives = 237/275 (86%), Gaps = 11/275 (4%)

Query: 3459 ctgtcgcccaggctggagtgcagtggtgcaatctcggctcactgcaagctctgcctcctg 3518
            ||||| ||||||||||||||||||||  |||||||||||||||||||  || || ||| |
Sbjct: 339  ctgtcacccaggctggagtgcagtggcacaatctcggctcactgcaacgtccgcttcccg 280

Query: 3519 ctttcatgccattctcctgcctcctgcctcagcctcctgagtagctgggactacaggagc 3578
              ||||  |||||||||       ||||||||||||||||||||||||||||||||| ||
Sbjct: 279  ggttcacaccattctcc-------tgcctcagcctcctgagtagctgggactacaggcgc 227

Query: 3579 ctgccaccatgcctggctaannnnnnnnnggatttttagtagagacgaggtttcaccatg 3638
            ||||||||| ||| ||||||         | ||||||||||||||||| ||||||||| ||
Sbjct: 226  ctgccaccacgcccggctaa---ttttttgtatttttagtagagacg-ggtttcaccgtg 171
```

Multiple sequence alignment

```
Mouse     CENVCYDRAFPISHIRFWALQIIFVSTPTLIYLGHVLHIVRMEEKKKEREEE
Rat       CENVCYDRAFPISHIRFWALQIIFVSTPTLIYLGHVLHIVRMEEKKKEREEE
Human     CENVCYDRAFPISHIRFWALQIIFVSTPTLIYLGHVLHIVRMEEKKKEREEE
Cow       CENVCYDRAFPISHVRFWVLQIIFVSTPTLIYLGHVLHLVRMEEKKKEREEE
Danio     CENVCYDEAFPISHIRFWVLQIIFVSTPTLIYLGHVLHIVRMEEKRKEREEE
          *******.******:***.*******************:******:******
```

Figure 4. Alignments between two or multiple sequences.

Pairwise alignment programs compare only two sets of nucleotide or amino acid sequence (e.g. BLAST, FASTA) [24]. BLAST is probably the most widely used tool and offers very fast local alignments by a heuristic search that is suited for the identification of similar regions within two sequences which could represent an evolutionary or functional relationship. Another program from the BLAST suite of software is called BLASTCLUST and can automatically join multiple pairs of similar sequences into single-linkage clusters (AB + BC = ABC) representing sequence families. This removes the need to compute all-vs-all pairwise comparisons since no further alignments are necessary after a sequence has been sorted into a certain cluster, thereby saving increasing amounts of time as the dataset grows.

Multiple sequence alignment programs (e.g. ClustalW, Muscle) can find similarities between several sequences using more complex algorithms [25]. The resulting alignment is frequently used as an input by software which tries to depict relationships between organisms in form of a phylogenetic tree.

In order to align large sequences of millions of bases (e.g. prokaryotic chromosomes) whole genome alignment programs are necessary (e.g. Mummer, Mugsy, Mauve, MAVID) [17,26-28]. These offer less sensitivity but can identify homologous regions between complete replicons on typical desktop hardware.

All of these tools allow the identification of sequences deemed to be similar according to a user-defined cutoff of identity and/or coverage which indicates a functional or phylogenetic relation (homology). Genes which resulted from a speciation event are called orthologs, while those which arose from duplication in one strain are referred to as paralogs [29]. The identification of orthologous genes in different organisms is frequently facilitated by the so called bidirectional best pair, which denotes those genes in two replicons that display the highest similarities when comparing all genes of one replicon to all genes of another replicon and vice versa. However, to reliably predict orthology, the comparison is often extended to the surrounding genes, whose conservation (synteny) can serve as an additional indicator of vertical inheritance.

2.4 Small non-coding regulatory RNA identification

While genome sequences can reveal the genetic repertoire of a strain, they do not correctly predict the expression of genes under various conditions due to the complexity of the sensing and regulatory networks involved, as well as of the impact minor genomic differences. Transcriptome-based approaches are directed at uncovering the amount of transcripts including protein-coding mRNAs or small non-coding regulatory RNAs (sRNA) using either microarray hybridization or RNA sequencing, thus uncovering a second layer of regulation [30]. The class of sRNA has recently moved into the focus of attention due to its involvement in different cellular processes, such as environmental adaptation and pathogenesis [31]. Bioinformatic *de novo* identification of new sRNAs based on the localization inside intergenic regions and the presence of promoter or terminator structures results in high false positive rates, since these traits do not necessarily result in transcription or a regulatory effect of the resulting RNA. The presence of the nucleotide sequence of known sRNAs can nonetheless be identified in other strains by homology searches [32]. While various software programs exist that serve to collect and compare published sRNA sequences, these are unable to quickly identify a batch of sRNAs in multiple genomes including the genomic neighborhood, a drawback that we tried to address with the development of a specialized software called sRNAdb [33,34].

2.5 Capturing diversity – genome, transcriptome, pan-genome, meta-genome

Selective pressure leads to constant adaptation of prokaryotes based on mutation (deletion, duplication, recombination), or horizontal gene transfer [35,36]. These processes can modify single nucleotides (e.g. single nucleotide polymorphisms = SNPs) or extensive areas of the affected replicons. Differentiating between an insertion in one strain and a deletion in another is not a trivial task, which has led to the use of the term “indel” to describe unclear situations.

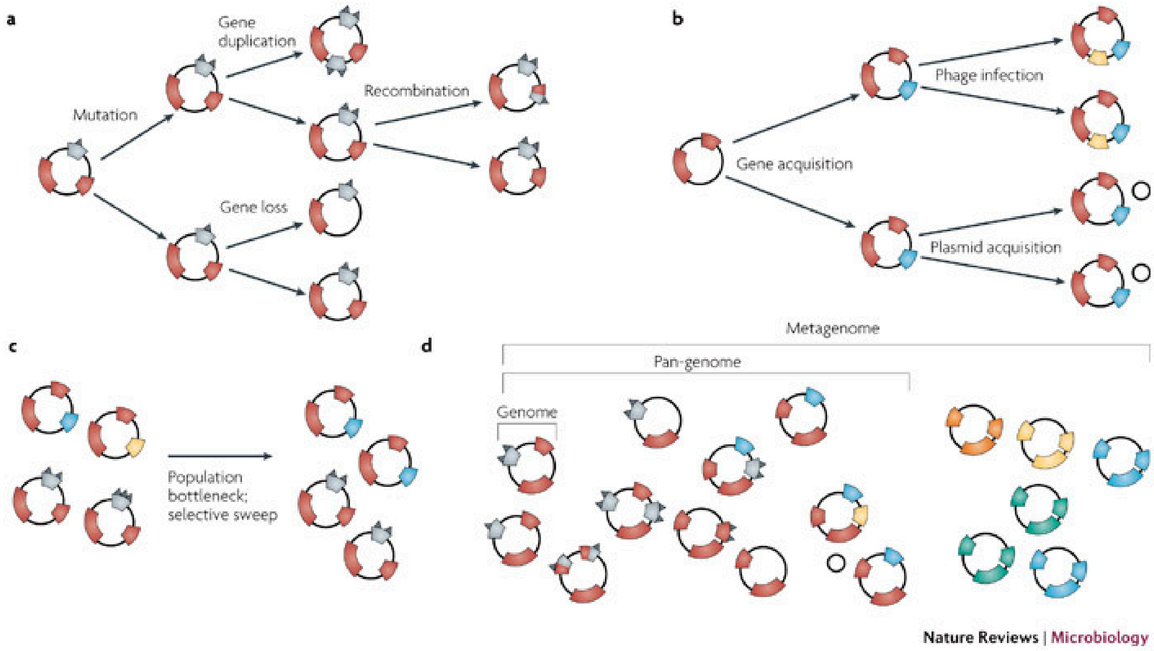


Figure 5. Subset of molecular evolutionary mechanisms based on mutation (a), horizontal gene transfer (b), and combinations thereof influencing populations (c, d) [35].

Horizontal gene transfer can result from a number of processes. Bacteriophages inject their DNA into bacterial hosts which may either lead to lysis of the bacterium or the inclusion in a lysogenic prophage state. Mobile elements bear transposition genes to extract themselves from replicons and reinsert at different positions or inside other replicons and thus spread genes with a high adaptive potential (e.g. pathogenicity factors, stress response, heavy metal resistance, host adaptation). They are frequently found on extrachromosomal plasmids which serve as storage and transfer facilities for mostly non-essential genes. Genomic islands represent combinations of genes originating from prophages, mobile elements and/or plasmids highlighting the fluidity of processes shaping bacterial genomes. These horizontally transferred regions can be identified because of differing sequence composition, the presence of mobilization genes or by comparison to closely related reference genomes that do not harbor them [37,38].

In order to capture the diversity of whole species or genera, the pan-genome concept was developed, which relies on a comparison of multiple available genomes per taxa [39]. These are used to divide genes into three categories: (i) core genome that is present in all strains and codes for basic housekeeping functions, (ii) accessory genes which are present in several but not all strains and contain a varying adaptive potential useful for molecular fingerprinting, and (iii) strain-specific genes which represent the youngest and most volatile genetic sequences. Other nomenclatures only distinguish between the core and dispensable genome, whereby the latter also includes strain-specific genes. The size of a pan-genome is associated with the niche and lifestyle of the respective taxa, since survival in diverse environments necessitates a larger amount of accessory genes (e.g. *Bacillus cereus, Escherichia coli*) than, for example, survival inside isolated niches or as parasitic or symbiotic organisms (e.g. *Streptococcus pyogenes, Haemophilus influenzae*) [38,40-42]. Communities of microorganisms found in an environmental niche are considered as a metagenome that will help to understand complex interactions between microbes and their environment in the future [43].

While software tools exist to identify differential conservation of regions inside a pan-genome, these were found to have poor visualization features, require high level bioinformatic abilities to be used effectively, are usually not locally installable, or are only helpful for specific sub-problems such as the identification of discriminatory regions for diagnosis [44,45]. In order to deal with these problems, various pan-genome analysis modules were apprehended to an existing software for comparative analysis of prokaryotic replicons called GECO [46] (see chapter 3.1.4).

2.6 Annotation

In order to understand the contents of a genome, its functional units such as protein-coding open reading frames (ORF) or RNAs (tRNA, rRNA) have to be identified by recognition of known sequences with an established function [47-49]. A host of software tools attempt to solve these problems using web-based or local installations, more or less manual supervision, by distributed or desktop computing, and with diverging degrees of modularity [50]. While fully automatic procedures exist for this task, they rarely produce optimal results. Gene names may not be identified at all (e.g. RAST) or disregard consistent nomenclature for modules of genes (e.g. GenDB) [51,52]. This artifact is a result of frequent reliance on the best BLAST hit inside a database bearing genes of many different organisms, which may use differing annotation schemes and where relation is based on inconsistent measures of sensitivity and specificity. Furthermore, truncated or fragmented genes are rarely recognized properly, leading to annotation based on spurious matches. To overcome this, many annotation suites offer interfaces which present all identified similarities to genes inside various databases (e.g. NCBI nr = non-redundant protein sequences, Pfam = protein domains, COG = clusters of orthologous groups encompassing functional categories) and allow a manual correction based on these data [53-55]. An external tool frequently used to assess and improve upon the putative function of problematic genes is the STRING database, which offers network correlations of genomic, transcriptomic and additional experimental data aggregating the evidence into a concise image for the user [56].

Use of these tools is hampered by the complexity of installation, lack of influence on the results and/or lack of maturity of many open source software program. Thus, GenDB or Ergatis pipeline software could not be installed locally due to minor differences amongst Linux distributions, despite personal involvement of authors of these projects. Recently, comparative analysis of syntenic regions of multiple strains has gained traction to improve upon quick and congruent identification of open reading frames or annotation [46,57,58].

2.7 Phylogenetic trees

Phylogenetic trees attempt to depict evolutionary relationships of genes or organisms based on (i) presence/absence of phenotypic markers (e.g. resistances, metabolic functions, surface antigens), (ii) nucleotide sequence similarities (microarray hybridization, sequence alignments of genes or proteins), or (iii) a combination thereof (gene content = presence/absence of genes which denotes presence/absence of functions) [35,59]. These data are computationally analyzed to infer putative ancestry and relative distance of elements based on their location in the resulting branching diagram. The tree can be presented as a phylogram, where distances are informative, or as a cladogram, which only depicts topology.

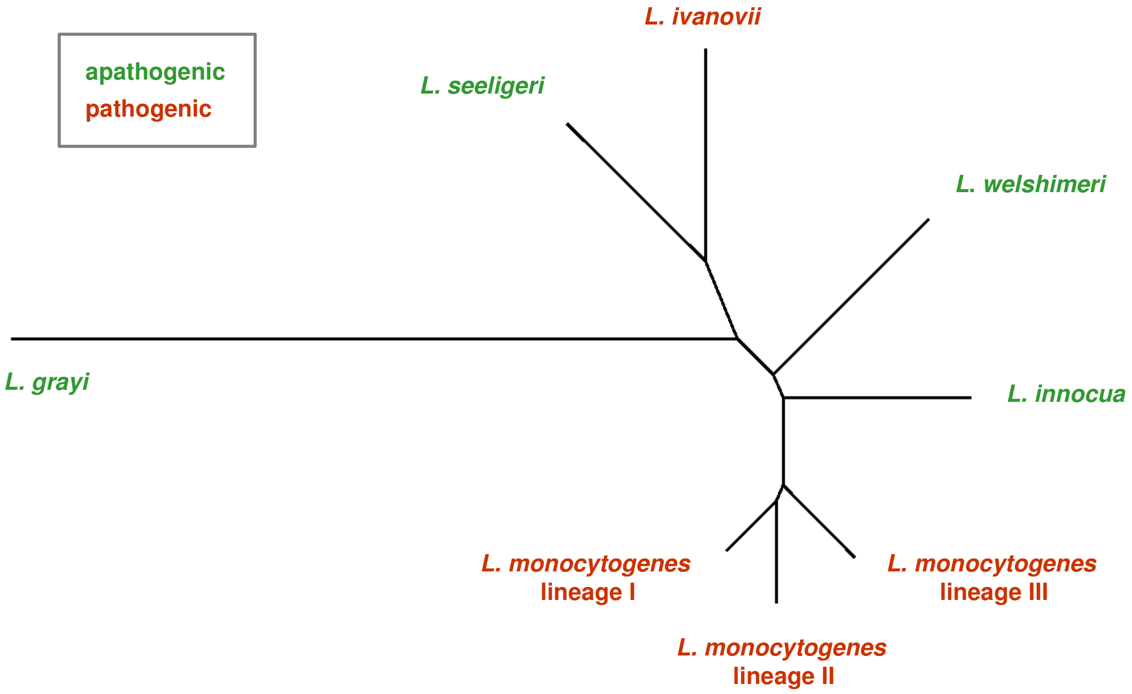

apathogenic
pathogenic
L. ivanovii
L. seeligeri
L. welshimeri
L. grayi
L. innocua
L. monocytogenes lineage I
L. monocytogenes lineage III
L. monocytogenes lineage II

Figure 6. Phylogram of species of genus *Listeria* based on a whole-genome alignment.

In the case of sequence-based comparisons, multiple sequence alignment programs are necessary to overlap the underlying sequences prior to construction of a phylogenetic tree based on the identified variances [60]. In short, distance based methods (e.g. neighbor joining, minimum evolution) rely on the computation of pairwise distances, followed by sequential clustering of taxa starting with the most related pair in order to identify the tree with the shortest total length, which is computationally the most efficient strategy. Character-based methods (maximum parsimony/likelihood) evaluate all possible trees, which is more sensitive but also time consuming and thus only applicable to smaller datasets. Finally, bootstrapping is used to check reproducibility of a tree by random sampling of positions from the multiple sequence alignment with replacements, followed by recreation of the tree. This process is repeated multiple times to discern the number of trees which confirm the original topology. These and other algorithms are available from phylogenetic inference suites (e.g. Mega, SplitsTree), while further software exists for the visualization of resulting trees in publication journals (e.g. Dendroscope) [61-63].

The significance of a phylogenetic tree correlates with the amount of underlying data, so that a whole-genome alignment covering millions of bases is considered most accurate, while reliance on only few genes largely reduces the number of informative sites, which are furthermore subject to stronger variation of evolutionary rates (e.g. 16s rRNA, MLST).

2.8 Comparative visualization

Comparative visualization is an important tool to quickly convey information about numerical (e.g. degree of sequence similarity, expression, GC content) or categorical data (e.g. functional category, prophage) of whole genomes or specific regions. While dozens of different software exists for this task, here the introduction will only focus on three of them, which include all features found to be valuable for the publications on which this thesis is based.

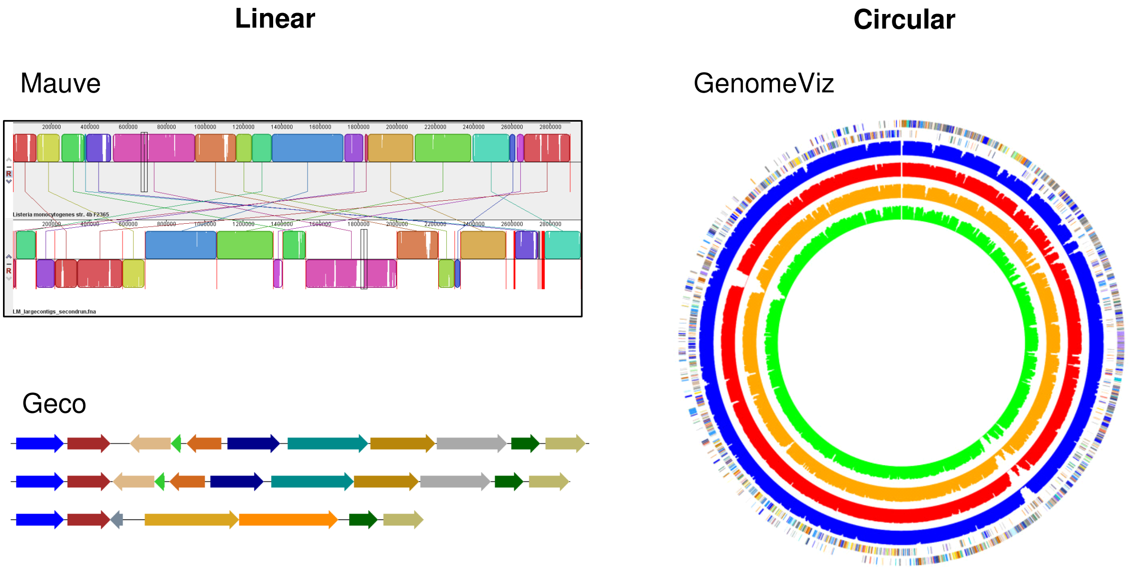

Linear
Circular
Mauve
GenomeViz
Listeria monocytogenes str. 4b F2365
LM_largecontigs_secondrun.fna
Geco

Figure 7. Examples for linear or circular visualizations. Mauve depicts homology of contigs versus a reference genome, GECO shows conservation of a locus in three replicons on gene level, and GenomeViz displays four whole-genome alignments versus a reference.

For example the whole-genome visualization software Mauve can quickly create multiple whole-genome alignments using the Muscle program and display these in a linear fashion (see Figure 7) [17]. Similar regions are color-coded, thereby allowing immediate visualization of insertions, deletions, and translocations, valuable in mapping contigs versus a reference replicon. It should be noted that Mauve can also give misleading results when encountering non-unique regions on the chromosome, which are displayed as regions with no similarity. GenomeViz allows circular display of numerical or categorical data, which must be pre-computed by the user [64]. While this requires basic knowledge in informatics, the resulting flexibility ensures a wide array of possible data visualizations (e.g. G/C content, alignment scores, RNAs, mobile elements). GECO focuses on limited genomic regions useful for the identification of syntenic, paralogous, fragmented or horizontally transferred genes [46]. Homology is pre-computed by BLASTCLUST and the client/server architecture does not necessitate a local installation, as long as only published replicons are compared.

2.9 Genus *Listeria*

Listeriae are Gram-positive Firmicutes related to the genera *Bacillus*, *Clostridium*, *Enterococcus*, *Streptococcus* and *Staphylococcus* all of whom have a low G+C content. The lifestyle of listeriae ranges from saprotrophic to opportunistic pathogenic [65]. Currently, there are eight species assigned to this genus (see Figure 6). However, only species *L. monocytogenes* has been described to be a human pathogen [66]. Non-pathogenic species are hypothesized to have developed following genome reduction of pathogenic progenitor strains of species *L. monocytogenes* [67,68].

In approximately 30% of all wild-type strains, large plasmid species have been described. These plasmids occur in both pathogenic as well as non-pathogenic strains [69-71]. An overrepresentation of listerial plasmids in strains isolated from the environment and food processing facilities implies that possession of a plasmid confers a greater selective advantage in these niches as compared to clinical isolates, where plasmids are rarely found [72,73].

2.10 Species *L. monocytogenes*

Human listeriosis caused by *L. monocytogenes* is one of the deadliest foodborne diseases presently known and several outbreaks in recent years resulted in high mortality rates of up to 30% [74-77]. Thus, for example, an outbreak of listeriosis between September and December of 2011 in the USA affected 146 individuals with 30 cases of mortality [78]. In contrast, the highly publicized *E. coli* HUSEC outbreak in Germany in 2011 involved 3400 affected individuals with 39 deaths [79]. Despite the low incidence rate in human infections, *L. monocytogenes* is thus a highly relevant germ, whose ubiquitous nature and resistance to acid, high salt concentrations and low temperatures lead to frequent contamination of food production plants.

The species *L. monocytogenes* is subdivided into four lineages and 12 serotypes [80]. Lineage I is overrepresented among human clinical isolates and clinical outbreaks, lineage II is associated with sporadic isolates from humans and animals, and lineages III and IV comprise strains that are rare and predominantly found in animals [81]. Strains of lineage II frequently display virulence-attenuated phenotypes due to deletions in major virulence genes [82-84]. Serotypes most commonly associated with human listerial infections are 1/2a, 1/2b and 4b [85]. Many genes found to be important for virulence or pathogenicity of the species code for surface-associated proteins, which modulate adhesion and interaction with the host using a number of recognizable protein domains (LPXTG, GW, P60, LysM, LRR) [86].

Prior to this thesis, only four complete genome sequences of genus *Listeria* were publicly available. This included two strains of species *L. monocytogenes* and one strain each of species *L. innocua* and *L. welshimeri*. Thus, genomic data was limited to two of the 12 serotypes of species *L. monocytogenes* [87]. The pan-genome of the species was previously assessed based on microarrays of 20 genomes (18 of these in draft-state) and was found to be a closed pan-genome, thus implying that *L. monocytogenes* has limited ability to effectively modify its genetic repertoire [88,89].

2.11 Resistance to bacteriophages

In order to protect their genome, bacteria have developed strategies dealing with invading foreign DNA. Restriction modification systems digest alien DNA [90] while abortive infection modules (Abi) encode for toxin/antitoxin systems that lead to cell death upon a modification of transcription or translation during a phage infection [91]. More complex frameworks resemble adaptive immune systems and were termed clustered regularly interspaced palindromic repeats (CRISPR) [92]. Here, the combination of spacer/repeat modules derived from phage sequences with CRISPR-associated genes (*cas*) facilitates a specific RNA-interference-like silencing of phage gene expression. The presence of CRISPR arrays inside a genome can be predicted by the identification of repeats using available software such as PILER-CR and CRT [93,94].

2.12 Main objectives of this study

The primary task of this thesis was the design, implementation, and application of software to create a pipeline to (i) assist complete sequencing of replicons included in a genome sample, (ii) comparatively annotate these replicons, (iii) comparatively analyze these in conjunction with publicly available data, (iv) and to create visualization tools for representation in publications.

The thesis aimed at examining evolutionary forces that have driven the differentiation of strains within the genus *Listeria*. This includes (v) the role of plasmids in niche adaptation, (vi) the size and distribution of the listerial pan-genome, (vii) forces shaping the compositions of chromosomes (mobile elements, hotspots), and finally (viii) the identification of new candidate genes required for its pathogenic lifestyle.

3 Results and Discussion

The bioinformatic pipelines created and developed in this study are the results of thesis-specific publications presented in this section. Since different approaches were used in these publications, these are briefly denoted as p1-p7 (publication 1-7).

3.1 Bioinformatic pipelines

This section describes bioinformatic pipelines for assembly, genome finishing, annotation and analysis.

3.1.1 Assembly and scaffolding of 454 pyrosequencing data

All genomes described in this study have been sequenced to closure, i.e. a single contig. This includes one chromosome and five plasmids that were finished by in-house sequencing (*Listeria monocytogenes 7* SLCC2482 chromosome and plasmid = p2 and p7, *Escherichia coli* HUSEC41 plasmids 1-4 = p4). Primary 454 shotgun data of genomic DNA was assessed for sequencing quality by checking control data such as the mean read length (LM7=232bp, HUS41=395bp) and the total amount of nucleotides sequenced (LM7=51m, HUS41=280m), which in turn was a measure of good sequencing quality. All reads were assembled *de novo* without a reference genome using the software programs SeqMan, CLC, and the 454 Newbler. The largest contigs were in all cases produced by the 454 Newbler assembler. Additional gaps introduced by the two other candidate assemblers were checked for read coverage quality and orthologous sequences found inside reference genomes and determined to be erroneous. Thus, the 454 Newbler assembler was employed for all subsequent studies.

Since closely related reference replicons were available for all genomes, most contigs could be scaffolded by comparison using the Mauve whole genome alignment software.

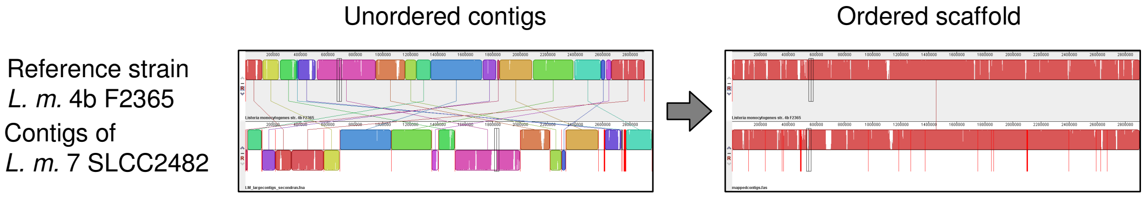


Figure 8. Example of scaffolded contigs of the chromosome of *L. monocytogenes* 7 SLCC2482 using Mauve. The ordered scaffold shows that Mauve was able to automatically order the large majority of contigs versus the reference replicon.

Due to known problems of Mauve considering assembly of non-unique sequences, an additional program was written in Java to allow the scaffolding of ambiguous contigs by harvesting BLAST searches versus reference replicons (BlastCollapser). This software summarizes data in a tabular form and includes information on all similar regions of a contig inside a reference to allow a concise overview of the possible mapping positions.

While the four plasmids in the genome of *E. coli* HUSEC41 had previously been identified by gel electrophoresis, the plasmid of strain *L. monocytogenes* 7 SLCC2482 was only recognized by the presence of two large contigs that did not fit into published chromosomes of species *L. monocytogenes*, but which resembled known plasmid sequences.

3.1.2 Finding specific primers for the finishing phase of genome sequencing

In order to rapidly predict specific primers in all gaps of an optimized user-defined scaffold of ordered contigs, a semi-automatic Java application called Minimap was written. Despite its name, it does not actually map contigs against a reference to identify their correct order and orientation, but relies on a predefined input to supply this data that can be produced by other software (e.g. Mauve, Projector2, Oslay). This enables the user to quickly adapt their scaffolding order in case of an error. Additionally, the user can supply a reference genome to automatically identify the putative size of gaps, which is helpful to determine optimal parameters for the subsequent PCR-based analysis.

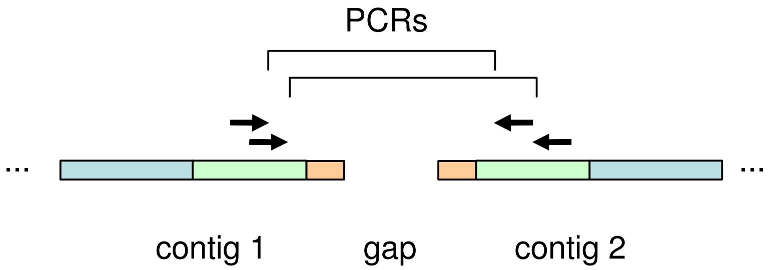


Figure 9. Minimap discards a selectable region from the contig borders (red) and predicts candidate primer pairs (black arrows) for PCRs from the following locus (green) for each gap in one run.

Minimap extracts the sequences bordering a gap (e.g. 450 bp of each contig) and discards the sequence closest to the gap (e.g. 150 bp) to permit an overlap of dedicated PCR sequence with the contig borders. In the example above, the resulting 300 bp of both contig borders are entered into Primer3, which predicts optimal primer pairs for each gap.

3.1.3 Joining contigs and gap closure PCRs

The proprietary software SeqMan (p2 and p7) and Geneious (p4) were used for gap-wise assembly of bordering contigs with gap-spanning PCR sequences.

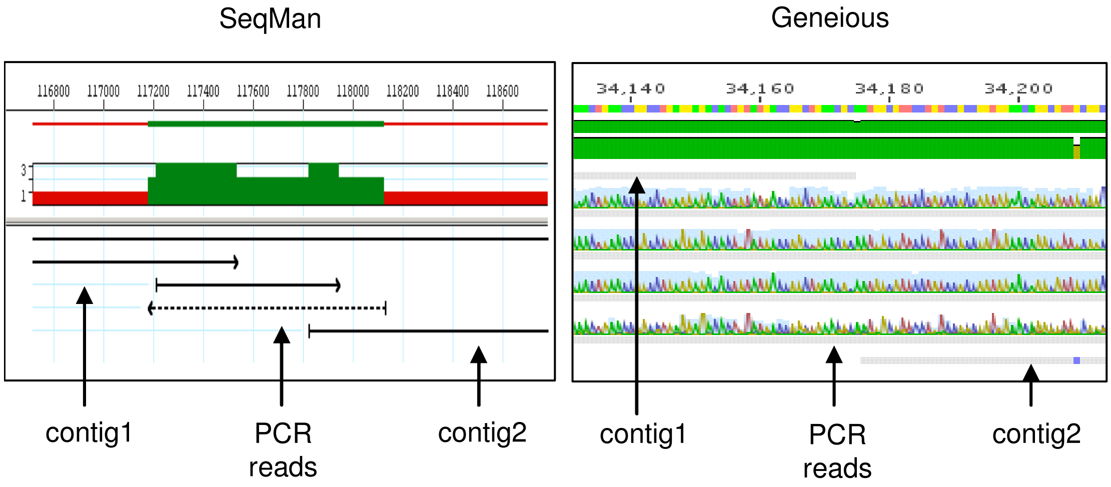


Figure 10. Gap assemblies using SeqMan and Geneious, depicting different visualization levels of gap-spanning PCR reads and adjacent contigs. The Geneious screenshot includes flowgram data to assess the quality of the basecalls.

Gap assemblies started with strict trimming of low quality bases of PCR reads and a high similarity to the bordering contig sequences. If no assembly was possible, the degree of trimming and minimum similarity was consecutively lowered to enable at least a partial inclusion of respective PCR reads.

Reads covering multiple highly similar regions (e.g. rRNA modules) are usually collapsed inside one contig by the assembler, resulting in gaps. In order to be separately resolved, gap-spanning PCRs of rRNA loci were produced and sequenced using published primers [95]. The putative locations of rRNA gaps were predicted based on the similarity to reference replicons, where these had already been identified.

Other misassemblies became apparent due to the presence of partially overlapping contigs or in differences when compared to reference replicons. One such error was identified in *Escherichia coli* HUSEC41 plasmid 4, which showed six contigs with overlaps that differed only by few SNPs and that could not be joined despite several attempts to produce gap-spanning PCRs. While this does not necessarily indicate an error, the identification of homopolymeric tracts that could not be resolved uniformly by the 454 sequencing and repeats in these contigs suggested a possible problem. A separate reassembly of all reads of the respective six contigs with a lower stringency resulted in three larger contigs without overlaps, which could be joined successfully.

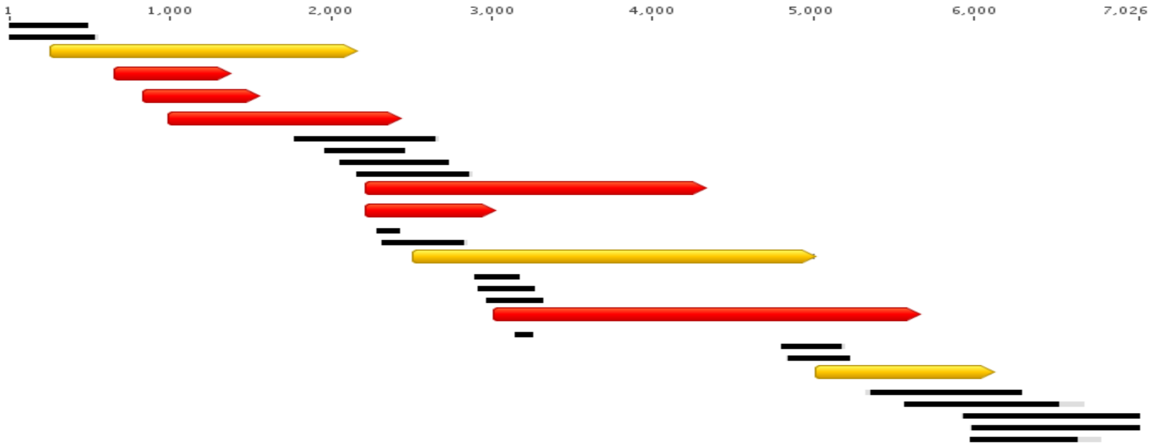
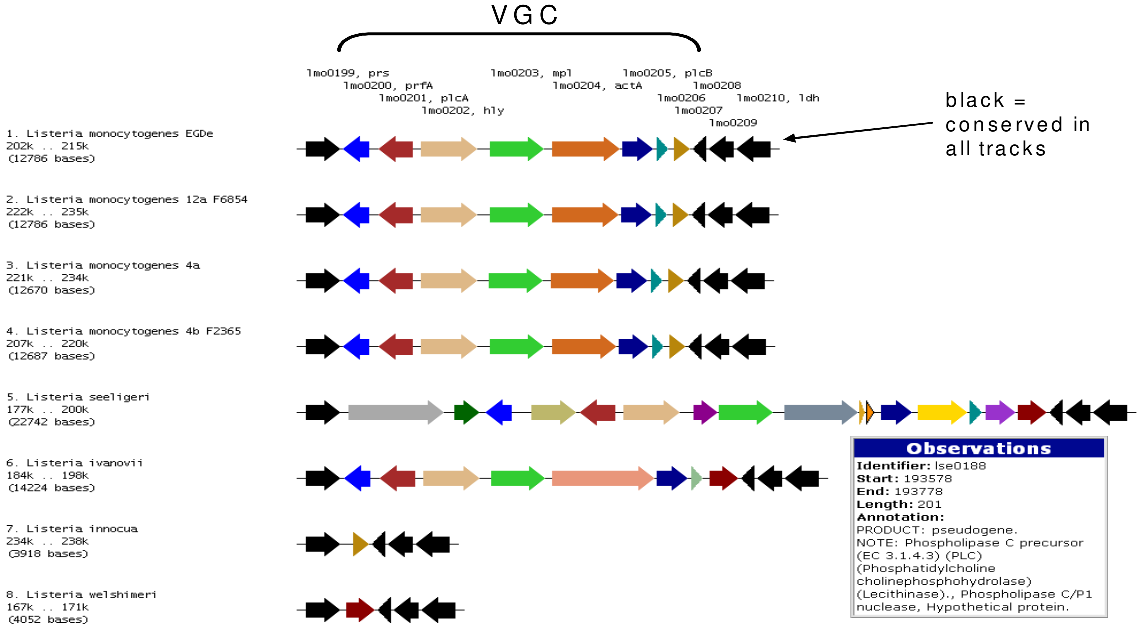


Figure 11. Geneious assembly of primary contigs (red), reassembled contigs (yellow) and PCR reads (black).

All contig positions with unclear base-calling were controlled by either manual assessment of the underlying read distribution, by comparison to reference replicons or polished by PCR to ensure a high quality. Finally, the sequence of each gap including surrounding contigs was exported per gap resulting in supercontigs which could be jointly assembled into one consecutive sequence per replicon.

3.1.4 Comparative analyses using GECO

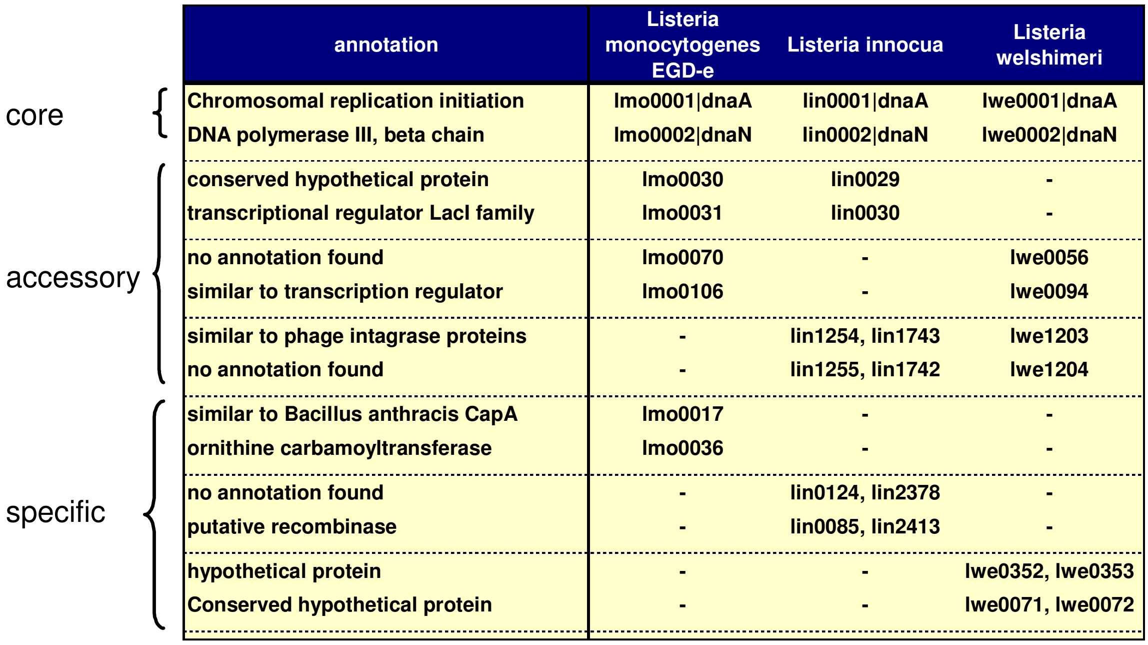
The locally installable GECO webserver was employed for comparative analyses in all publications described here, and can be accessed through any web browser. The abilities of the software package have been extended and improved considerably to allow faster and more extensive analyses since its publication in 2007. The original server has been partially upgraded since then, but only the most recent installation contains the full functionality [96,97].



Observations
Identifier: lse0188
Start: 193578
End: 193778
Length: 201
Annotation:
PRODUCT: pseudogene.
NOTE: Phospholipase C precursor (EC 3.1.4.3) (PLC) (Phosphatidylcholine cholinephosphohydrolase) (Lecithinase)., Phospholipase C/P1 nuclease, Hypothetical protein.

Figure 12. Example of a GECO visualization of the *Listeria* virulence gene cluster (VGC). Homologous genes are color-coded (black = core) and putative HGT predicted by deviating sequence composition is framed in black. Mutation or loss of virulence genes in apathogenic species (*L. seeligeri, L. ivanovii, L. innocua, L. welshimeri*) in relation to putative ancestor *L. monocytogenes* is clearly visible.

Among the additions to GECO since its original publication in 2007 is a direct link for the extraction of sequence data connected to each gene in the visualization servlet. This feature accelerates access to various sequence-based databases (e.g. NCBI, Pfam, STRING). The primary advantage in relation to the older version is represented by the introduction of multiple powerful homology lists and matrices, which are intended for assistance in annotation, pan-genome prediction and analysis, and batch-wise extraction of nucleotide or amino acid sequences for groups of genes (p1, p2, p4, p6, p7).



	annotation	Listeria monocytogenes EGD-e	Listeria innocua	Listeria welshimeri
core	Chromosomal replication initiation	lmo0001|dnaA	lin0001|dnaA	lwe0001|dnaA
	DNA polymerase III, beta chain	lmo0002|dnaN	lin0002|dnaN	lwe0002|dnaN
accessory	conserved hypothetical protein	lmo0030	lin0029	-
	transcriptional regulator LacI family	lmo0031	lin0030	-
	no annotation found	lmo0070	-	lwe0056
	similar to transcription regulator	lmo0106	-	lwe0094
	similar to phage intagrase proteins	-	lin1254, lin1743	lwe1203
	no annotation found	-	lin1255, lin1742	lwe1204
specific	similar to Bacillus anthracis CapA	lmo0017	-	-
	ornithine carbamoyltransferase	lmo0036	-	-
	no annotation found	-	lin0124, lin2378	-
	putative recombinase	-	lin0085, lin2413	-
	hypothetical protein	-	-	lwe0352, lwe0353
	Conserved hypothetical protein	-	-	lwe0071, lwe0072

Figure 13. Schematic pan-genome distribution of a homology matrix of three strains offered by GECO (cluster-mode) based on a user-defined homology cutoff. Each line represents genes of a homologous cluster functionally equivalent to a protein family, which may include multiple paralogous genes per replicon. Annotation is taken from the first gene inside a cluster starting from the left.

These allow quick large-scale comparisons of homologies of complete replicons inside spreadsheets. Lists can be sorted in different ways: (i) by degree of conservation ranging from core to specific genes (cluster-mode), (ii) by syntenic layout to identify chromosomal hotspots and insertions (synteny-mode), (iii) and finally by homology to a reference strain including separate annotations of all strains (annotation-mode). Furthermore, indels between two strains including paralogous expansion/compression (gain/loss-mode) and mean/median numbers useful for the description of size and conservation of a total pan-genome (pangenome-mode) can be produced.

When a list of interesting genes are identified using taxonomic or functional distribution, simple batch-extraction of clusters including nucleotide or amino acid sequences based on homology to these is possible. This data can serve as input for multiple sequence alignment software (e.g. for the extraction of MLST genes for a phylogenetic tree) or for other databases.

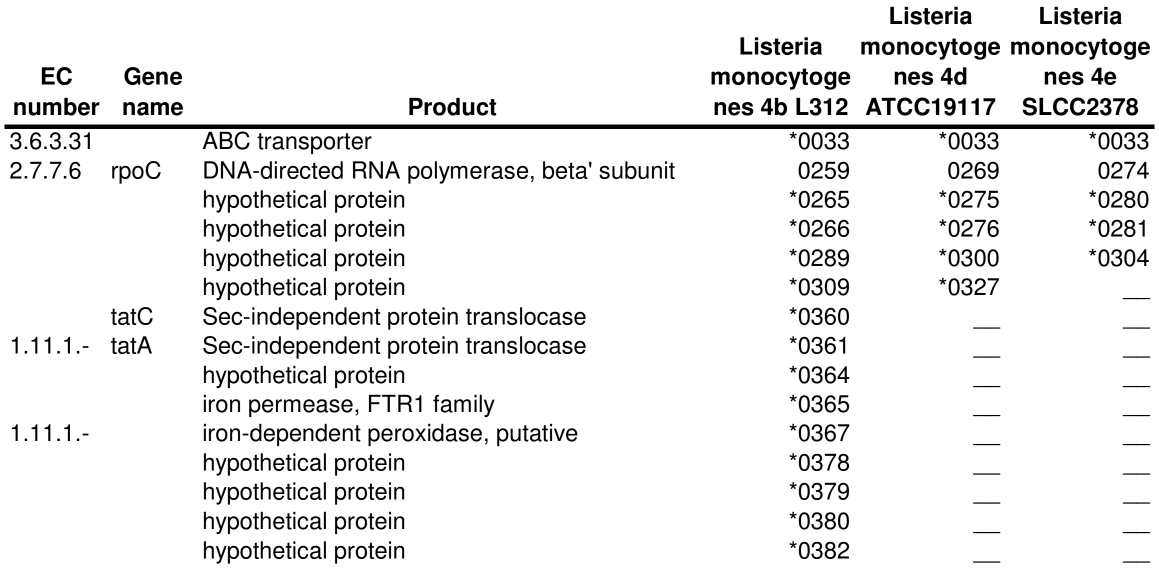
Batch-extraction of nucleotide sequences based on absolute replicon coordinates enables comparison of sequences overlapping multiple genes or inter-genic sequences. This can be employed to comparatively identify junction regions that still contain sequence remnants left inside a locus following a deletion event.

3.1.5 Annotation pipeline

The annotation pipelines used for publications in this thesis have changed considerably over time to become more independent from monolithic annotation suites. Initially, we used the GenDB software established at the CEBITEC Bielefeld as a central platform for annotation of genomes. As this thesis progressed, this dependence on the GenDB platform was found to be a bottleneck in the processing of data and a local pipeline was established to enable rapid annotation and analysis of genomes.

For example, in 2008/2009 (p1, p6) sequencing data was exported and integrated into the automatic annotation environment of GenDB. These data had then to be directly imported into the database by the authors of GenDB. Further collaborative effort was necessary to annotate remaining genes inside GenDB. The export of the final annotation had to be modified manually by GenDB staff in Bielefeld in order to satisfy submission criteria for the annotation of truncated genes, tRNA, and rRNA as requested by the EBI. Thus, at this stage, it was only possible to effectively use GenDB in collaboration with the University of Bielefeld.

In order to facilitate local analysis of data deriving from genome sequences available in Giessen, we implemented a modular approach (p2, p4, p7). In this analysis, a primary first-pass automatic annotation by RAST or GenDB was imported into GECO and an annotation matrix of highly similar genes in reference genomes was exported (bidirectional best pair annotation mode). The remaining genes were comparatively annotated by homology to multiple reference replicons inside the GECO visualizer ensuring consistent nomenclature for modules of genes. In the case of inconsistent references, live-queries versus public databases (NCBI, Pfam, STRING) were included to update spreadsheet information. Similarly, identification of surface-associated proteins by Augur enabled the use of consistent notation by adding the type of domain in brackets (e.g. cell wall surface anchor family protein (LPXTG motif)). Recognition of truncated genes by browsing genomes in search for homologous regions that displayed a gene with differing lengths in one or several replicons completed the annotation process. The resulting matrix was integrated into a database created in this thesis, overlaying and extending the primary automatic annotation. Finally, data was automatically checked for errors (double spaces, non-consecutive numbering of identifiers, etc.) before export of an annotated flatfile to be submitted to the EBI database prior to publication.



				Listeria	Listeria
			Listeria	monocytoge	monocytoge
EC	Gene		monocytoge	nes 4d	nes 4e
number	name	Product	nes 4b L312	ATCC19117	SLCC2378
3.6.3.31		ABC transporter	*0033	*0033	*0033
2.7.7.6	rpoC	DNA-directed RNA polymerase, beta' subunit	0259	0269	0274
		hypothetical protein	*0265	*0275	*0280
		hypothetical protein	*0266	*0276	*0281
		hypothetical protein	*0289	*0300	*0304
		hypothetical protein	*0309	*0327	__
	tatC	Sec-independent protein translocase	*0360	__	__
1.11.1.-	tatA	Sec-independent protein translocase	*0361	__	__
		hypothetical protein	*0364	__	__
		iron permease, FTR1 family	*0365	__	__
1.11.1.-		iron-dependent peroxidase, putative	*0367	__	__
		hypothetical protein	*0378	__	__
		hypothetical protein	*0379	__	__
		hypothetical protein	*0380	__	__
		hypothetical protein	*0382	__	__

Figure 14. Excerpt of the final annotation matrix produced for strains of *L. monocytogenes* (p7). Numbers inside replicon columns refer to the respective locustags to be overwritten (e.g. *lmo0001*). All genes that could not be automatically annotated by high bi-directional best pair homology to a well-annotated reference were marked with an asterisk. If only one gene inside a cluster did not pass this cutoff, the complete cluster was manually assessed in detail.

The new comparative annotation approach is most useful for the simultaneous annotation of multiple replicons and in the presence of well-annotated and closely related references. Since only one annotation per cluster is necessary, it is not only faster but also ensures a congruent annotation of orthologous genes. Identification of diverging or fragmented genes is simplified and only basic knowledge of correlative functions of spreadsheet software is necessary.

3.1.6 Total replicon visualizations using Mauve and GenomeViz identify global similarities

Visualization of homologous regions of complete replicons is a common first step in assessing the degree of relationship connected with evolutionary patterns of insertions and deletions.

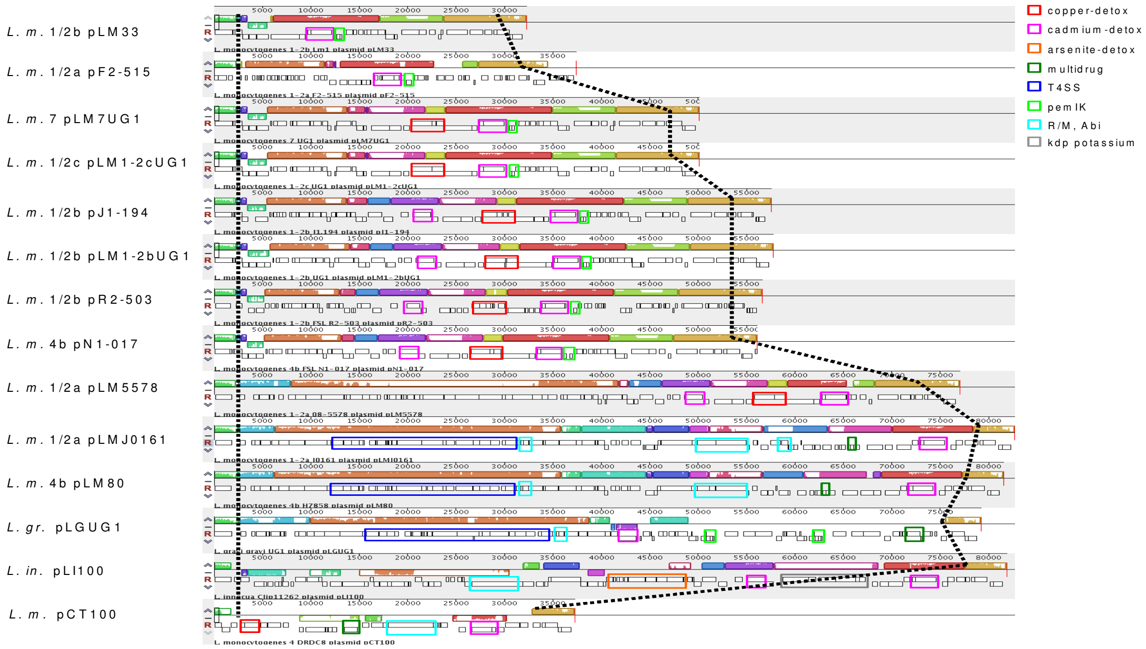


Figure 15. Mauve alignment of listerial plasmids including functional modules (p2). All plasmids share a common theta-type replicon (left and right of the dotted lines, respectively).

Thus, the ancestral relationship of all of the *Listeria* plasmids could easily be visualized, together with the various degrees of conservation resulting from multiple indels (p2). Subsequently, functional modules were clustered to reveal the impact of these adaptations i.e. on the ability of the host to resist heavy metal or oxidative stress, as well as invasion of foreign DNA. A similar approach was undertaken for the depiction of listerial chromosomes, but the higher degree of similarity of these replicons resulted in a figure bearing only little informative value.

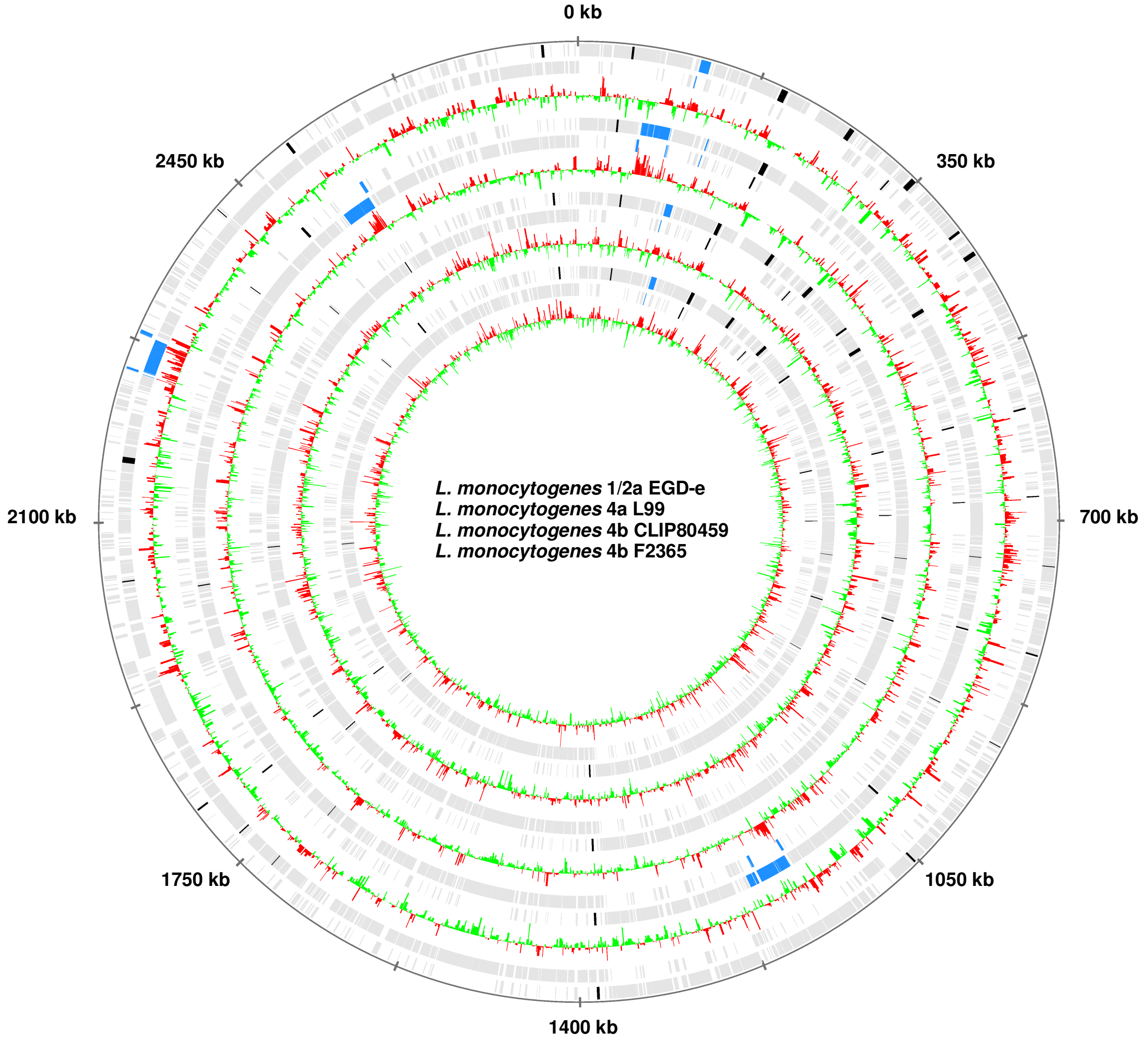


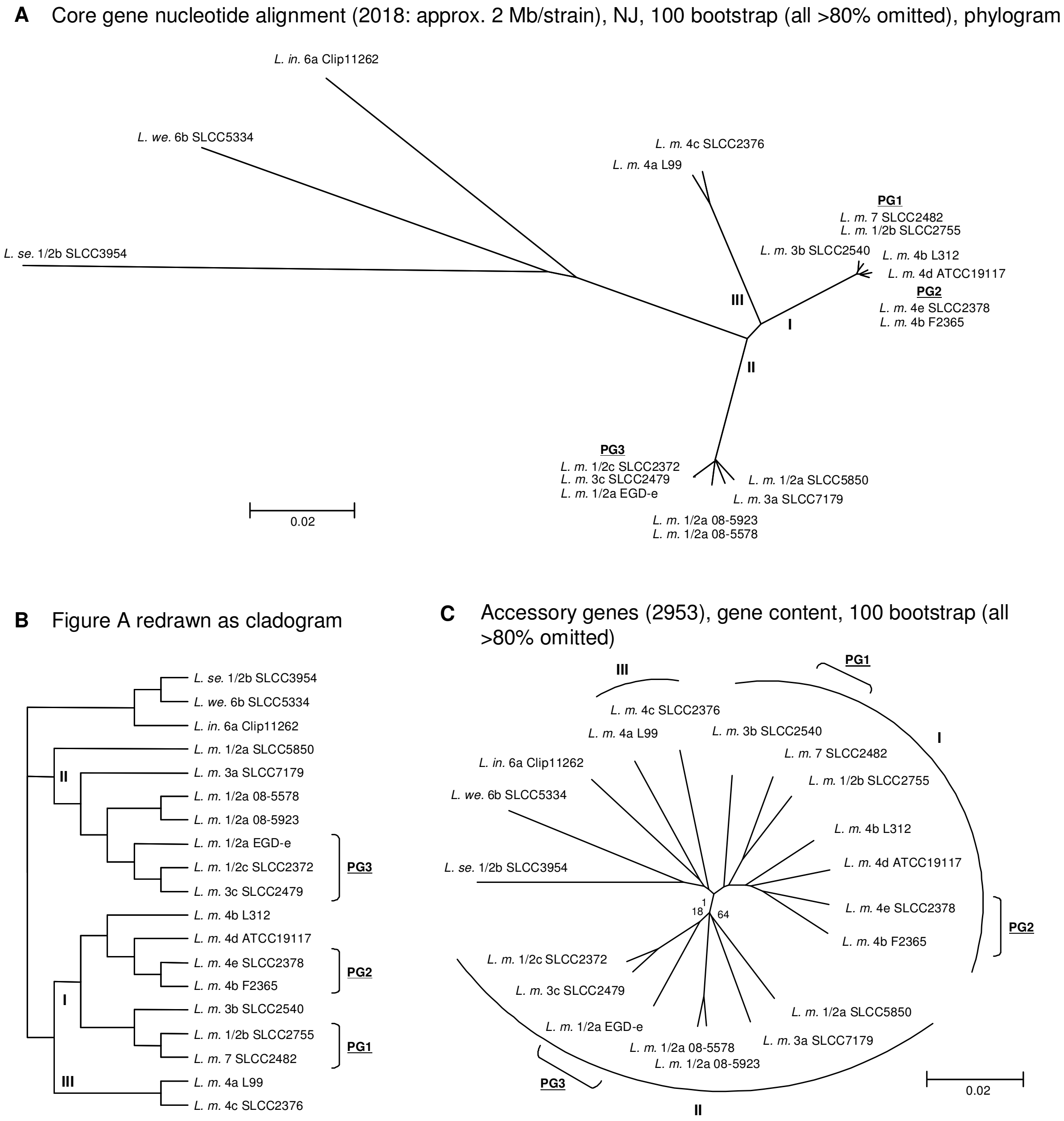
Figure 16. GenomeViz depiction of transcriptome data showing coding sequences (gray), phage genes (blue) and virulence genes (black), as well as their degree of intracellular regulation (red=up, green=down) inside four listerial strains from outside to inside (p6).

GenomeViz is more suitable for the display of similar genomes due to its ability to include various types of discriminating data. Thus, the intracellular up-regulation of prophage genes in listerial chromosomes was immediately recognizable (p6). The software package GenomeViz is suited for the display and comparison of similar genomes because of its ability to include varying types of data and for plotting this data in common formats. By zooming onto the areas of difference or similarities, information relating to individual genes can now be represented and analysed in greater detail. Similarly, any set of data e.g. SNPs, can now be mapped onto these genomes and analysed in a similar fashion.

3.1.7 Creating and assessing phylogenies

A further option is to depict relations between organisms using taxonomic trees. Here, minor sequence diversion (phylogenetic) or the presence/absence of genes approximately representing functions that are frequently the result of horizontal gene transfer (phenotypic) is possible.

In order to compare the difference between these approaches, multiple trees were generated to assess the evolutionary distances between the sequenced listerial chromosomes (p7).


A Core gene nucleotide alignment (2018: approx. 2 Mb/strain), NJ, 100 bootstrap (all >80% omitted), phylogram

B Figure A redrawn as cladogram

C Accessory genes (2953), gene content, 100 bootstrap (all >80% omitted)

Figure 17. Phylogenies of listerial chromosomes displaying (A) a phylogram of a 2 Mb alignment of core genes, (B) a cladogram of the same data to allow the recognition of branchings, (C) a phylogram of the accessory gene content (p7). Lineages of *L. monocytogenes* are denoted by roman numbers (I-III) and only bootstrap support values below 80% are acknowledged. Phylogenomic groups (PG) mark strains that are closely related according to their core genome SNPs.

In order to visualize a representative set of SNPs between strains, which could roughly represent temporal divergence of strains when assuming comparable rates of evolutionary adaptation, nucleotide sequences of all core genes were included. Thus, a whole-genome cluster matrix of 19 listerial strains, harbouring the most representative pan-genomic distribution of genes of genus *Listeria* currently available, was created with GECO. In order to limit the comparison to completely conserved orthologs, all non-core clusters or clusters with paralogous genes were removed from this list. Identifiers of all remaining core genes of the strain *Listeria monocytogenes* EGD-e were extracted and nucleotide sequences of homologues inside all strains were exported by GECO, aligned, and transformed to phylogenomic trees using multiple algorithms. These resulted in identical topologies with consistent bootstrap support values of >80%, implying a high stability of the observed branching.

Another tree was built based only on the distribution (presence/absence) of accessory genes to identify the impact of indels, which represent gain and loss of specific functions often based on horizontal gene transfer. A whole-genome matrix of all strains was extracted from GECO and the presence or absence of a homologue inside a cluster was transformed to “1” and “0”, respectively. This matrix was supplied to the GeneContent software and a tree was inferred, which revealed low bootstrap support at the central genus junction indicating indels of early accessory genes that lead to contradictory topologies.

The resulting trees showed a largely conserved topology similar to known MLST phylogenies based on housekeeping genes, but differed considerably when considering branch lengths. Long common histories of strains of each lineage as displayed by the core-genome tree have resulted in only a small number of conserved lineage-specific genes. Gene content furthermore implies relatively homogeneous distances between strains of different lineages of *L. monocytogenes* and obligate apathogenic species, implying a lack of conservation of accessory genes often associated with horizontal gene transfer. Thus, the observed differences in phenotype of apathogenic species and lineages of *L. monocytogenes* seem to result mostly from adaptation due to SNPs and a relatively small number of gene-scale indels. In order to assess the impact of groups of genes on tree topology the respective indels could be easily removed from the underlying matrices. This revealed a slightly obfuscating influence of prophages and mobile genetic elements, because of the insertion of multiple genes in one evolutionary event, leading to a bias in the "true" phylogenetic signal represented by the SNP-based core genome tree.

While the steps included in this analysis could also be facilitated in a more manual fashion using a combination of other tools, the ability of GECO to instantly reveal pan-genomic distributions of genes, which can be easily filtered for clusters of interest followed by extraction of the respective nucleotide or amino acids, greatly simplifies the application.

3.1.8 Pan-genome distribution and prediction of total size

In order to understand the diversity present in a pan-genome and predict its total size, the extent of conservation can be depicted in the form of graphical plots.

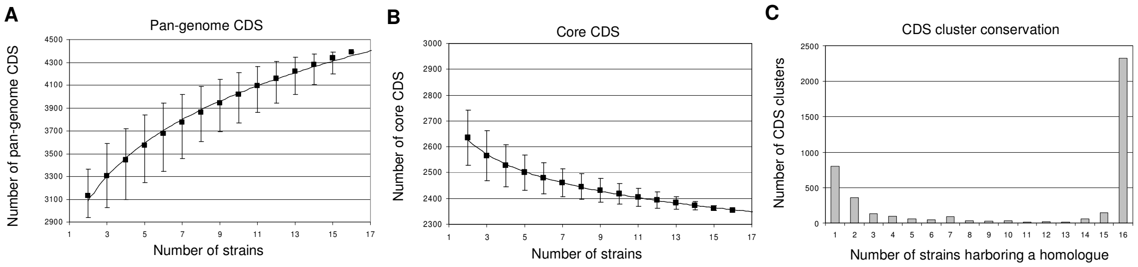


Figure 18. Pan-genome analysis showing (A) pan-genome size after consecutive addition of strains, (B) mutually conserved core genes, (C) total degree of conservation in 1-16 strains (p7).

A module was added to GECO that consecutively adds homology distributions of 1-x strains to assess changes in the number of pan-genome and core genes within an increasing number of included strains. For the species *L. monocytogenes* pan-genome, chromosomes of 16 strains were added 10000 times without replacement and in a randomized order. The size of the total species pan-genome is extrapolated by power law regression that was previously described to reflect the rate of discovery of new genes in a pan-genome based on currently available data [98]. Furthermore, lists produced by GECO can be filtered to reveal the distribution of genes in the current pan-genome by considering the exact degree of conservation ranging from core to strain-specific genes (GECO whole-genome cluster matrix).

The results indicate that *L. monocytogenes* is a highly conserved species with a core genome comprising approximately 75% of each strain, and is thus similar to species *Streptococcus pyogenes, Streptococcus agalactiae*, and *Haemophilus influenzae*. An abundant core genome seems to be the hallmark of specialized parasites as well as primarily saprotrophic commensal bacteria such as listeriae*,* which imply that these genomes are highly adapted to the respective growth niches. Limited natural competence, which has been described for the genus [66], and the presence of restriction modification systems and adaptive immune systems in some strains of the pan-genome support the observed stability. Nevertheless, it is capable of adapting and integrating horizontally transferred genes. The pan-genome prediction of 6000 genes after 100 strains have been sequenced, i.e. twice as large as the genome of each individual strain, indicates that the pan-genome of *Listeria monocytogenes* is larger than previously suggested [88,89].

Thus, implementation of this functionality in GECO reduces the amount of time necessary to obtain an overview of pan-genomic populations to a matter of minutes.

3.1.9 Delineating the impact of duplication and horizontal gene transfer

Gene duplication and horizontal gene transfer are crucial processes for the evolution of genomes leading to variation of gene dosage and the acquisition of new functions [99].

Putative duplicates can be detected based on GECO homology matrices, which automatically join paralogous genes of one replicon inside a cluster. The ages of the respective duplications are related to the observed degrees of sequence similarity, since duplicates are frequently subject to relaxed selective pressure and heightened mutation. Different measures of homology used for the creation of matrices can thus differentiate between predicted ages of duplication events. HGT is recognizable as a deviation of the sequence composition of a gene from that of the remaining genes of the host chromosome (e.g. using SIGI-HMM). While these methods result in reasonable approximations, highly similar genes horizontally transferred from a related strain can still be misclassified as duplication. Furthermore, the codon-usage of HGT genes is subsequently adapted to that of the host strain, thereby masking the origin of older insertions.

In order to delineate the impact on the evolution of species *L. monocytogenes*, these processes were identified and correlated with putative functions (p6). Most duplications in listerial genomes seem to be ancient since they predate speciation and the number of highly similar genes was found to be low. Classification of duplicated genes by metabolic pathways showed that carbohydrate metabolism was primarily affected, which probably represents an adaptation of the last common ancestor of listeriae to nutrients available in its growth environments. Putative horizontally transferred genes were rarely classifiable, implying that these relatively recent events had little apparent impact on housekeeping, but may have supplementary function in promoting growth and/or virulence in specific niches.

3.1.10 Comparative identification of hyperdynamic hotspots and mobile genetic elements

The colinearity of chromosomes of *L. monocytogenes* permitted the use of an elegant method to identify regions of dissimilarity analogous to a previously published approach based on phylogenetic conservation [38] (p4, p7). Apart from mobile genetic elements and prophages, this analysis also uncovered hyperdynamic loci frequently associated with multiple insertion elements, suggesting reduced purifying selection at these sites, which create hotspots for the insertion of additional content.

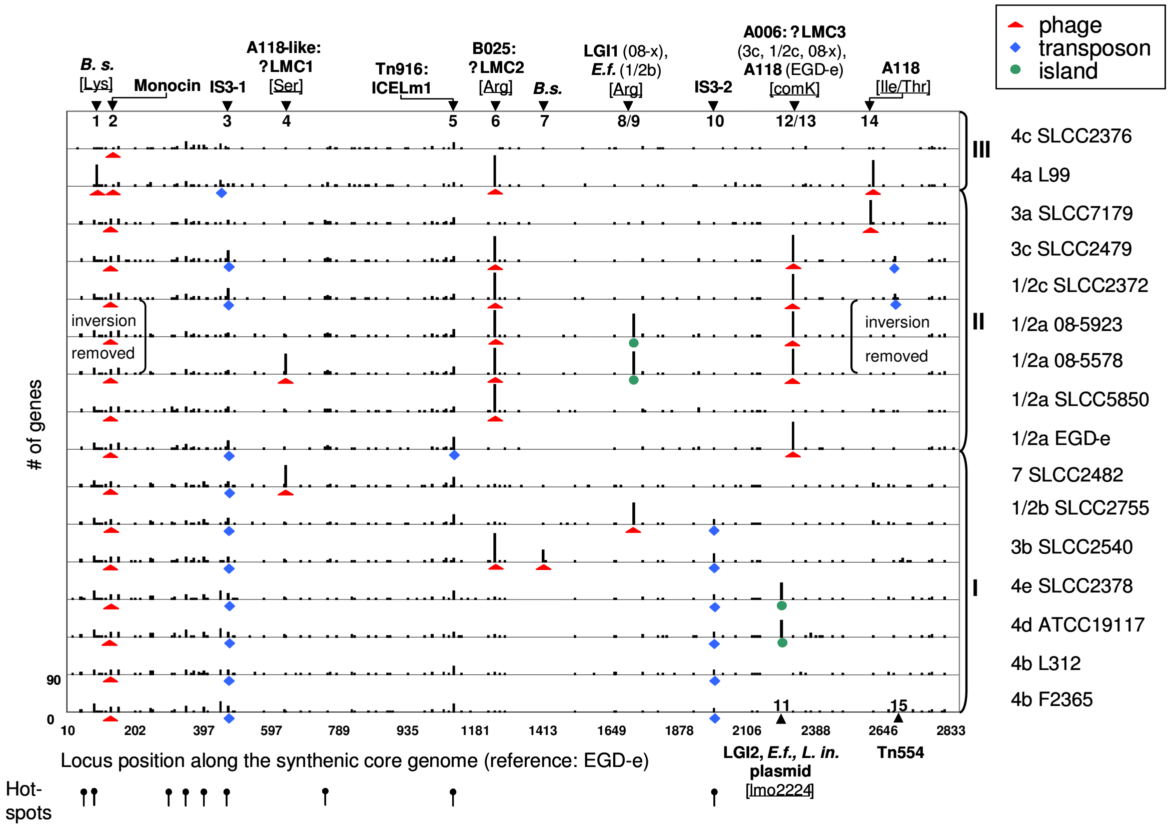


Figure 19. Bar chart depiction of non-core regions of 16 chromosomes of *L. monocytogenes* relative to the layout of reference strains EGD-e (p6). An inversion found in strains 08-5923 and 08-5578 was removed previously. Identified mobile elements and their putative target genes are noted if applicable, as well as hyperdynamic hotspots.

A synteny homology matrix exported from GECO identifies commonly conserved core genes, as well as genes located in between. All core genes showing a break in the synteny (translocation, inversion) relative to a reference strain are removed from the pool. The final result is the number of genes located between syntenic core genes, which can be plotted as a bar chart for each individual strain using standard spreadsheet software. The exact borders of mobile genetic elements were later refined based on annotation, deviation of GC-content and comparative analysis with sequenced bacteriophages and strains of genus *Listeria* using the GECO visualization interface.

Within the species *L. monocytogenes,* insertions of foreign DNA and other modifications were highly associated with nine highly variable regions, which contain one fourth of all accessory genes. Apart from well-known major pathogenicity factors such as internalin A/B and Listeriolysin S, these loci also show a large number of transposases, restriction-modification systems and surface-associated genes, which represent numerous separate events of insertion, deletion, and recombination. Mobile genetic elements are rare in the species (three transposons, two islands), but display some adaptive potential as seen by the integration of genes for heavy metal resistance, as well as cell-surface-associated proteins. Furthermore, nine different prophages were detected that seem to bear a function in intracellular survival. These encompass one third of the accessory genome. Strains of serogroup 4 were previously described to be resistant to phages due to a differing cell wall composition [100]. This is supported by comparative data on 4b strains, but excludes strain 4a L99, that harbours three prophages.

This strategy can identify putative single insertional events implicating mobile elements (indicated by high peaks in a subset of strains), as well as hyperdynamic hotspots (indicated by clustered sets of lower peaks in all strains. Nonetheless, distinction between insertion and deletion necessary for the reliable detection of horizontal gene transfer is not a trivial task and requires manual inspection. A gene present in only one strain of a species may also have been vertically inherited and lost in an ancestor of all other strains of the species. While the phylogenetic distribution is a helpful first indicator, delineation of these processes necessitates an inspection of homologous regions in strains showing a putative deletion for remnant nucleotide sequences of the missing gene (junction sequence). This task can be accelerated by using the sequence extraction interface included in GECO, which retrieves any number of nucleotides from a replicon based on the location, followed by comparison with multiple sequence alignment software.

3.1.11 Finding and describing genes which support taxonomic or phenotypical divisions

Identification of genes that are determinants for phylogenetic or phenotypic groups is a significant step of pan-genomic analysis that is simplified when using GECO homology matrices (p1, p6, p7). Columns bearing strains of the respective taxa (e.g. lineages, serotypes, pathogenic/apathogenic) can be filtered using any spreadsheet software to reveal specifically conserved indels.

In order to gain an overview of the general functional category of genes, these can be classified according to sequence-clusters collected by public databases (e.g. COG, KEGG). The presence of a surface-associated domain that may indicate a protein-interaction frequently associated with virulence or pathogenicity can also be predicted based on sequence similarity (e.g. using Augur).

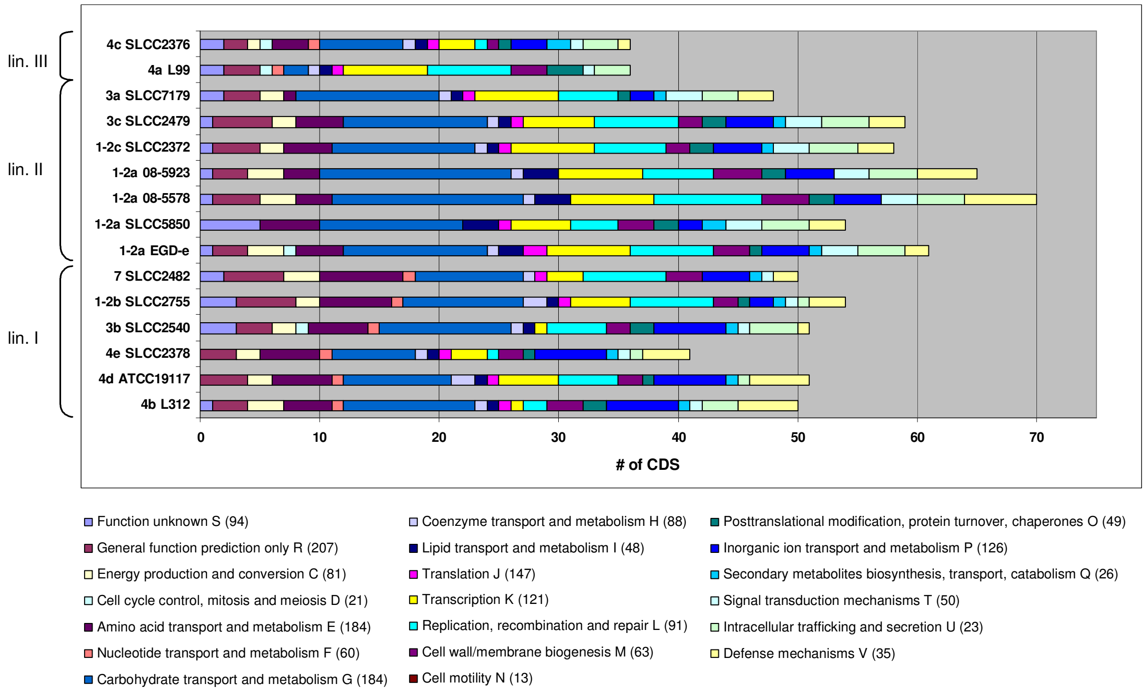


Figure 20. Functional classification of accessory genes according to the COG database can reveal the direction of gene-scale adaptation of different strains (core genes are denoted in brackets inside the legend).

Using these strategies, 45 genes were found to be conserved in pathogenic lineages I/II and absent from less pathogenic lineage III, and frequently indicated in functions which may be advantageous for survival inside a host (e.g. metabolism, stress response, invasion), supporting a mostly reductive evolution of strains of lineage III.

We furthermore identified 33 lineage-III-specific, 22 lineage-II-specific and 14 lineage-I-specific core genes. Interestingly, the distinct lineage core repertoire of lineage II includes predominantly genes involved in carbohydrate metabolism organized in three operon-like islands, while those of lineages I and III mainly consist of hypothetical and surface-associated proteins scattered over the respective chromosomes. This indicates that ancestral strains of lineages I and III diverged from lineage II by gene loss related to carbohydrate metabolism and gain of surface-associated genes serving different needs of interaction with the environment.

Serogroups or –types rarely displayed specifically conserved core genes, with the exception of serogroup 4, which were already described to be responsible for differences in teichoic acid composition [101]. Thus, observed variable antigenes of other serogroups or –types either result from minor changes inside coding genes, from differences located in intergenic regions, or a combination thereof.

Similarly, no specific core indels could be found for strains of serotypes most commonly associated with human listeriosis (1/2abc, 4b), implying multiple varying genes or minor mutations to be effectors of virulence and pathogenicity.

Analysis of surface-associated proteins displayed conserved lineage-backbones with frequent strain-specific adaptations. Internalin-like proteins revealed a relatively homogeneous distribution in lineage I. Lineage II and III show more profound variation implying complex lifestyles for these strains. A total of nine new putative internalins were identified in this study, which represent future candidate genes for research into virulence factors.

3.1.12 Bioinformatic identification of small non-coding regulatory RNAs

While wet-lab technologies such as microarray hybridization or RNA-Seq are most reliable in discovery of new sRNAs, the presence of known sRNA sequences can be identified in other strains using bioinformatic approaches. The sRNAdb software was designed to allow identification of sRNAs based on sequence homology (BLASTN) to published, predicted and experimentally verified sRNAs (p5).

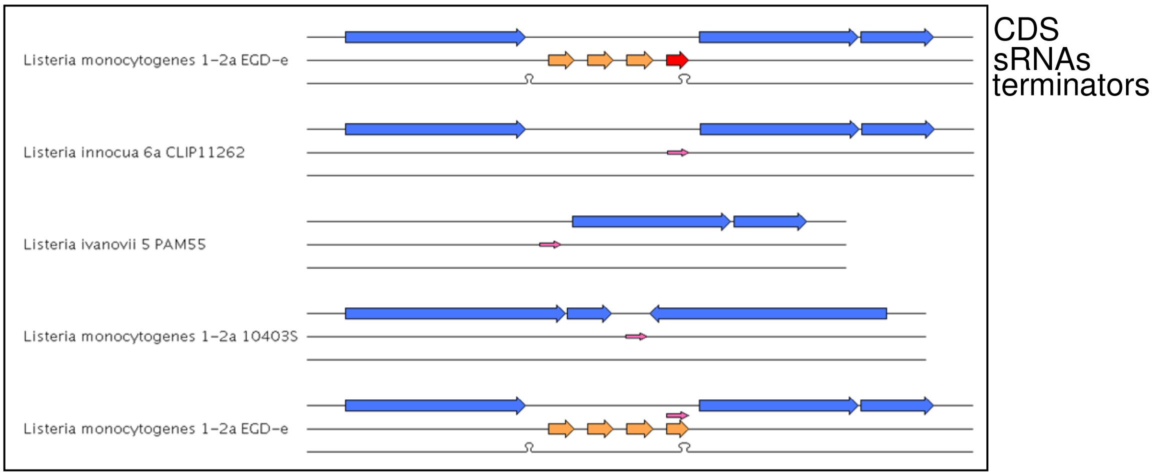

Listeria monocytogenes 1-2a EGD-e
Listeria innocua 6a CLIP11262
Listeria ivanovii 5 PAM55
Listeria monocytogenes 1-2a 10403S
Listeria monocytogenes 1-2a EGD-e
CDS
sRNAs
terminators

Figure 21. sRNAdb comparative regional visualization of sRNAs including genomic neighbourhood. Comparative display of loci surrounding homologs (thin red arrow) of a reference sRNA (broad red arrow) including other known sRNAs (orange arrow), protein-coding genes (blue arrow) and terminators.

The software developed can be locally installed and focuses on the ability to search for large numbers of sRNAs inside whole genomes with different degrees of detail. These include (i) a query interface for tabular data of known sRNA, (ii) an alignment interface to extract a matrix of presence/absence of multiple sRNAs in multiple replicons based on a defined measure of homology, or to identify differences between sRNAs down to the level of SNPs, and (iii) a regional visualization interface to find and compare sRNAs including the surrounding locus of genes and associated promoters and terminators.

Complete genome sequences of listerial chromosomes and plasmids were searched with sRNAdb for all experimentally verified sRNAs of strain *L. monocytogenes* EGD-e. This analysis identified a high degree of conservation of chromosomal sRNAs and the presence of additional sRNAs on listerial plasmids, which may therefore serve as sRNA transfer vectors (p2, p7). Nonetheless, difference in the distribution of *trans*-encoded RNA was seen on the chromosomes, hinting at a diverse range of regulatory adaptations. This is in contrast to highly conserved *cis*-regulatory elements that were considered to be present in the last common ancestor of *L. monocytogenes*.

3.1.13 Identification and visualization of CRISPR adaptive immune systems

CRISPR systems comprise of multiple Cas (CRISPR associated) genes, which encode for proteins that process exogenous DNA for insertion into CRISPR arrays (spacer), which when transcribed subsequently silence the incoming homologous DNA. In order to assess the impact of CRISPR systems in the genus *Listeria*, the homology and distribution of Cas genes and arrays were identified and compared.

CRISPR spacer/repeat-arrays were identified with available software (e.g. PILER-CR, CRT), complemented by BLASTN searches of consensus repeat sequences and manual assessment to remove false positive predictions. In order to identify possible targets, resulting spacers of CRISPR arrays were compared to published sequence data using BLAST. Cas genes were identified by sequence similarity to previously annotated Cas genes in the NCBI or Pfam databases, and by searching the annotation of genes located in the vicinity of CRISPR arrays using GECO.

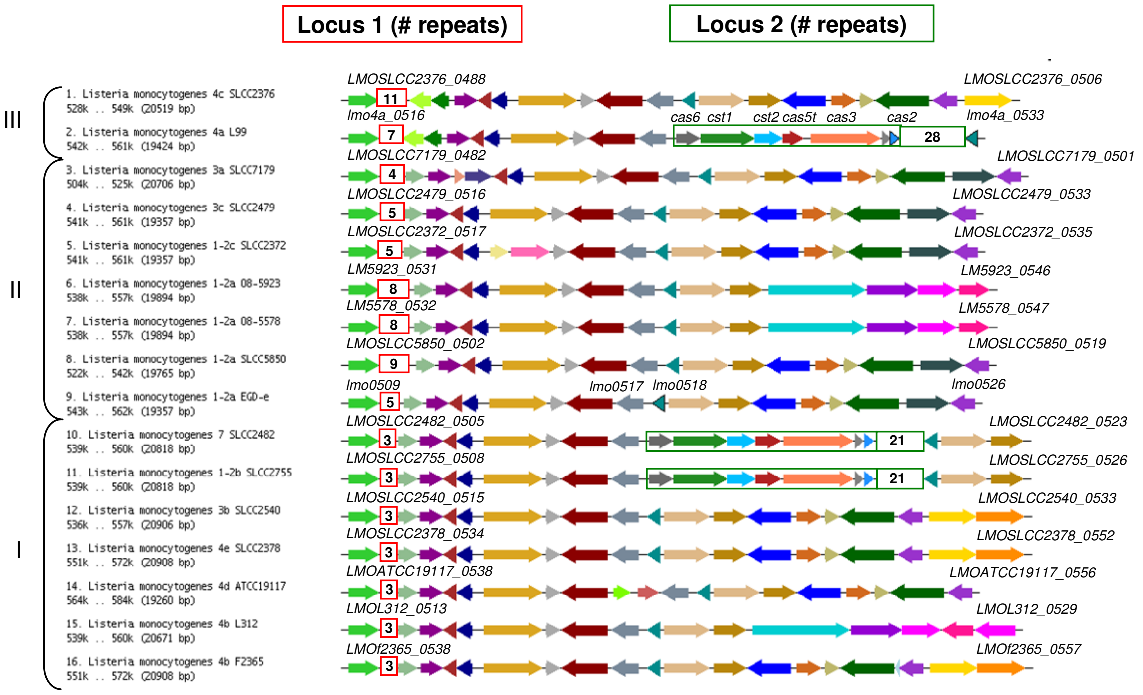
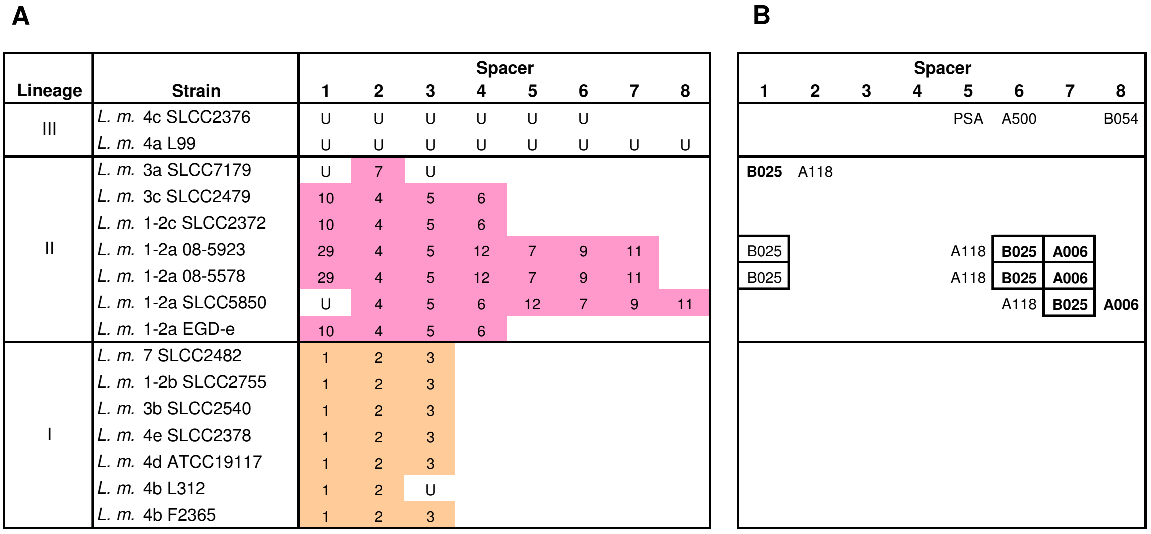


Figure 22. GECO homology depiction of genes of a chromosomal region of 16 strains of *L. monocytogenes* strains harbouring two identified CRISPR arrays (number of spacers depicted inside box) (p7). Locus 1 lost its associated *cas* genes and is likely dysfunctional. Locus 2 was putatively inserted into an ancestor of strains *L. m.* 4a L99, 7 SLCC2482 and 1/2b SLCC2755, since no remnant sequence of this module could be identified in other strains.

A detailed analysis and visualization of these data necessitated the combination of homology and positional information. Thus, a Java-based program called BlastclustToMatrix was written to combine the clusters created from a homology analysis of spacers by BLASTCLUST with the order of spacers in the respective replicons. The result of this analysis is a tab-delimited spreadsheet, which permits the identification of specific and homologous spacers, as well as their position in the respective CRISPR arrays.



A

Lineage	Strain	Spacer 1	2	3	4	5	6	7	8
III	*L. m.* 4c SLCC2376	U	U	U	U	U	U		
	L. m. 4a L99	U	U	U	U	U	U	U	U
II	*L. m.* 3a SLCC7179	U	7	U					
	L. m. 3c SLCC2479	10	4	5	6				
	L. m. 1-2c SLCC2372	10	4	5	6				
	L. m. 1-2a 08-5923	29	4	5	12	7	9	11	
	L. m. 1-2a 08-5578	29	4	5	12	7	9	11	
	L. m. 1-2a SLCC5850	U	4	5	6	12	7	9	11
	L. m. 1-2a EGD-e	10	4	5	6				
I	*L. m.* 7 SLCC2482	1	2	3					
	L. m. 1-2b SLCC2755	1	2	3					
	L. m. 3b SLCC2540	1	2	3					
	L. m. 4e SLCC2378	1	2	3					
	L. m. 4d ATCC19117	1	2	3					
	L. m. 4b L312	1	2	U					
	L. m. 4b F2365	1	2	3					

B

Spacer 1	2	3	4	5	6	7	8
				PSA	A500		B054
B025	A118						
B025				A118	**B025**	**A006**	
B025				A118	**B025**	**A006**	
					A118	**B025**	**A006**

Figure 23. Extract of CRISPR array locus 1 of 16 strains of *L. monocytogenes* (p7). (A) Depiction of identical spacers with the same numbers and unique spacers with "U" sorted by distance to Cas genes (column 1=youngest, 8=oldest). Modules bearing related spacers were colour coded. (B) Putative target of CRISPR spacers to be silenced with a maximum of one nucleotide mismatch (perfect matches = bold). Spacers targeting prophages located in the same chromosome are shown with a border.

Analysis of 16 *L. monocytogenes* strains revealed, that five strains exhibited putatively functional CRISPR/Cas systems of two subtypes. The heterogeneous distribution, which was not correlated to the presence of prophages in the respective strains, suggests a supplementary function for these systems. All identifiable spacers targeted known bacteriophages, including 11 instances of spacers targeting prophages located in the same chromosome, which may lead to active silencing. The only commonly present array (dysfunctional locus 1) displayed a clear separation between spacers of all lineages indicating that (i) the last common ancestor of *L. monocytogenes* harboured a CRISPR system, and that (ii) it was lost in a common ancestor of lineage I and also during multiple independent events in strains comprising lineages II and III. Nine strains were found to possess other restriction modification systems, which could also contribute to degradation of invading DNA. Nonetheless, for five strains no identifiable defence against bacteriophages was detected. Therefore, observed chromosomal stability of listeriae may result from (i) limited natural competence, (ii) the presence of yet unknown genes serving a purifying function, or (iii) a largely destructive effect of HGT that may disrupt regulatory structures or increase the energetic burden leading to a loss of fitness for the respective highly adapted strain.

Thus, the type and degree of relatedness of CRISPR arrays between various strains, putative age of spacers, target bacteriophages, and possible silencing of prophages already present in the chromosomes could be recognized. The diversity of spacers inside these loci may also be useful for future typing systems. Use of the described prediction and analysis strategy relies on open-source software that was partially developed for publications underlying this thesis (GECO, BlastclustToMatrix).

3.1.14 Problems encountered and possible solutions to further streamline analyses

Presently, comparative annotation with GECO incorporates multiple manual steps including visual interpretation of syntenic regions, semi-automatic connections to third-party databases (Pfam, STRING, COG) for spurious matches, assembly of information within external spreadsheet software, and reliance on an unpublished software component to update annotation and export flatfiles. Replacement of these with concise web-based interfaces with precomputed data inside GECO will improve the annotation process.

The lengthy gap closure phases pose another problem, since they necessitate constant transfer of annotation to the most recent sequence version. While these data can already be mapped semi-automatically using available GECO annotation matrix functions, the process should be extended to distinguish between current and former contig borders to improve sensitivity and specificity.

Delineation between insertion and deletion is currently based on extensive manual assessment to identify putative loci by phylogenetic distribution and deviating sequence composition, and to compare resulting nucleotide alignments to recognize junction sequences. Automation of these steps will be highly useful to permit efficient analysis of large datasets.

Since some homology matrices (e.g. synteny mode) demand large processing capacities from the webserver due to the use of live BLAST searches, GECO is unable to support many large numbers of concurrent users. Future plans include the support of distributed computing solutions to relieve the webserver.

3.2 Summaries of Publications

The following text offers a brief summary and complete of results of publications employing the software packages described above.

3.2.1 Complete genome sequence of *Listeria seeligeri*, a non-pathogenic member of the genus *Listeria*

3.2.1.1 Methods and Contribution

Annotation: The author was a lead participant in annotation according to principles outlined above.

Bioinformatic analyses: The author performed the phylogenomic analysis using MAVID and MEGA, functional mapping by COG classification using Augur and comparative analysis with related species with the GECO analysis server.

Draft: The author drafted the manuscript.

3.2.1.2 Summary

Analysis of non-pathogenic type strain *L. seeligeri* SLCC3954 isolated from soil identified common loss of genes related to intracellular survival combined with limited gene gain focused on metabolic pathways relevant for a saprophytical lifestyle, indicating a common evolutionary path towards genome shrinkage already described for other apathogenic species. Thus, an ancestor of species *L. monocytogenes* may have consecutively lost determinants for pathogenicity thereby spawning other apathogenic species.

3.2.2 Comparative analysis of plasmids in the genus *Listeria*

3.2.2.1 Methods and Contribution

Genome finishing: The author conducted all bioinformatic steps involved in the complete sequencing of plasmid pLM7UG1 from a genomic 454 pyrosequencing run as already described.

Annotation: The author comparatively (re-)annotated 14 plasmids according to principles outlined above.

Bioinformatic analyses: The author comparatively analysed and visualized the data including global synteny analysis using Mauve, creation of a phylogenetic tree based on the replication initiation protein using Clustalw and Dendroscope, as well as comparative analyses using GECO and installation of a public server (http://bioinfo.mikrobio.med.uni-giessen.de/geco2plasmids).

Draft: The author drafted the manuscript.

3.2.2.2 Summary

In order to assess the evolutionary origin and functional contribution of plasmids in genus *Listeria*, four new plasmids were sequenced to closure and annotated, thus extending the public data available by the first plasmids of *L. monocytogenes* serotypes 1/2c and 7 as well as the first plasmid of species *L. grayi*. These were comparatively analyzed in conjunction with 10 additional publically available listerial plasmid sequences. Prior to the incorporation into the analysis a complete reannotation of these plasmids was performed in order to remove artefacts based on different gene calling strategies.

All plasmids displayed a common theta replicon-type consisting of replication and partitioning genes (*repA-C*) and a gene encoding an error-prone translesion DNA polymerase IV that probably functions as an adaptive mutator generating genetic diversity. Common to all plasmids was also the presence of a diverse range of transposition-related genes. Listerial plasmids showed further homologies indicating ancestral relations and ongoing horizontal gene transfer towards broad-host range plasmids of other Firmicutes like *Streptococcus*, *Enterococcus,* and *Bacillus*, which are typical inhabitants of the gut microbiota, thus implying the eukaryotic host, but also soil and water as niches for growth and transfer of these elements.

Interestingly, four putative sRNAs previously identified in the chromosome of *L. monocytogenes* EGD-e were also present on various plasmids, implying the dissemination of non-coding regulatory RNAs by plasmids.

A number of genes were identified which may serve a role in defence against bacteriophages, oxidative stress response, multidrug efflux, and heavy metal detoxification which can increase bacterial fitness in the environment (e.g. soil and food processing facilities) or the eukaryotic host and may result from selective pressures prevalent to these niches.

Taken together, the vast majority of plasmid-borne content represents supplemental functionality that may induce a non-effective energetic burden when constantly implemented in the chromosome, thus facilitating survival under stress conditions.

3.2.3 Genome-wide identification of small RNAs in the opportunistic pathogen *Enterococcus faecalis*

3.2.3.1 Methods and Contribution

Bioinformatic analyses: Creation and application of a software (Crumbs) to automatically design intergenic tiling-arrays for the detection of sRNAs optimized for optimal coverage according to the available size of the chip.

3.2.3.2 Summary

The first experimental genome-wide identification of sRNAs in the Gram positive opportunistic pathogen *E. faecalis* V583 was facilitated based on tiling microarray analysis. Thus, a software tool named CRUMBS was written in order to select probes for custom microarray designs.

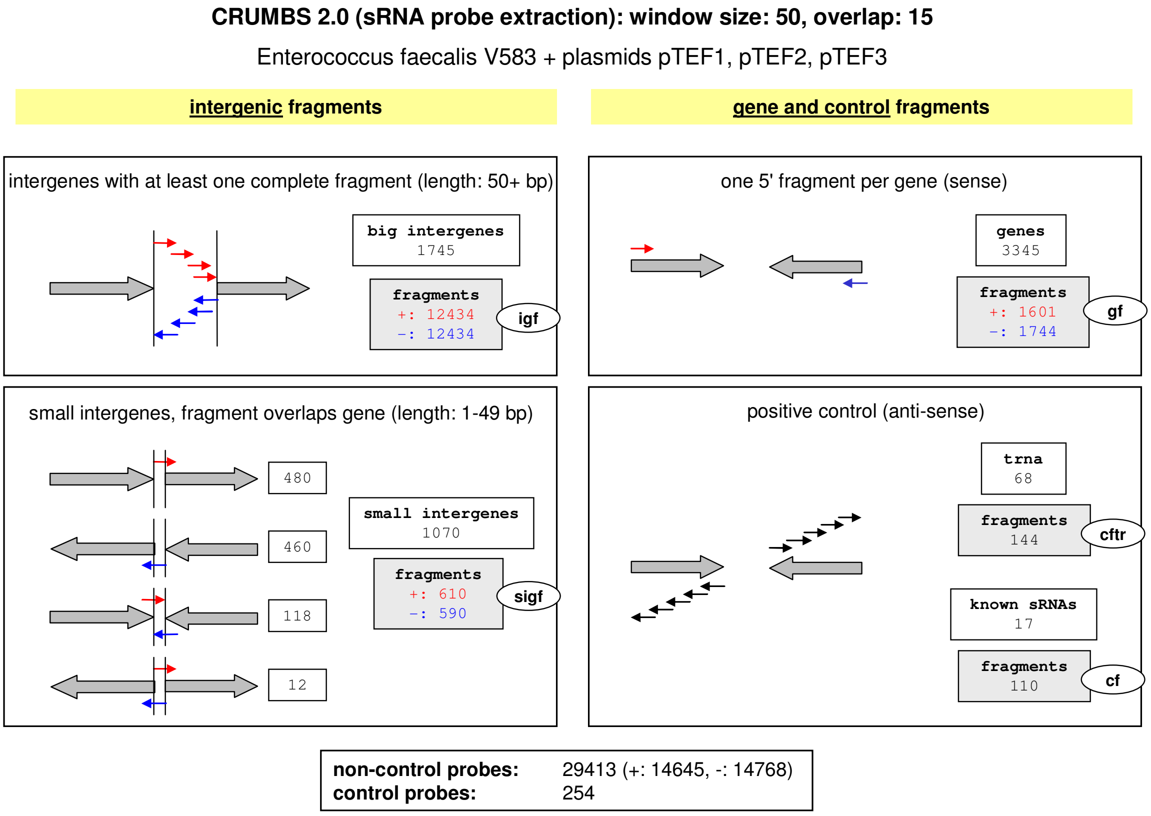

CRUMBS 2.0 (sRNA probe extraction): window size: 50, overlap: 15
Enterococcus faecalis V583 + plasmids pTEF1, pTEF2, pTEF3
intergenic fragments
gene and control fragments
intergenes with at least one complete fragment (length: 50+ bp)
big intergenes 1745
fragments +: 12434 -: 12434
igf
one 5' fragment per gene (sense)
genes 3345
fragments +: 1601 -: 1744
gf
small intergenes, fragment overlaps gene (length: 1-49 bp)
480
460
118
12
small intergenes 1070
fragments +: 610 -: 590
sigf
positive control (anti-sense)
trna 68
fragments 144
cftr
known sRNAs 17
fragments 110
cf
non-control probes: 29413 (+: 14645, -: 14768)
control probes: 254

Figure 24. CRUMBS schematic overview of probe selection intended for the detection of intergenic sRNA transcripts and positive control.

The selection process attempts to pick overlapping probes (fragments) from intergenic regions while avoiding including mRNA of protein-coding genes. Therefore, only those probes are considered valid, which do not overlap with a protein-coding gene in the anti-sense direction, which necessitates a different handling of intergenic regions that cannot accommodate one probe without overlap. Additionally, a 5'-sense overlap of sRNAs with protein-coding genes was also included. Positive control probes can be selected from known sRNA, tRNA, and rRNA genes. Size and overlap of probes are variable and can accommodate different types of microarray chips. Since the software relies on public flatfiles that include gene locations to automatically select probes, it can be used to rapidly design custom microarrays for any prokaryote.

Thus, 11 sRNAs could be successfully characterized in *Enterococcus faecalis* V583, thereby providing an impetus to the understanding of gene regulation in this important human pathogen.

3.2.4 Complete sequences of plasmids from the hemolytic-uremic syndrome-associated *Escherichia coli* strain HUSEC41

3.2.4.1 Methods and Contribution

Genome finishing: The author conducted all bioinformatic steps involved in the complete sequencing of four plasmids of *E. coli* HUSEC41 from a genomic 454 pyrosequencing run.

Annotation: The author performed the majority of the comparative annotation processes for these four plasmids according to principles outlined above.

Bioinformatic analyses: The author comparatively analysed the data in conjunction with reference replicons using Mauve and GECO.

Draft: The author drafted the manuscript.

3.2.4.2 Summary

Four plasmids from a historical enteroaggregative Shiga toxin-producing Escherichia coli (HUSEC) serotype O104:H4 strain, HUSEC41/01-09591, isolated in 2001 in Germany, were completely sequenced and annotated. Two of these were found to be mobilizable and homologies to published plasmids were established in brief. Among the encoded functions are resistances towards streptomycin and sulphonamides. Future sequencings and analyses are supposed to reveal the degree of relation to the *E. coli* HUSEC strain responsible for an outbreak in Germany in 2011.

3.2.5 sRNAdb: A small non-coding RNA database for gram-positive bacteria

3.2.5.1 Methods and Contribution

Design and implementation: The author conceived and designed major aspects of the sRNAdb database, application and interfaces, and implemented core structures in JAVA.

Draft: The author assisted in drafting the manuscript.

3.2.5.2 Summary

In order to identify and compare sRNAs inside prokaryotic genomes based on sequence similarity, sRNAdb was developed. The main advantage of this client/server based software is its ability to rapidly detect and visualize multiple sRNAs within complete replicons including the genomic neighbourhood of protein-coding genes, promoters and terminators to clearly recognize differences between homologous regions. The public server contains all published experimentally verified sRNAs of gram-positive bacteria, as well as *in silico* predictions. sRNAdb can be locally installed to include user-generated data.

3.2.6 Comparative genomics and transcriptomics of lineages I, II, and III strains of *Listeria monocytogenes*

3.2.6.1 Methods and Contribution

Annotation: The author was a major participant in the annotation of *L. monocytogenes* 4a L99 according to principles outlined above.

Bioinformatic analyses: The author comparatively analysed the data in conjunction with reference replicons using GECO and created figures of several loci (monocin, internalin modules GHE-C2DE), performed the predictions of horizontally transferred genes using SIGI-HMM, identified putative duplications using several homology analyses and correlated these data with metabolic pathway categories, and assisted in the creation of two GenomeViz figures depicting genome- and transcriptome-based similarities.

Draft: The author assisted in drafting the manuscript.

3.2.6.2 Summary

The main objective of this study was research on genome and transcriptome differences between phylogenetic lineages of species *L. monocytogenes,* especially of the less virulent lineage III. We completely sequenced virulence-attenuated strain 4a L99, being the first strain of lineage III and serotype 4a, as well as a second strain of serotype 4b (CLIP80459) and compared these to public data.

Loss and divergence of surface- and virulence-associated genes was identified as a recurrent pattern for strains of lineage III.

Correlation of duplications and HGT with metabolic pathways revealed that the former may have been used by an ancestor of all listeriae to extend carbohydrate metabolism, while the latter has resulted in the insertion of genes putatively coding for supplementary functions for specific niches (e.g. virulence).

All of the compared strains display remnants of an ancestral dysfunctional bacteriophage resistance CRISPR system, while strains of serotype 4a harboured a second putatively functional CRISPR locus. These adaptive bacterial immune systems were thus putatively present in an early ancestor of the genus, have then been lost in later early ancestors, and finally reinserted by horizontal gene transfer into an ancestor of strains of serotype 4a.

All strains displayed intracellular up-regulation of genes implied in virulence, stress response and metabolism, which represent the core intracellular response of *Listeria monocytogenes* in order to adapt to available nutrients and challenges of this niche. Interestingly, pervasive intracellular up-regulation extended to prophage genes, an effect that was confirmed by attenuated growth of chromosomal deletion mutants of genes *lmaB* and *lmaD* of the only common prophage locus (monocin) in the murine infection model. The specific intracellular function of listerial prophage genes remains unknown.

Intracellular transcriptomic differences between strains were manifest at intracellular regulation of carbohydrate flux and flagellar genes. Strain 4b F2365, which does not down-regulate glycolysis in favour of the pentose phosphate pathway, as seen with its more pathogenic cousins, suggesting that this property contributes to poor intracellular growth of this strain. Strain 4a L99 also seems to be unable to down-regulate its intracellular transcription of flagellin, which may lead to increased detection by the host and promote rapid clearance.

In combination with observations of phenotypical attenuation of virulence in the mouse model [102], these data imply a reductive evolution of strain 4a L99 and in extension lineage III, which may signal an ongoing shift from a facultatively pathogenic lifestyle to an obligate saprophytical one to establish a less stressful and thus less contested presence in the host.

3.2.7 Dynamic integration hotspots and mobile genetic elements shape the genome structure of the species *Listeria monocytogenes*

It should be noted that the supplementary material of this publication is currently accessible at the following website until the publication process has been completed:

http://bioinfo.mikrobio.med.uni-giessen.de/dl/thesis_ck

user: thesisck321

password: thesisck321

3.2.7.1 Methods and Contribution

Genome finishing: The author conducted all bioinformatic steps involved in the complete sequencing of the chromosome of *L. monocytogenes* 7 SLCC2482 from a genomic 454 pyrosequencing run as already described.

Annotation: The author comparatively annotated 12 chromosomes of *L. monocytogenes* according to principles outlined above.

Bioinformatic analyses: The author comparatively analysed and visualized the data including global synteny analyses using Mauve, creation of phylogenetic trees based on core gene alignments and gene content using GECO, Mugsy, MEGA and GeneContent, current and future pan-genomic distribution of genes inside the species using GECO and power law regression by Excel, identification and analysis of prophages and mobile elements using GECO, identification of surface-associated genes using Augur, identification and analysis of CRISPR/Cas systems using PILER-CR, CRT, BLASTN, BLASTCLUST and BlastclustToMatrix, and installation of a public GECO server (http://bioinfo.mikrobio.med.uni-giessen.de/geco2lisdb).

Draft: The author drafted the manuscript.

3.2.7.2 Summary

In order to fully access the genetic potential for survival in the environment and the infected host, and thus evolutionary forces shaping the species, 11 new strains encompassing all serotypes of *L. monocytogenes* were completely sequenced.

The species pan-genome was extrapolated based on 16 finished chromosomes, thereby including the largest number of different genes currently available for species comparison. *L. monocytogenes* displays a large core-genome (78% of genes) indicative of a highly conserved species. The majority of accessory genes are located in nine hypervariable hotspots and 15 mobile genetic elements or prophages, highlighting few chromosomal regions as carriers for most horizontally transferred genes. The observed chromosomal stability is not correlated with the presence of CRISPR/Cas systems or other defences versus foreign DNA (RM, Abi), hinting at other mechanisms to maintain strain integrity.

Differences observed between phylogenomic and accessory gene content trees imply that observed phenotypes of listerial species may largely result from deletions, adaptation due to SNPs and a relatively small number of gene-scale insertions.

Deletion of genes described to be vital for virulence or pathogenicity was amplified in strains of lineages III and II, confirming previous observations. Phenotypical experiments found that only half of the compared strains could efficiently invade Caco-2 cells, which serve as a primary intestinal barrier. Virulence characteristics were additionally assessed by injection into *Galleria mellonella* larvae, which identified five strains as being virtually apathogenic under these circumstances, including both strains of lineage III.

The distribution of small non-coding RNAs implicated the presence of partially conserved regulatory RNA repertoires in all lineages, which may contribute to survival in the environment or the host. Further candidate genes bearing pathogenic or diagnostic potential were predicted based on the presence of functional domains and phylogenetic distribution.

Summarily, this study presented a multitude of genes that are common or disparately conserved considering lineages, serogroups, serotypes, and strains that will assist diagnosis and future research of species *L. monocytogenes*.

4 Summary

In this thesis, a general approach to complete *de novo* sequencing, annotation and bioinformatic analyses of bacterial genomes and pan-genomes was devised and implemented. The tools developed and implemented were used to process and analyse genome- and transcriptome-based data that allowed a reconstruction on the phylogeny of the genus *Listeria* and the evolution of adaptation traits enabling its survival in the intracellular environment.

An efficient finishing pipeline was constructed using a combination of commercial and open source components connected by self-written software that was successfully employed for the assembly of one bacterial chromosome and five plasmids. In order to achieve a congruent and exact annotation for 12 chromosomes and 18 plasmids, a modular comparative annotation system was created including rapid primary automatic annotation followed by manual corrections that can be easily used to predict functions for a large amount of prokaryotic replicons simultaneously.

Considerable extensions to the author's previously published GECO comparative analysis suite were implemented to assess relations of prokaryotic strains inside pan-genomes. These can be employed to (i) predict the total size of a pan-genome, as well as the distribution of conservation to identify indels supporting taxonomic or phenotypic divisions, (ii) rapidly export data necessary for the construction of phylogenomic and phenotypical trees to delineate between minor mutations and horizontal gene transfer, (iii) to identify and visualize diverging regions inside related replicons to recognize hyperdynamic hotspots, mobile elements and prophages, (iv) and to comparatively visualize limited regions for detailed assessment. A public GECO-LisDB comparative analysis server was set up bearing the largest number of completely sequenced listeriae available to date [97].

In order to allow comparative identification of small non-coding RNAs, a client/server software called sRNAdb was devised, which can rapidly detect the presence of sRNAs inside complete replicons.

These pipelines were employed for evolutionary analysis of genus *Listeria* focusing on pathogenic species *L. monocytogenes*. All compared chromosomes of the species displayed extensive similarities indicative of a highly conserved functionality. Observed

differences between phylogenetic trees indicate that minor SNP variations may have a profound impact on adaptation and thus available niches, while gene-scale indels often associated with horizontal gene transfer are rarely conserved. Most duplications found inside the genus are ancient and do not contribute to current evolutionary adaptation of listeriae.

Differentially distributed genes were predominantly found in nine highly variable regions or as a part of rare mobile genetic elements or prophages, highlighting the stability of the backbone of the species. Interestingly, only a subset of strains displayed putatively functional CRISPR adaptive immune systems of different types, or other identifiable defences against aggressive foreign DNA, indicating other factors to sustain the observed chromosomal stability, which now deserves further attention.

A further factor expanding the genomes in a subset of strains is represented by plasmids, which putatively descended from a common ancestor. Plasmid-encoded genes with identifiable functions frequently belonged to the category of stress response, which may be most beneficial in defence against disinfectants applied in food processing environments.

New candidate genes that may influence virulence of *Listeria* were predicted based on their phylogenetic distribution, functional domains and intracellular expression. Prophage-related genes had a major effect on intracellular survival of listeriae as recognized by attenuated virulence of deletion mutants.

Loss of virulence was previously identified as a recurring pattern of listerial evolution and could be seen in genomes, intracellular expression and phenotypes of several compared strains of all lineages, especially lineage III. This indicates a drift towards an obligate saprophytic or commensal lifestyle within the eukaryotic host, which may support exchange of genetic information in the nutrient-rich gastrointestinal tract and serve as a vector for transportation to a new location.

In conclusion, these studies aimed at the development of pipelines intended for the identification of evolutionary patterns within the pan-genome of *Listeria monocytogenes* and related species and uncovered new candidate genes valuable for diagnostics and virulence research.

5 Zusammenfassung

Im Verlauf dieser Dissertation wurden Werkzeuge entworfen und implementiert, die die vollständige *de novo* Sequenzierung, Annotation und bioinformatische Analyse von bakteriellen Genomen und Pan-Genomen erlauben. Diese wurden eingesetzt um Genom- und Transkriptomdaten zum Zwecke der phylogenetischen Rekonstruktion des Genus *Listeria* zu untersuchen und die evolutionäre Anpassung an die intrazelluläre Nische zu analysieren.

Hierzu wurde eine effiziente Pipeline konstruiert, die die nötigen Schritte der Sequenzierung von der Assemblierung bis zum Lückenschluss umfasst. Diese basiert auf einer Kombination kommerzieller und freier Software, die durch eigens entwickelte Programme verbunden wurden und erfolgreich zur Sequenzierung eines bakteriellen Chromosoms und fünfer Plasmide eingesetzt werden konnten. Weiterhin wurde ein modulares komparatives Annotationssystem erstellt, welches eine gleichmäßige parallele Annotation mehrerer bakterieller Genome ermöglicht. Mit dessen Hilfe wurden 12 prokaryotische Chromosomen und 18 Plasmide annotiert.

Die bereits publizierte GECO Plattform zur komparativen Genomanalyse wurde um mehrere leicht zugängliche Werkzeuge zur Pangenom-Analyse erweitert. Diese können eingesetzt werden, (i) um die erwartete Größe eines Pangenoms vorherzusagen und den Grad der Konservierung einzelner Insertionen und Deletionen zu bestimmen, welche ausschlaggebend für taxonomische oder phenotypische Gruppen sind, (ii) um rasch Daten zu exportieren, die zur Erstellung phylogenetischer oder phenotypischer Bäume genutzt werden können, um die möglichen Auswirkungen von geringfügigen Mutationen und horizontalem Gentransfer gegenüberzustellen, (iii) um divergente Regionen in mehreren Genomen zu identifizieren, was zur Erkennung hyperdynamischer Hotspots, mobiler Elemente und Prophagen genutzt werden kann, (iv) und um begrenzte Regionen im Detail komparativ zu visualisieren. Ein öffentlicher GECO Webserver wurde aufgesetzt, der diese Funktionalität für die zur Zeit größte Anzahl vollständig sequenzierter Listerien-Genome zur komparativen Analyse bereitstellt [97].

Weiterhin wurde eine Software (sRNAdb) entwickelt, welche die rasche Identifizierung und Visualisierung von kleinen nicht-kodierenden RNAs aufgrund von Sequenzähnlichkeiten erlaubt.

Diese Pipelines wurden zur Analyse der Evolution des Genus *Listeria* und besonders der pathogenen Spezies *L. monocytogenes* eingesetzt.

Alle verglichenen Chromosomen der Spezies zeigten weitreichende Ähnlichkeiten, was auf einen hohen Grad an Verwandtschaft hindeutet. Die beobachteten Unterschiede zwischen phylogenetischen Bäumen zeigen, dass geringfügige Mutationen starke Auswirkungen auf die Anpassung an verschiedene Nischen haben, während Insertionen und Deletionen vollständiger Gene häufig mit flüchtigem horizontalem Gentransfer assoziiert sind. Die große Mehrzahl der Duplikationen im Genus stammt aus der Zeit vor der Aufspaltung in einzelne Spezies, welche somit zur laufenden evolutionären Anpassung der Listerien wenig beitragen.

Differentiell verteilte Gene wurden hauptsächlich in wenigen hyperdynamischen Regionen oder als Bestandteil von Prophagen und mobilen Elementen gefunden, was die hohe Stabilität der listeriellen Chromosomen hervorhebt. Diese beruht jedoch nicht auf der Präsenz von CRISPR Immunsystemen oder anderen identifizierbaren Gegenmaßnahmen, die den Einbau von fremder DNA verhindern, sondern könnte eine Folge einer bereits erreichten annähernd optimalen Anpassung sein.

Die Genome einiger Stämme werden außerdem durch Plasmide erweitert, welche vermutlich von einem gemeinsamen Vorfahren abstammen. Auf diesen extrachromosomalen Elementen fanden sich Gene, die häufig mit der Anpassung an Streßsituationen (Desinfektionsmittel, Antibiotika) assoziiert waren und so zum Überleben in Produktionsanlagen der Lebensmittel verarbeitenden Industrie beitragen können.

Weiterhin wurden mit Hilfe der phylogenetischen Verteilung, der Präsenz funktionaler Domänen, sowie der intrazellulären Expression neue Gene identifiziert, welche möglicherweise zur Virulenz oder Pathogenität von *Listerien* beitragen. Eingeschränkte Virulenz von Deletions-Mutanten hat gezeigt, dass einige Prophagen-assoziierte Gene auf noch unbekannte Art zur intrazellulären Anpassung beisteuern.

Die Verringerung der Virulenz wurde bereits als verbreitetes evolutionäres Muster von Spezies *L. monocytogenes* beschrieben, was die vorliegende Arbeit durch Genom-, Transkriptom- und phenotypische Daten bestätigen konnte, insbesondere hinsichtlich der Abstammungslinie III. Diese weist auf eine Verschiebung in Richtung eines obligat saprophytischen oder kommensalen Lebensstiles innerhalb eines eukaryotischen Wirtes hin, was den Austausch von genetischer Information im nährstoff- und bakterienreichen

Gastrointestinaltrakt begünstigen könnte. Der Wirt dient vermutlich außerdem als Transportmittel um eine Verteilung der Listerien zu gewährleisten.

Zusammenfassend läßt sich sagen, daß in den zugrundeliegenden Publikationen bioinformatische Pipelines erstellt und angewandt wurden, welche zur Aufklärung evolutionärer Muster des Pangenoms von *Listeria monocytogenes* und verwandter Spezies genutzt werden konnten. Weiterhin wurden neue Gene identifiziert, welche zu zukünftigen diagnostischen Methoden beitragen können oder möglicherweise mit der Virulenz einzelner Stämme im Zusammenhang stehen.

6 Abbreviations

Abi	abortive infection
asRNA	antisense small non-coding RNA
bp	basepairs
cas	CRISPR-associated
CDS	coding sequence
COG	clusters of orthologous groups
CRISPR	clustered regularly interspaced short palindromic repeats
ddNTP	dideoxyribonucleoside triphosphates
DNA	deoxyribonucleic acid
FIC	filamentation induced by cyclic adenosine monophosphate
HGT	horizontal gene transfer
HMM	hidden markov model
indel	insertion or deletion
LIPI	*Listeria* pathogenicity island
LRR	leucine-rich repeat
Mb	megabases
MLST	multi locus sequence typing
NCBI	National Center for Biotechnology Information
ncRNA	non-coding RNA
ORF	open reading frame (protein coding DNA sequence)
PCR	polymerase chain reaction
RNA	ribonucleic acid
SNP	single nucleotide polymorphism
spp.	subspecies

sRNA	small non-coding regulatory RNA
VGC	virulence gene cluster

7 References

1. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* 1977; 74: 5463-7.
2. Sanger F, Air GM, Barrell BG *et al*. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* 1977; 265: 687-95.
3. Fleischmann RD, Adams MD, White O *et al*. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 1995; 269: 496-512.
4. 10,000 Microbe Genome Plan Starts in Shenzhen. Chinese Academy of Sciences 2012. http://english.cas.cn/Ne/CASE/200908/t20090805_44705.shtml
5. Morozova O. Sequencing: how is DNA read? Genome British Columbia 2009. http://www.genomebc.ca/education/articles/sequencing
6. Rothberg JM, Leamon JH. The development and impact of 454 sequencing. *Nat Biotechnol* 2008; 26: 1117-24.
7. Madabhushi RS. Separation of 4-color DNA sequencing extension products in noncovalently coated capillaries using low viscosity polymer solutions. *Electrophoresis* 1998; 19: 224-30.
8. Margulies M, Egholm M, Altman WE *et al*. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 2005; 437: 376-80.
9. MacLean D, Jones JD, Studholme DJ. Application of 'next-generation' sequencing technologies to microbial genetics. *Nat Rev Microbiol* 2009; 7: 287-96.
10. Branton D, Deamer DW, Marziali A *et al*. The potential and challenges of nanopore sequencing. *Nat Biotechnol* 2008; 26: 1146-53.
11. Metzker ML. Sequencing technologies - the next generation. *Nat Rev Genet* 2010; 11: 31-46.
12. Zhang J, Chiodini R, Badr A, Zhang G. The impact of next-generation sequencing on genomics. *J Genet Genomics* 2011; 38: 95-109.
13. National Center for Biotechnology Information. NCBI 2012. http://www.ncbi.nlm.nih.gov
14. European Bioinformatics Institute. EBI 2012. http://www.ebi.ac.uk/
15. Miller JR, Koren S, Sutton G. Assembly algorithms for next-generation sequencing data. *Genomics* 2010; 95: 315-27.

16. Kumar S, Blaxter ML. Comparing de novo assemblers for 454 transcriptome data. *BMC Genomics* 2010; 11: 571.

17. Darling AE, Mau B, Perna NT. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* 2010; 5: e11147.

18. Kim PG, Cho HG, Park K. A scaffold analysis tool using mate-pair information in genome sequencing. *J Biomed Biotechnol* 2008; 2008: 675741.

19. Latreille P, Norton S, Goldman BS *et al*. Optical mapping as a routine tool for bacterial genome sequence finishing. *BMC Genomics* 2007; 8: 321.

20. Richter DC, Schuster SC, Huson DH. OSLay: optimal syntenic layout of unfinished assemblies. *Bioinformatics* 2007; 23: 1573-9.

21. van Hijum SA, Zomer AL, Kuipers OP, Kok J. Projector 2: contig mapping for efficient gap-closure of prokaryotic genome sequence assemblies. *Nucleic Acids Res* 2005; 33: W560-W566.

22. Koressaar T, Remm M. Enhancements and modifications of primer design program Primer3. *Bioinformatics* 2007; 23: 1289-91.

23. Gordon D. Viewing and editing assembled sequences using Consed. *Curr Protoc Bioinformatics* 2003; Chapter 11: Unit11.

24. Altschul SF, Madden TL, Schaffer AA *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997; 25: 3389-402.

25. Larkin MA, Blackshields G, Brown NP *et al*. Clustal W and Clustal X version 2.0. *Bioinformatics* 2007; 23: 2947-8.

26. Kurtz S, Phillippy A, Delcher AL *et al*. Versatile and open software for comparing large genomes. *Genome Biol* 2004; 5: R12.

27. Angiuoli SV, Salzberg SL. Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics* 2011; 27: 334-42.

28. Bray N, Pachter L. MAVID: constrained ancestral alignment of multiple sequences. *Genome Res* 2004; 14: 693-9.

29. Sarkar A, Soueidan H, Nikolski M. Identification of conserved gene clusters in multiple genomes based on synteny and homology. *BMC Bioinformatics* 2011; 12 Suppl 9: S18.

30. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009; 10: 57-63.

31. Storz G, Altuvia S, Wassarman KM. An abundance of RNA regulators. *Annu Rev Biochem* 2005; 74: 199-217.
32. Mraheil MA, Billion A, Mohamed W *et al*. The intracellular sRNA transcriptome of *Listeria monocytogenes* during growth in macrophages. *Nucleic Acids Res* 2011; 39: 4235-48.
33. Cao Y, Wu J, Liu Q *et al*. sRNATarBase: a comprehensive database of bacterial sRNA targets verified by experiments. *RNA* 2010; 16: 2051-7.
34. Gardner PP, Daub J, Tate JG *et al*. Rfam: updates to the RNA families database. *Nucleic Acids Res* 2009; 37: D136-D140.
35. Medini D, Serruto D, Parkhill J *et al*. Microbiology in the post-genomic era. *Nat Rev Microbiol* 2008; 6: 419-30.
36. Ochman H, Lawrence JG, Groisman EA. Lateral gene transfer and the nature of bacterial innovation. *Nature* 2000; 405: 299-304.
37. Waack S, Keller O, Asper R *et al*. Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. *BMC Bioinformatics* 2006; 7: 142.
38. Touchon M, Hoede C, Tenaillon O *et al*. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet* 2009; 5: e1000344.
39. Tettelin H, Masignani V, Cieslewicz MJ *et al*. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc Natl Acad Sci U S A* 2005; 102: 13950-5.
40. Lapidus A, Goltsman E, Auger S *et al*. Extending the Bacillus cereus group genomics to putative food-borne pathogens of different toxicity. *Chem Biol Interact* 2008; 171: 236-49.
41. Lefebure T, Stanhope MJ. Evolution of the core and pan-genome of Streptococcus: positive selection, recombination, and genome composition. *Genome Biol* 2007; 8: R71.
42. Hogg JS, Hu FZ, Janto B *et al*. Characterization and modeling of the *Haemophilus influenzae* core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains. *Genome Biol* 2007; 8: R103.
43. Tringe SG, von MC, Kobayashi A *et al*. Comparative metagenomics of microbial communities. *Science* 2005; 308: 554-7.

44. Laing C, Buchanan C, Taboada EN *et al*. Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions. *BMC Bioinformatics* 2010; 11: 461.

45. Ostlund G, Schmitt T, Forslund K *et al*. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res* 2010; 38: D196-D203.

46. Kuenne CT, Ghai R, Chakraborty T, Hain T. GECO--linear visualization for comparative genomics. *Bioinformatics* 2007; 23: 125-6.

47. Delcher AL, Bratke KA, Powers EC, Salzberg SL. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* 2007; 23: 673-9.

48. Schattner P, Brooks AN, Lowe TM. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res* 2005; 33: W686-W689.

49. Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res* 2007; 35: 3100-8.

50. Siezen RJ, van Hijum SA. Genome (re-)annotation and open-source annotation pipelines. *Microb Biotechnol* 2010; 3: 362-9.

51. Meyer F, Goesmann A, McHardy AC *et al*. GenDB--an open source genome annotation system for prokaryote genomes. *Nucleic Acids Res* 2003; 31: 2187-95.

52. Aziz RK, Bartels D, Best AA *et al*. The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 2008; 9: 75.

53. Sayers EW, Barrett T, Benson DA *et al*. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2012; 40: D13-D25.

54. Punta M, Coggill PC, Eberhardt RY *et al*. The Pfam protein families database. *Nucleic Acids Res* 2012; 40: D290-D301.

55. Tatusov RL, Fedorova ND, Jackson JD *et al*. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 2003; 4: 41.

56. Szklarczyk D, Franceschini A, Kuhn M *et al*. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res* 2011; 39: D561-D568.

57. Blom J, Albaum SP, Doppmeier D *et al*. EDGAR: a software framework for the comparative analysis of prokaryotic genomes. *BMC Bioinformatics* 2009; 10: 154.

58. Poptsova MS, Gogarten JP. Using comparative genome analysis to identify problems in annotated microbial genomes. *Microbiology* 2010; 156: 1909-17.
59. Gu X, Huang W, Xu D, Zhang H. GeneContent: software for whole-genome phylogenetic analysis. *Bioinformatics* 2005; 21: 1713-4.
60. Nei M. Phylogenetic analysis in molecular evolutionary genetics. *Annu Rev Genet* 1996; 30: 371-403.
61. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* 2011; 28: 2731-9.
62. Huson DH. SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics* 1998; 14: 68-73.
63. Huson DH, Richter DC, Rausch C, Dezulian T, Franz M, Rupp R. Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics* 2007; 8: 460.
64. Ghai R, Chakraborty T. Comparative microbial genome visualization using GenomeViz. *Methods Mol Biol* 2007; 395: 97-108.
65. Duffy LL, Vanderlinde PB, Grau FH. Growth of *Listeria monocytogenes* on vacuum-packed cooked meats: effects of pH, aw, nitrite and ascorbate. *Int J Food Microbiol* 1994; 23: 377-90.
66. den Bakker HC, Cummings CA, Ferreira V *et al*. Comparative genomics of the bacterial genus *Listeria*: Genome evolution is characterized by limited gene acquisition and limited gene loss. *BMC Genomics* 2010; 11: 688.
67. Chakraborty T, Hain T, Domann E. Genome organization and the evolution of the virulence gene locus in *Listeria* species. *Int J Med Microbiol* 2000; 290: 167-74.
68. Hain T, Steinweg C, Kuenne CT *et al*. Whole-genome sequence of *Listeria welshimeri* reveals common steps in genome reduction with Listeria innocua as compared to Listeria monocytogenes. *J Bacteriol* 2006; 188: 7405-15.
69. Perez-Diaz JC, Vicente MF, Baquero F. Plasmids in *Listeria*. *Plasmid* 1982; 8: 112-8.
70. Peterkin PI, Gardiner MA, Malik N, Idziak ES. Plasmids in *Listeria monocytogenes* and other Listeria species. *Can J Microbiol* 1992; 38: 161-4.
71. Glaser P, Frangeul L, Buchrieser C *et al*. Comparative genomics of *Listeria* species. *Science* 2001; 294: 849-52.

72. Lebrun M, Loulergue J, Chaslus-Dancla E, Audurier A. Plasmids in *Listeria monocytogenes* in relation to cadmium resistance. *Appl Environ Microbiol* 1992; 58: 3183-6.

73. McLauchlin J, Hampton MD, Shah S, Threlfall EJ, Wieneke AA, Curtis GD. Subtyping of *Listeria monocytogenes* on the basis of plasmid profiles and arsenic and cadmium susceptibility. *J Appl Microbiol* 1997; 83: 381-8.

74. Taillefer C, Boucher M, Laferriere C, Morin L. Perinatal listeriosis: Canada's 2008 outbreaks. *J Obstet Gynaecol Can* 2010; 32: 45-8.

75. Fretz R, Sagel U, Ruppitsch W *et al*. Listeriosis outbreak caused by acid curd cheese Quargel , Austria and Germany 2009. *Euro Surveill* 2010; 15.

76. Smith B, Larsson JT, Lisby M *et al*. Outbreak of listeriosis caused by infected beef meat from a meals-on-wheels delivery in Denmark 2009. *Clin Microbiol Infect* 2011; 17: 50-2.

77. Farber JM, Peterkin PI. *Listeria monocytogenes*, a food-borne pathogen. *Microbiol Rev* 1991; 55: 476-511.

78. Centers for Disease Control and Prevention. CDC 2012. http://www.cdc.gov/listeria/

79. Robert Koch Institut. RKI 2012. http://www.rki.de/

80. Ragon M, Wirth T, Hollandt F *et al*. A new perspective on *Listeria monocytogenes* evolution. *PLoS Pathog* 2008; 4: e1000146.

81. Orsi RH, den Bakker HC, Wiedmann M. *Listeria monocytogenes* lineages: Genomics, evolution, ecology, and phenotypic characteristics. *Int J Med Microbiol* 2011; 301: 79-96.

82. Gaillard JL, Berche P, Frehel C, Gouin E, Cossart P. Entry of *L. monocytogenes* into cells is mediated by internalin, a repeat protein reminiscent of surface antigens from gram-positive cocci. *Cell* 1991; 65: 1127-41.

83. Poyart C, Trieu-Cuot P, Berche P. The inlA gene required for cell invasion is conserved and specific to *Listeria monocytogenes*. *Microbiology* 1996; 142 (Pt 1): 173-80.

84. Roche SM, Gracieux P, Milohanic E *et al*. Investigation of specific substitutions in virulence genes characterizing phenotypic groups of low-virulence field strains of *Listeria monocytogenes*. *Appl Environ Microbiol* 2005; 71: 6039-48.

85. Vazquez-Boland JA, Kuhn M, Berche P *et al*. *Listeria* pathogenesis and molecular virulence determinants. *Clin Microbiol Rev* 2001; 14: 584-640.

86. Bierne H, Cossart P. *Listeria monocytogenes* surface proteins: from genome predictions to function. *Microbiol Mol Biol Rev* 2007; 71: 377-97.

87. Borucki MK, Call DR. *Listeria monocytogenes* serotype identification by PCR. *J Clin Microbiol* 2003; 41: 5537-40.

88. Phillippy AM, Deng X, Zhang W, Salzberg SL. Efficient oligonucleotide probe selection for pan-genomic tiling arrays. *BMC Bioinformatics* 2009; 10: 293.

89. Deng X, Phillippy AM, Li Z, Salzberg SL, Zhang W. Probing the pan-genome of *Listeria monocytogenes*: new insights into intraspecific niche expansion and genomic diversification. *BMC Genomics* 2010; 11: 500.

90. Wilson GG. Organization of restriction-modification systems. *Nucleic Acids Res* 1991; 19: 2539-66.

91. Fineran PC, Blower TR, Foulds IJ, Humphreys DP, Lilley KS, Salmond GP. The phage abortive infection system, ToxIN, functions as a protein-RNA toxin-antitoxin pair. *Proc Natl Acad Sci U S A* 2009; 106: 894-9.

92. Makarova KS, Haft DH, Barrangou R *et al*. Evolution and classification of the CRISPR-Cas systems. *Nat Rev Microbiol* 2011; 9: 467-77.

93. Edgar RC. PILER-CR: fast and accurate identification of CRISPR repeats. *BMC Bioinformatics* 2007; 8: 18.

94. Bland C, Ramsey TL, Sabree F *et al*. CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* 2007; 8: 209.

95. Wang Y, Qian PY. Conservative fragments in bacterial 16S rRNA genes and primer design for 16S ribosomal DNA amplicons in metagenomic studies. *PLoS One* 2009; 4: e7401.

96. Kuenne C. GECO comparative genome analysis server (prokaryote). JLU Giessen 2012. http://bioinfo.mikrobio.med.uni-giessen.de/geco2

97. Kuenne C. GECO comparative analysis server (listeriae). JLU Giessen 2012. http://bioinfo.mikrobio.med.uni-giessen.de/geco2lisdb

98. Tettelin H, Riley D, Cattuto C, Medini D. Comparative genomics: the bacterial pan-genome. *Curr Opin Microbiol* 2008; 11: 472-7.

99. Treangen TJ, Rocha EP. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet* 2011; 7: e1001284.

100. Promadej N, Fiedler F, Cossart P, Dramsi S, Kathariou S. Cell wall teichoic acid glycosylation in *Listeria monocytogenes* serotype 4b requires gtcA, a novel, serogroup-specific gene. *J Bacteriol* 1999; 181: 418-25.

101. Nelson KE, Fouts DE, Mongodin EF *et al*. Whole genome comparisons of serotype 4b and 1/2a strains of the food-borne pathogen *Listeria monocytogenes* reveal new insights into the core genome components of this species. *Nucleic Acids Res* 2004; 32: 2386-95.

102. Chakraborty T, Ebel F, Wehland J, Dufrenne J, Notermans S. Naturally occurring virulence-attenuated isolates of *Listeria monocytogenes* capable of inducing long term protection against infection by virulent strains of homologous and heterologous serotypes. *FEMS Immunol Med Microbiol* 1994; 10: 1-9.

8 Erklärung zur Dissertation

„Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig und ohne unzulässige Hilfe oder Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Alle Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder nichtveröffentlichten Schriften entnommen sind, und alle Angaben, die auf mündlichen Auskünften beruhen, sind als solche kenntlich gemacht. Bei den von mir durchgeführten und in der Dissertation erwähnten Untersuchungen habe ich die Grundsätze guter wissenschaftlicher Praxis, wie sie in der „Satzung der Justus-Liebig-Universität Gießen zur Sicherung guter wissenschaftlicher Praxis“ niedergelegt sind, eingehalten sowie ethische, datenschutzrechtliche und tierschutzrechtliche Grundsätze befolgt. Ich versichere, dass Dritte von mir weder unmittelbar noch mittelbar geldwerte Leistungen für Arbeiten erhalten haben, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen, oder habe diese nachstehend spezifiziert. Die vorgelegte Arbeit wurde weder im Inland noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde zum Zweck einer Promotion oder eines anderen Prüfungsverfahrens vorgelegt. Alles aus anderen Quellen und von anderen Personen übernommene Material, das in der Arbeit verwendet wurde oder auf das direkt Bezug genommen wird, wurde als solches kenntlich gemacht. Insbesondere wurden alle Personen genannt, die direkt und indirekt an der Entstehung der vorliegenden Arbeit beteiligt waren. Mit der Überprüfung meiner Arbeit durch eine Plagiatserkennungssoftware bzw. ein internetbasiertes Softwareprogramm erkläre ich mich einverstanden.“

Ort, Datum

Unterschrift

9 Acknowledgements

First I want to express my gratitude to my supervisor, Prof. Dr. Trinad Chakraborty for innumerable ideas, fruitful discussions and constant encouragement.

Thanks are also extended to Dr. Torsten Hain who offered unwavering support in stressful times.

Furthermore, I would like to acknowledge Mobarak Mraheil, Nelli Schklarenko, and Claudia Zörb, whose excellent technical abilities permitted me to completely focus on the bioinformatic analysis.

Last but not least I want to thank my parents and my brother for their enduring support that helped me to stay on course.

10 Supplementary material

The following pages include the original publications or manuscripts in preparation.

Journal of Bacteriology, Mar. 2010, p. 1473–1474 Vol. 192, No. 5
0021-9193/10/$12.00 doi:10.1128/JB.01415-09

Complete Genome Sequence of *Listeria seeligeri*, a Nonpathogenic Member of the Genus *Listeria*[▽]

Christiane Steinweg,[1]§ Carsten T. Kuenne,[1]§ André Billion,[1] Mobarak A. Mraheil,[1] Eugen Domann,[1] Rohit Ghai,[1] Sukhadeo B. Barbuddhe,[1] Uwe Kärst,[2] Alexander Goesmann,[3] Alfred Pühler,[4] Bernd Weisshaar,[5] Jürgen Wehland,[2] Robert Lampidis,[6] Jürgen Kreft,[6] Werner Goebel,[6] Trinad Chakraborty,[1]* and Torsten Hain[1]*

Institute of Medical Microbiology, Justus-Liebig-University, Frankfurter Strasse 107, D-35392 Giessen, Germany[1]; Helmholtz-Centre for Infection Research, Inhoffenstrasse 7, D-38124 Braunschweig, Germany[2]; Bioinformatics Resource Facility, Centrum für Biotechnologie, University of Bielefeld, Universitätsstrasse 25, D-33615 Bielefeld, Germany[3]; Lehrstuhl für Genetik, University of Bielefeld, Universitätsstrasse 25, D-33615 Bielefeld, Germany[4]; Lehrstuhl für Genomforschung, University of Bielefeld, Universitätsstrasse 25, D-33615 Bielefeld, Germany[5]; and Lehrstuhl für Mikrobiologie, University of Würzburg, Am Hubland/Biozentrum, D-97074 Würzburg, Germany[6]

Received 29 October 2009/Accepted 23 December 2009

We report the complete and annotated genome sequence of the nonpathogenic *Listeria seeligeri* SLCC3954 serovar 1/2b type strain harboring the smallest completely sequenced genome of the genus *Listeria*.

Listeria seeligeri is one of seven species of the genus *Listeria*, a group of Gram-positive, motile, facultative anaerobic, low-GC-content, nonsporulating rods (3, 10, 12). To obtain a better understanding of the evolution of this nonpathogenic *Listeria* species, the type strain, SLCC 3954 (serovar 1/2b), a soil isolate from Germany (11), was sequenced using Sanger technology. Two small (1.5- to 2.5-kb) insert plasmid libraries were constructed with the TOPO Shotgun Subcloning Kit (Invitrogen) as described previously (9). Additionally, a small insert plasmid library (~1.0 to 1.5 kb) and a medium insert plasmid library (~5 kb) were constructed in the pUC19 cloning vector (New England Biolabs) by Qiagen (Hilden, Germany). A fosmid library harboring fragments of around 40 kb was created using the CopyControl Fosmid Library Production Kit (Epicentre) as published before (9).

Sequencing was performed by Agowa (Berlin, Germany), Qiagen (Hilden, Germany), and the Max-Planck-Institut (Köln, Germany) by using ABI Big Dye Terminator technology. A total of 30,961 and 1,305 high-quality reads from the shotgun, and fosmid libraries were used to generate a draft assembly with an overall coverage of ~7-fold by using the Phred/Phrap/Consed assembly package (4, 5, 7). Contigs were linked by primer walking on shotgun clones and fosmids as well as by PCR gap closure followed by sequencing of the PCR product. Genome annotation was performed as described previously (9).

The genome of *L. seeligeri* consists of a circular chromosome of 2,797,636 bp and hence is slightly smaller than those of previously sequenced listerial strains (6, 9). Functional classification of genes obtained by mapping against clusters of orthologous groups (14) was predicted by Augur (1). Comparative analysis of these clusters indicates gene loss in categories such as amino acid/carbohydrate transport and metabolism as well as in transcription, thus confirming a general trend toward genome shrinkage. *L. seeligeri* harbors no plasmid and carries only a single copy of a prophage and no transposon in its genome. The mean G+C content of the *L. seeligeri* genome is 37.4%, which is close to the average value of all known *Listeria* strains (6, 9). G/C skew analysis revealed a bidirectional replication mechanism, and the origin of replication (oriC) is located close to the *dnaA* gene, which is positioned diametrically opposite to the replication terminus. We identified six 16S-23S-5S rRNA operons, all of which are located on the leading strand, two on the right and four on the left replichore. Additionally, a total of 67 tRNA genes were detected. We used MAVID (2) for phylogenetic analysis of the genomes of *L. monocytogenes*, *L. innocua*, *L. welshimeri*, and *L. seeligeri* and conclude that the "phylogenomic" relationship corresponds exactly to phylogenetic analysis based on either 16S rRNA genes (8, 15) or other additional specific marker genes (8, 13).

The genome sequence of *L. seeligeri* is the fourth species of genus *Listeria* to be reported.

Nucleotide sequence accession number. The genome sequence of *L. seeligeri* serovar 1/2b (SLCC3954) reported here has been deposited in the EMBL database under accession number FN557490.

We thank Alexandra Amend, Claudia Zörb, Nelli and Juri Schklarenko, and Prisca Viehoever for excellent technical assistance.

This work was supported by the fund obtained from the BMBF through the Competence Network PathoGenoMik (031U213B) and the ERA-NET PathoGenoMics program grant SPATELIS to T.C. and T.H.

REFERENCES

1. **Billion, A., R. Ghai, T. Chakraborty, and T. Hain.** 2006. Augur—a computational pipeline for whole genome microbial surface protein prediction and classification. Bioinformatics **22:**2819–2820.

* Corresponding author. Mailing address: Institute of Medical Microbiology, Justus-Liebig-University, Frankfurter Strasse 107, D-35392 Giessen, Germany. Phone: 49-641 99 46400. Fax: 49-641 99 46409. E-mail for T. Hain: Torsten.Hain@mikrobio.med.uni-giessen.de. E-mail for T. Chakraborty: Trinad.Chakraborty@mikrobio.med.uni-giessen.de.

§ Both coauthors contributed equally to the work.

[▽] Published ahead of print on 8 January 2010.

2. **Bray, N., and L. Pachter.** 2004. MAVID: constrained ancestral alignment of multiple sequences. Genome Res. **14:**693–699.
3. **Collins, M. D., S. Wallbanks, D. J. Lane, J. Shah, R. Nietupski, J. Smida, M. Dorsch, and E. Stackebrandt.** 1991. Phylogenetic analysis of the genus *Listeria* based on reverse transcriptase sequencing of 16S rRNA. Int. J. Syst. Bacteriol. **41:**240–246.
4. **Ewing, B., and P. Green.** 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. **8:**186–194.
5. **Ewing, B., L. Hillier, M. C. Wendl, and P. Green.** 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. **8:**175–185.
6. **Glaser, P., L. Frangeul, C. Buchrieser, C. Rusniok, A. Amend, F. Baquero, P. Berche, H. Bloecker, P. Brandt, T. Chakraborty, A. Charbit, F. Chetouani, E. Couve, A. de Daruvar, P. Dehoux, E. Domann, G. Dominguez-Bernal, E. Duchaud, L. Durant, O. Dussurget, K. D. Entian, H. Fsihi, F. Garcia-del Portillo, P. Garrido, L. Gautier, W. Goebel, N. Gomez-Lopez, T. Hain, J. Hauf, D. Jackson, L. M. Jones, U. Kaerst, J. Kreft, M. Kuhn, F. Kunst, G. Kurapkat, E. Madueno, A. Maitournam, J. M. Vicente, E. Ng, H. Nedjari, G. Nordsiek, S. Novella, B. de Pablos, J. C. Perez-Diaz, R. Purcell, B. Remmel, M. Rose, T. Schlueter, N. Simoes, A. Tierrez, J. A. Vazquez-Boland, H. Voss, J. Wehland, and P. Cossart.** 2001. Comparative genomics of *Listeria* species. Science **294:**849–852.
7. **Gordon, D., C. Abajian, and P. Green.** 1998. Consed: a graphical tool for sequence finishing. Genome Res. **8:**195–202.
8. **Graves, L. M., L. O. Helsel, A. G. Steigerwalt, R. E. Morey, M. I. Daneshvar, S. E. Roof, R. H. Orsi, E. D. Fortes, S. R. Milillo, H. C. den Bakker, M. Wiedmann, B. Swaminathan, and B. D. Sauders.** 2009. *Listeria marthii* sp. nov., isolated from the natural environment, Finger Lakes National Forest. Int. J. Syst. Evol. Microbiol. doi 10.1099/ijs.0.014118–0.
9. **Hain, T., C. Steinweg, C. T. Kuenne, A. Billion, R. Ghai, S. S. Chatterjee, E. Domann, U. Karst, A. Goesmann, T. Bekel, D. Bartels, O. Kaiser, F. Meyer, A. Puhler, B. Weisshaar, J. Wehland, C. Liang, T. Dandekar, R. Lampidis, J. Kreft, W. Goebel, and T. Chakraborty.** 2006. Whole-genome sequence of *Listeria welshimeri* reveals common steps in genome reduction with *Listeria innocua* as compared to *Listeria monocytogenes*. J. Bacteriol. **188:**7405–7415.
10. **Rocourt, J.** 1999. The genus *Listeria* and *Listeria monocytogenes*: phylogenetic positions, taxonomy, and identification, p. 1–20. *In* E. T. Ryser and E. Marth (ed.), *Listeria*, listeriosis, and food safety. Macel Dekker Inc., New York, NY.
11. **Rocourt, J., and P. Grimont.** 1983. *Listeria welshimeri* sp. nov. and *Listeria seeligeri* sp. nov. Int. J. Syst. Bacteriol. **33:**866–869.
12. **Sallen, B., A. Rajoharison, S. Desvarenne, F. Quinn, and C. Mabilat.** 1996. Comparative analysis of 16S and 23S rRNA sequences of *Listeria* species. Int. J. Syst. Bacteriol. **46:**669–674.
13. **Schmid, M. W., E. Y. Ng, R. Lampidis, M. Emmerth, M. Walcher, J. Kreft, W. Goebel, M. Wagner, and K. H. Schleifer.** 2005. Evolutionary history of the genus *Listeria* and its virulence genes. Syst. Appl. Microbiol. **28:**1–18.
14. **Tatusov, R. L., D. A. Natale, I. V. Garkavtsev, T. A. Tatusova, U. T. Shankavaram, B. S. Rao, B. Kiryutin, M. Y. Galperin, N. D. Fedorova, and E. V. Koonin.** 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res. **29:**22–28.
15. **Vaneechoutte, M., P. Boerlin, H. V. Tichy, E. Bannerman, B. Jager, and J. Bille.** 1998. Comparison of PCR-based DNA fingerprinting techniques for the identification of *Listeria* species and their use for atypical *Listeria* isolates. Int. J. Syst. Bacteriol. **48:**127–139.

Comparative Analysis of Plasmids in the Genus *Listeria*

Carsten Kuenne[1], Sonja Voget[2], Jordan Pischimarov[1], Sebastian Oehm[3], Alexander Goesmann[3], Rolf Daniel[2], Torsten Hain[1]*, Trinad Chakraborty[1]*

1 Institute of Medical Microbiology, Justus-Liebig University, Giessen, Germany, **2** Goettingen Genomics Laboratory, Institute for Microbiology and Genetics, Georg-August University Goettingen, Goettingen, Germany, **3** Bioinformatics Resource Facility, Center for Biotechnology, Bielefeld University, Bielefeld, Germany

Abstract

Background: We sequenced four plasmids of the genus *Listeria*, including two novel plasmids from *L. monocytogenes* serotype 1/2c and 7 strains as well as one from the species *L. grayi*. A comparative analysis in conjunction with 10 published *Listeria* plasmids revealed a common evolutionary background.

Principal Findings: All analysed plasmids share a common replicon-type related to theta-replicating plasmid pAMbeta1. Nonetheless plasmids could be broadly divided into two distinct groups based on replicon diversity and the genetic content of the respective plasmid groups. *Listeria* plasmids are characterized by the presence of a large number of diverse mobile genetic elements and a commonly occurring translesion DNA polymerase both of which have probably contributed to the evolution of these plasmids. We detected small non-coding RNAs on some plasmids that were homologous to those present on the chromosome of *L. monocytogenes* EGD-e. Multiple genes involved in heavy metal resistance (cadmium, copper, arsenite) as well as multidrug efflux (MDR, SMR, MATE) were detected on all listerial plasmids. These factors promote bacterial growth and survival in the environment and may have been acquired as a result of selective pressure due to the use of disinfectants in food processing environments. MDR efflux pumps have also recently been shown to promote transport of cyclic diadenosine monophosphate (c-di-AMP) as a secreted molecule able to trigger a cytosolic host immune response following infection.

Conclusions: The comparative analysis of 14 plasmids of genus *Listeria* implied the existence of a common ancestor. Ubiquitously-occurring MDR genes on plasmids and their role in listerial infection now deserve further attention.

Citation: Kuenne C, Voget S, Pischimarov J, Oehm S, Goesmann A, et al. (2010) Comparative Analysis of Plasmids in the Genus *Listeria*. PLoS ONE 5(9): e12511. doi:10.1371/journal.pone.0012511

Editor: Niyaz Ahmed, University of Hyderabad, India

Received July 20, 2010; **Accepted** August 10, 2010; **Published** September 2, 2010

Funding: The project was funded by the German Ministry of Education and Research through the ERANET program grants sncRNAomics and SPATELIS to T.H. and T.C. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: Trinad.Chakraborty@mikrobio.med.uni-giessen.de (TC); Torsten.Hain@mikrobio.med.uni-giessen.de (TH)

Introduction

The genus *Listeria* comprises six non-pathogenic species *L. marthii*, *L. innocua*, *L. welshimeri*, *L. seeligeri*, *L. grayi*, and *L. rocourtiae*, and two species with pathogenic potential viz. *L. monocytogenes* and *L. ivanovii*, which can cause both human and animal infections [1–7]. Since *L. monocytogenes* exhibits resistance towards heat and cold stress it can proliferate in food processing environments [8] and thus colonize dairy and meat products which have caused several outbreaks as well as sporadic cases of listeriosis [9]. Three serotypes of the species *L. monocytogenes* viz. 1/2a, 1/2b and 4b are responsible for 95% of all human clinical infections [10].

Extrachromosomal DNA was previously detected in many *L. monocytogenes* wildtype strains with rates of isolation ranging from 0–79% with an overall average of 30% [11–15]. Two studies which examined 173 [14] and 322 [16] isolates of *L. monocytogenes* respectively found an overrepresentation of plasmids in strains from food and the environment in comparison to those obtained from clinical cases. It was shown that plasmids were found more frequently (75%) in recurrent *L. monocytogenes* strains sampled from food/processing environments than in those from sporadic strains (35%) [17]. Plasmids were also more frequently associated with serogroup 1 strains compared to those from serogroup 4. It was determined that 95% of the *L. monocytogenes* plasmid-positive strains were resistant towards cadmium versus only 13% of the plasmid-negative strains [14] and that the *cadAC* genes were similar to those previously detected in *Staphylococcus aureus* [18]. Only in two cases antibiotic resistance of *L. monocytogenes* could be traced to a plasmid [19,20]. Plasmids were also previously described for *L. innocua* [2] and *L. grayi* [12]. Furthermore plasmids pAMbeta1 and pIP501 of *Streptococcus* could be transferred to *L. monocytogenes* where they stably replicated underlining the broad-host range of these replicons and their potential for horizontal transfer between strains of these genera [12,21]. The contribution of plasmids to the infectious process has not been examined and their evolutionary history is not yet well understood apart from homologies to other gram-positive plasmids such as with plasmid pXO2 from *Bacillus anthracis* which is required for the pathogenic properties of this species [22–24].

Results and Discussion

Listeria plasmids overview

We determined the entire sequences of plasmids from *L. monocytogenes* 7 UG1 SLCC2482, *L. monocytogenes* 1/2c UG1

SLCC2372, *L. monocytogenes* 1/2b UG1 SLCC2755 and *L. grayi* subspecies *grayi* UG1 DSM20601. For comparative analysis we included sequences of the plasmids pLM33 of *L. monocytogenes* Lm1, pCT100 of *L. monocytogenes* DRDC8, pLM80 of *L. monocytogenes* H7858, pLM5578 of *L. monocytogenes* 08-5578 and pLI100 of *L. innocua* Clip11262 which were downloaded from the NCBI website as well as further five gapped *L. monocytogenes* plasmid sequences from strains FSL J1.194, FSL R2-503, FSL N1-017, FSL F2-515 and J0161 which were retrieved from the Broad Institute (http://www.broad.mit.edu) database. All plasmid contigs were remapped and reannotated.

It should be noted that plasmids sequenced by the Broad Institute were found to contain a large number of SNPs leading to truncated genes. A recent study assumed that higher selective pressure was responsible for this phenomenon [23], but other studies with this data have also indicated truncations in many essential housekeeping genes on the chromosomes of these strains [25] indicating an alternative explanation i.e. sequencing errors. Indeed the average sizes of coding sequences from *L. monocytogenes* plasmids sequenced in this study vary between 260 and 264 while those obtained from the Broad study range from 131 to 245 amino acids, respectively (Table 1).

Since it is not feasible to include locus tags for up to 14 homologs of a gene we decided to only include a gene name or annotation in the text which can be used in conjunction with a homology matrix (Table S1) to identify the respective loci. Furthermore a public Geco server [26] including all plasmids of this study as well as their reference annotations was set up (http://bioinfo.mikrobio.med.uni-giessen.de/geco2plasmids/).

Phylogenetic clustering based on replication protein

All plasmids contained a similar minimal replicon consisting of three genes necessary for replication (*repA*) and partitioning (*repB, repC*) as well as the origin of replication [27–32] and a gene encoding a DNA polymerase IV. The replicon is a member of the pAMbeta1 family of theta-replicating plasmids and its proteins are most closely related to plasmids from the genera *Bacillus* (pXO2, pAW63, pBT9727), *Streptococcus* (pSM19035) and *Enterococcus* (pRE25, pVEF1, pVEF2) with protein identities ranging from between 36–56%. An exception to this homology was found to be RepC which shows no sequence similarity but a similar location, size and orientation as its putative functional homologs in the plasmids of the aforementioned genera. The genes encoding *repB*/C are overlapping indicating an operon. Interestingly, the translesion DNA polymerase has previously been suggested to stimulate spontaneous deletions during DNA repair [33,34] and could thus contribute to variation and adaptation of both plasmid and host genes when present.

To examine the relationship of the plasmid backbones we joined all fragments of the RepA proteins and used CLUSTALW [35] to create a phylogenetic tree (Figure 1). This methodology clearly confirms the relatedness of these plasmids to those present in other gram-positive strains and separated the plasmids of genus *Listeria* into two distinct phylogenetic groups consisting of *L. monocytogenes* serotypes 7, 1/2c, 1/2b, 4b FSL N1-017 and serogroup 4 DRDC8 in one cluster and *L. monocytogenes* serotype 1/2a, 4b H7858, *L. innocua* and *L. grayi* in the other. The plasmid of *L. monocytogenes* 1/2b strain F2-515 was an exception to this since it clustered with group 2 confirming observations from a previous study [23]. Plasmid sizes generally correlate with the clustering based on the replication initiation protein ranging from 32–57 kb in group 1 and 77–83 kb in group 2 (Table 1), again with the exception of F2-515 which belongs to group 2 but has a size similar to plasmids of group 1.

Comparative genetic analysis

The replicon-based distinction is mirrored by the gene content to some extent, which indicates a highly similar set of genes for

Table 1. General features of 14 plasmids of genus *Listeria*.

Host	Plasmid	Isolation	Status	Length [bp]	ORFs[a]	MGEs[b]	Mean Number of Amino Acids per CDS	Source/ Accession
L. monocytogenes 1/2b Lm1	pLM33	cheese	closed	32307	36	9	258	GU244485
L. monocytogenes 1/2a FSL F2-515	pF2-515	meat	contigs (11)	37163	61	12	131	Broad Institute[c]
L. monocytogenes 7 UG1 SLCC2482	pLM7UG1	human	closed	50100	55	13	260	FR667690
L. monocytogenes 1/2c UG1 SLCC2372	pLM1-2cUG1	human	closed	50100	54	13	264	FR667691
L. monocytogenes 1/2b FSL J1.194	pJ1-194	human	contigs (1)	57536	69	16	223	Broad Institute[c]
L. monocytogenes 1/2b UG1 SLCC2755	pLM1-2bUG1	human	closed	57780	63	16	261	FR667692
L. monocytogenes 1/2b FSL R2-503	pR2-503	human	contigs (3)	56540	86	20	159	Broad Institute[c]
L. monocytogenes 4b FSL N1-017	pN1-017	trout	contigs (3)	56037	62	13	245	Broad Institute[c]
L. monocytogenes 1/2a 08-5578	pLM5578	human	closed	77054	76	11	291	CP001603
L. monocytogenes 1/2a J0161	pLMJ0161	human	contigs (2)	82700	90	10	266	Broad Institute[c]
L. monocytogenes 4b H7858	pLM80	meat	contigs (2)	81588	88	11	264	AADR01000010, AADR01000058
L. grayi subspecies *grayi* UG1 DSM20601	pLGUG1	chinchilla	closed	79249	99	8	224	FR667693
L. innocua 6a Clip11262	pLI100	cheese	closed	81905	84	24	273	AL592102
L. monocytogenes 4 DRDC8	pCT100	milk	closed	37279	34	6	292	U15554

[a]Open Reading Frames.
[b]Mobile Genetic Elements.
[c]http://www.broad.mit.edu/annotation/genome/listeria_group.
Plasmid length does not include spacers which were inserted between contigs. All genes automatically predicted by GenDB [58] to encode a recombinase, transposase, integrase, invertase or resolvase are denoted as mobile genetic element (MGE).
doi:10.1371/journal.pone.0012511.t001

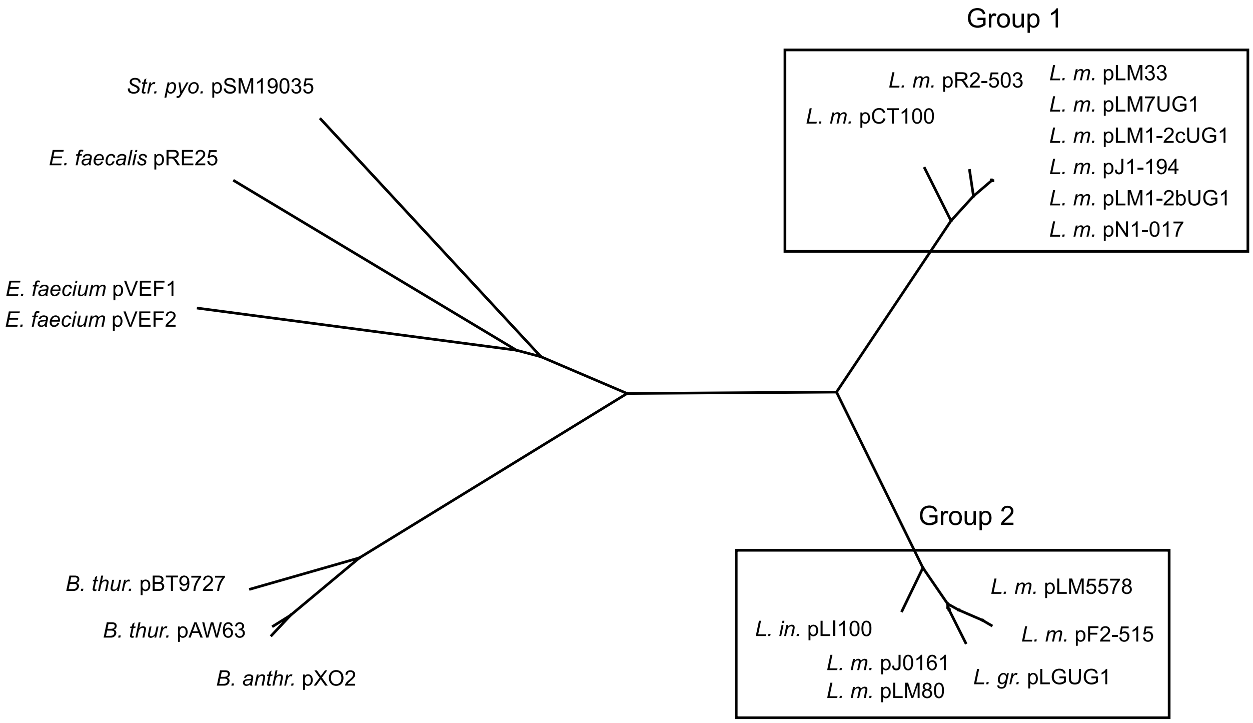


Figure 1. Phylogenetic tree of the replication initiation protein. Phylogenetic tree based on the replication initiation protein RepA of plasmids of genus *Listeria* and related genera. In the case of pF2-515 and pN1-017 multiple proteins had to be merged due to premature stopcodons. Clustalw [35] was used to create the multiple sequence alignment which was visualized using Dendroscope [60]. The clustering of the replication initiation proteins shows a clear separation into two phylogenetic groups.
doi:10.1371/journal.pone.0012511.g001

most plasmids of group 1, with a more heterogenous distribution for group 2 (Figure 2). Genes were considered homologs if BlastP found a sequence identity of at least 30% covering more than 80% of both proteins (Table S2). Apart from the replicon no other feature is conserved overall, but all plasmids contain a cadmium resistance operon (*cadA/C*) [18] with the exception of pLGUG1 which lacks *cadC* and harbors a transposase at that relative position.

Apart from pCT100, plasmids of group 1 are closely related to each other and mainly differ by two putative indel events. Most of them can be grouped into two types which will be referred to as pLisI (pLM7UG1, pLM1-2cUG1) and pLisII (pJ1-194, pR2-503, pLM1-2bUG1, pN1-017). The smallest plasmid pLM33 was previously described to contain multiple transposases, remnants of *Listeria* phage A006, CRISPR associated protein Cas5 implied in phage defense and a Clp protease related to *Lactobacillus* which is involved in environmental stress response [23]. Despite clustering with group 2 according to its rep-protein, the sequence of plasmid pF2-515 shows a much higher homology to plasmids of group 1 which is not reflected by its gene-content due to a large number of false stop codons leading to a computed average protein length of only 131 amino acids. The indel between pLM33 and pLisI is 18 kb in size and contains multiple transposases, a copper-transporting P-type ATPase and a multi-copper oxidase (MCO) implied in copper detoxification [36]. Either one or both copper resistance genes show putatively premature stopcodons in pLM7UG1, pJ1-194, pR2-503 and pN1-017. All plasmids excepting the smallest plasmids, pLM33, pF2-515 and pCT100 contained a protein with a FIC domain (filamentation induced by cyclic adenosine monophosphate) which is implicated in the disruption of cellular functions following transfer to the host cell cytoplasm during infection [37]. The plasmids of type pLisI and pLisII differ by 6 kb. This region encodes a transposase, a cadmium-transporting ATPase as well as an NADH peroxidase and a periplasmic component of an ABC-type glycine/betaine transport system closely related to *Aerococcus viridans* ATCC 11563 with an identity of 97% and 55% respectively. NADH peroxidases are described as being necessary for decomposition of hydrogen peroxide accumulated during aerobic growth [38]. This could play a role in intracellular survival against hydrogen peroxide stress [39] as well as in defense against disinfectants which are also known to induce general and oxidative stress responses [40]. All plasmids of group 1 apart from pCT100, and pLGUG1 of group 2, carry the sequence of the PemIK toxin/antitoxin stable maintenance system described for *Lactobacillus salivarius* UCC118 plasmid pSF118–20 [41]. In most cases one of the genes (pLM7UG1, pLM1-2cUG1, pJ1-194, pLM1-2bUG1, pR2-503) could not be identified by the gene prediction, which implies a decay of this functionality in those strains.

Only half of the sequence of pCT100 is shared with any other plasmid of the genus *Listeria*, including the replicon, a cadmium resistance system, a copper transporter and an insertion of 6 kb shared with pLI100 related to 12 kb plasmid pEW104 of *Lactococcus lactis subsp. cremoris* W10. Sequences present in this 6 kb fragment include two genes probably involved in replication and a single-gene type I restriction and modification (R/M) system called *Lla*GI which was shown to confer decreased bacteriophage sensitivity to its host [42]. The other half of pCT100 consists of multiple genes indicated in copper detoxification and a Na^+-driven multidrug efflux pump which belongs to the MATE family (multidrug and toxic compound extrusion) [43].

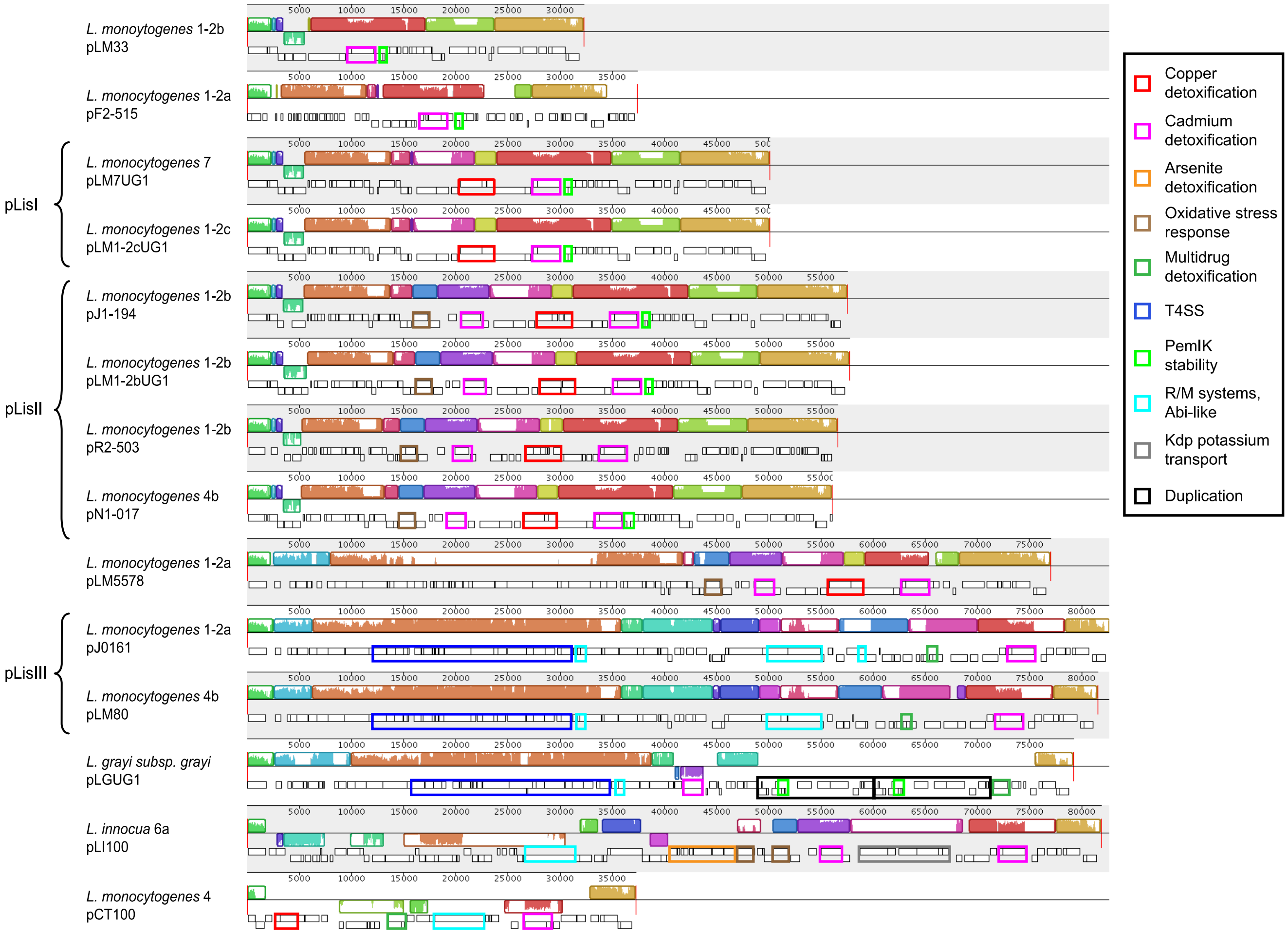


Figure 2. Complete alignment of plasmid sequences. Whole-sequence alignment of 14 plasmids of genus *Listeria* as computed by Mauve [59]. Relevant loci are marked specifically. These include genes involved in heavy metal detoxification (copper, cadmium, arsenite), multidrug resistance (MDR: SMR, MATE), phage defense (R/M systems, Abi), oxidative stress response, an incomplete type IV secretion system (T4SS), a PemIK stable inheritance module, a Kdp-type potassium transport system and a sequence duplication. It should be noted that Mauve was not able to identify all homologies due to algorithmic limitations.
doi:10.1371/journal.pone.0012511.g002

All plasmids of group 2 excepting pLGUG1 showed extensive homology to plasmids of group 1 further indicating a common ancestor. Plasmids pLM5578 and pLI100 share 76% and 45% nucleotide identity with pLisII respectively, with only 18% identity for the plasmids pJ0161 and pLM80. The latter two plasmids are closely related to each other and will be referred to as type pLisIII. Both contain a 40 kb region that is similar to plasmid pXO2 from *B. anthracis* which can also be found in pLGUG1 and to some extent in pLM5578 which is thought to include an incomplete type IV secretion system [22]. This system was shown to be insufficient for conjugation in pXO2 [31]. They also share a locus of 20 kb containing 24 genes including multiple transposases and two restriction modifications systems, one of them related to type III R/M system LlaFl of *Lactococcus lactis* [44]. This locus also harbors a gene encoding a triphenylmethane reductase described for the degradation of toxic synthetic dyes in *Citrobacter* [45] and a system of two genes *ebrAB*, which create a heterodimer channel involved in multidrug efflux in *Bacillus subtilis* [46]. The latter system belongs to the small multidrug resistance family (SMR) and is implied in resistance towards ethidium bromide and quarternary ammonium compounds [47] commonly found in disinfectants and could support persistence in food processing environments. Plasmid pLMJ0161 contains a specific insertion of a gene encoding an Abi-like protein implied in phage resistance [48]. The second half of pLGUG1 harbors a specific insertion spanning 24 kb which consists of a duplicated sequence comprising 17 genes most of them hypothetical. In addition this locus includes a Tn552-family transposase and a PemI/PemK post segregational killing system. Other pLGUG1 specific genes encode a MATE family multidrug efflux pump distantly related to the one found in pCT100. In pLI100 regions homologous to pLisII are interrupted by multiple pLI100-specific insertions. One of these consists of six genes related to potassium transport which span a region of 10 kb and are located on the same strand implying an operon. This Kdp-ATPase system [49] consists of a two-component signal transduction system (*kdpD/E*) and a potassium-transporting ATPase (*kdpA/B/C*). It is widely distributed among bacteria and archeae and plays a vital role in osmotic adaptation and pH regulation [50]. A homologue of this system exists in the chromosome of all fully sequenced strains of genus *Listeria* (data not shown) and contributes to growth during osmotic stress and low temperature in *L. monocytogenes* [51]. Another pLI100-specific insertion is an arsenite resistance operon related to integrative conjugative element ICESde3396 of *Streptococcus dysgalactiae subsp. equisimilis* strain

NS3396 (EU142041) spanning 12 kb and consisting of seven genes which may contribute to the survival of *L. innocua* in the environment. Adjacent to this region a coenzyme A disulfide reductase gene was identified which is implied in oxidative stress response in *Borrelia burgdorferi* bb0728 [52].

In general, the plasmids of genus *Listeria* harbor a large and diverse number of mobile genetic elements. Between 6 and 24 genes per plasmid were annotated as transposase, resolvase, integrase, recombinase or invertase. This suggests that plasmids may act as an evolutionary sink for mobile genetic elements which may have shaped the diversity and evolution of plasmids in the genus *Listeria*.

Small non-coding RNA

Recently small non-coding RNAs (sRNAs) have become a focus of research because of their roles in bacterial regulatory mechanisms [53]. Interestingly we could identify multiple putative sRNAs on listerial plasmids. One class of RNA was already described for plasmids where they are predominantly implied in replication control, segregation and conjugation [54]. To identify additional sRNAs a software called sRNAdb (J. Pischimarov, unpublished), which employs BlastN, was used to find sequence similarities previously described in *L. monocytogenes*. Using a cutoff of 60% identity and 80% coverage, four putative sRNAs [55] were identified on plasmids of genus *Listeria* (Table S2). These comprise two pairs of homologues being *rli28/rli50* and *rli44/rli46*. Homologous sequences to *rli28/rli50* could be identified in pLM80-like, pLI100 and pCT100 while *rli44/rli46* could be found in pLM33, pF2-515, pLisII, pLisIII and pLGUG1.

Conclusion

Here we report on the completion of four new plasmid sequences, including two novel plasmids from *L. monocytogenes* serotype 1/2c and 7 strains as well as one from the species *L. grayi*. In the comparative analysis presented here we compared sequences of 14 plasmids from three species using additional sequences either previously published or deposited in databases. We found that all plasmids share a common replicon-type related to theta-replicating plasmid pAMbeta1 [56] implying a common ancestor. Nonetheless a phylogenetic division must have occurred when considering the replication initiation protein. This division was mostly mirrored by the genetic content which showed clear distinctions between those groups apart from two atypical plasmids (pF2-515, pCT100). Based on regions of synteny, we are able to trace diversification and evolution driven by indels that account for the range of plasmid sizes detected. The presence of a commonly occurring translesion repair DNA polymerase on all plasmids suggests a mechanism by which genetic deletions are generated. Since plasmids of genus *Listeria* are related to *Bacillus*, *Enterococcus* and *Streptococcus* and were described to be transferable between some of these genera [12,21], it is likely that exchange amongst these bacteria takes place in many different environmental niches e.g. gut and soil. Also, the unexpected detection of a large number of mobile genetics elements present on these plasmids imply that these could be involved in increasing genetic diversity or even altering gene expression both at chromosomal and episomal sites. Furthermore we found multiple independent systems involved in defense against phages (type I and III restriction systems, Abi-like) in group 2 implying a role for plasmids in the dissemination of these genes to ward off bacteriophage infection. The detection of small non-coding RNAs on a number of plasmids that were homologous to those present on the chromosome of *L. monocytogenes* EGD-e suggests that sRNAs might be transfered via plasmid conjugation.

The overrepresentation of plasmids in studies examining strains from food and the environment [14,16] and in recurrent *L. monocytogenes* strains sampled from food/processing facilities [17] is an intriguing observation. However, the presence of multiple genes involved in heavy metal resistance (cadmium, copper, arsenite) as well as multidrug efflux (MDR, SMR, MATE) and oxidative stress response (peroxidase, reductase) on listerial plasmids could assist survival and their presence may have resulted from selective pressure due to the use of disinfectants in food processing environments. Finally, we note that MDR efflux pumps have recently been shown to promote cyclic diadenosine monophosphate (c-di-AMP) as a secreted molecule able to trigger the cytosolic host response following infection [57]. The implication for the presence of MDR genes on plasmids and their role in listerial infection now deserves further scrutiny.

Availability

Four plasmid sequences from this article have been deposited in the EMBL/GenBank database under accession numbers FR667690 (pLM7UG1), FR667691 (pLM1-2cUG1), FR667692 (pLM1-2bUG1) and FR667693 (pLGUG1). An EMBL-formatted version of all plasmids can be downloaded (http://bioinfo.mikrobio.med.uni-giessen.de/publications/listeria_plasmids/listeria_plasmids_embl.tar.gz). The data can also be compared and retrieved using Geco (http://bioinfo.mikrobio.med.uni-giessen.de/geco2plasmids/).

Materials and Methods

Public data sources

Contigs of five gapped *L. monocytogenes* plasmids were downloaded from the homepage of the Broad Institute (http://www.broad.mit.edu/annotation/genome/listeria_group) originating from strains FSL J1.194 (2.44), FSL R2-503 (2.52, 2.53, 2.54), FSL N1-017 (2.75, 2.76, 2.77), FSL F2-515 (2.1405, 2.1406, 2.1407, 2.1408, 2.1409, 2.1410, 2.1411, 2.1412, 2.1413, 2.1414, 2.1415) and J0161 (1.50, 1.51). The plasmids pLM33 of *L. monocytogenes* Lm1 (GU244485), pCT100 of *L. monocytogenes* DRDC8 (U15554), pLM80 of *L. monocytogenes* H7858 (AADR01000010, AADR01000058), pLM5578 of *L. monocytogenes* 08-5578 (CP001603) and pLI100 of *L. innocua* Clip11262 (AL592102) were downloaded from the GenBank database (http://www.ncbi.nlm.nih.gov/genbank/index.html).

Isolation and sequencing

The remaining strains from *L. monocytogenes* 7 UG1 SLCC2482, *L. monocytogenes* 1/2c UG1 SLCC2372, *L. monocytogenes* 1/2b UG1 SLCC2755 and *L. grayi subspecies grayi* UG1 DSM20601 were isolated using Epicentre's MasterPure gram-positive DNA purification kit as recommended by the manufacturer. The DNA was sequenced on a 454 GS-FLX System to coverages between 16–57x. The resulting reads were assembled *de novo* with the 454 Newbler assembler and mapped vs. published plasmids to identify homologous contigs. PCR-based techniques were used to close the remaining gaps which were sequenced with Sanger ABI Big Dye technology. The sequencing was performed by Roche (Germany), Goettingen Genomics Laboratory (Goettingen, Germany) and Agowa (Berlin, Germany).

Bioinformatics

All contigs of gapped plasmids were scaffolded according to finished plasmids and joined to a consecutive sequence using the spacer "nnnnnttaattaattaannnnn" to prevent the gene prediction from crossing contig borders. All sequences were then reordered to a putative origin adjacent to the replication initiation gene (*repA*) as described for the homolog replicon of *B. anthr*acis plasmid pXO2 [29] and automatically annotated using the GenDB system [58]. The annotation was corrected based on a comparative syntheny analysis as offered by Geco [26]. In order to compute a phylogenetic tree Clustalw [35] was applied on replication initiation proteins using standard parameters. A multiple sequence alignment of the complete plasmid sequences was created with the Mauve software [59] using a progressive alignment including seed families to increase sensitivity. Mauve was not able to identify all homologies correctly with any combination of parameters. The chosen alignment is the optimal result considering false positives/negatives (data not shown).

Supporting Information

Table S1 This matrix shows a single-linkage clustering of all proteins of 14 plasmids of the genus *Listeria* using a minimum of 30% amino acid identity and 80% coverage. Annotation was included from the first protein of each cluster starting from the left. Clusters were sorted according to their size to ensure that mutually conserved proteins can be found at the top of the list while specific ones are moved to the bottom.

Found at: doi:10.1371/journal.pone.0012511.s001 (0.07 MB XLS)

Table S2 Using BlastN with a cutoff of 60% identity and 80% coverage four sRNAs of Listeria monocytogenes EGD-e [55] could be identified on various plasmids.

Found at: doi:10.1371/journal.pone.0012511.s002 (0.02 MB XLS)

Acknowledgments

We thank Claudia Zörb, Nelli Schklarenko and Burkhard Linke for excellent technical assistance.

Author Contributions

Conceived and designed the experiments: TH TC. Performed the experiments: SV. Analyzed the data: CK. Contributed reagents/materials/analysis tools: CK JP SO AG RD. Wrote the paper: CK TH TC.

References

1. Graves LM, Helsel LO, Steigerwalt AG, Morey RE, Daneshvar MI, et al. (2010) *Listeria marthii* sp. nov., isolated from the natural environment, Finger Lakes National Forest. Int J Syst Evol Microbiol 60: 1280–1288.
2. Glaser P, Frangeul L, Buchrieser C, Rusniok C, Amend A, et al. (2001) Comparative genomics of *Listeria* species. Science 294: 849–852.
3. Hain T, Steinweg C, Kuenne CT, Billion A, Ghai R, et al. (2006) Whole-genome sequence of *Listeria welshimeri* reveals common steps in genome reduction with *Listeria innocua* as compared to *Listeria monocytogenes*. J Bacteriol 188: 7405–7415.
4. Steinweg C, Kuenne CT, Billion A, Mraheil MA, Domann E, et al. (2010) Complete genome sequence of *Listeria seeligeri*, a nonpathogenic member of the genus *Listeria*. J Bacteriol 192: 1473–1474.
5. Leclercq A, Clermont D, Bizet C, Grimont PA, Le Fleche-Mateos A, et al. (2009) *Listeria rocourtiae* sp. nov. Int J Syst Evol Microbiol.
6. Chakraborty T, Hain T, Domann E (2000) Genome organization and the evolution of the virulence gene locus in *Listeria* species. Int J Med Microbiol 290: 167–174.
7. Guillet C, Join-Lambert O, Le MA, Leclercq A, Mechai F, et al. (2010) Human listeriosis caused by *Listeria ivanovii*. Emerg Infect Dis 16: 136–138.
8. Azizoglu RO, Osborne J, Wilson S, Kathariou S (2009) Role of growth temperature in freeze-thaw tolerance of *Listeria* spp. Appl Environ Microbiol 75: 5315–5320.
9. McLauchlin J, Mitchell RT, Smerdon WJ, Jewell K (2004) *Listeria monocytogenes* and listeriosis: a review of hazard characterisation for use in microbiological risk assessment of foods. Int J Food Microbiol 92: 15–33.
10. Kathariou S (2002) *Listeria monocytogenes* virulence and pathogenicity, a food safety perspective. J Food Prot 65: 1811–1829.
11. Fistrovici E, Collins-Thompson DL (1990) Use of plasmid profiles and restriction endonuclease digest in environmental studies of *Listeria* spp. from raw milk. Int J Food Microbiol 10: 43–50.
12. Perez-Diaz JC, Vicente MF, Baquero F (1982) Plasmids in *Listeria*. Plasmid 8: 112–118.
13. Peterkin PI, Gardiner MA, Malik N, Idziak ES (1992) Plasmids in *Listeria monocytogenes* and other *Listeria* species. Can J Microbiol 38: 161–164.
14. Lebrun M, Loulergue J, Chaslus-Dancla E, Audurier A (1992) Plasmids in *Listeria monocytogenes* in relation to cadmium resistance. Appl Environ Microbiol 58: 3183–3186.
15. Kolstad J, Caugant DA, Rorvik LM (1992) Differentiation of *Listeria monocytogenes* isolates by using plasmid profiling and multilocus enzyme electrophoresis. Int J Food Microbiol 16: 247–260.
16. McLauchlin J, Hampton MD, Shah S, Threlfall EJ, Wieneke AA, et al. (1997) Subtyping of *Listeria monocytogenes* on the basis of plasmid profiles and arsenic and cadmium susceptibility. J Appl Microbiol 83: 381–388.
17. Harvey J, Gilmour A (2001) Characterization of recurrent and sporadic *Listeria monocytogenes* isolates from raw milk and nondairy foods by pulsed-field gel electrophoresis, monocin typing, plasmid profiling, and cadmium and antibiotic resistance determination. Appl Environ Microbiol 67: 840–847.
18. Lebrun M, Audurier A, Cossart P (1994) Plasmid-borne cadmium resistance genes in *Listeria monocytogenes* are similar to *cadA* and *cadC* of *Staphylococcus aureus* and are induced by cadmium. J Bacteriol 176: 3040–3048.
19. Poyart-Salmeron C, Carlier C, Trieu-Cuot P, Courtieu AL, Courvalin P (1990) Transferable plasmid-mediated antibiotic resistance in *Listeria monocytogenes*. Lancet 335: 1422–1426.
20. Hadorn K, Hachler H, Schaffner A, Kayser FH (1993) Genetic characterization of plasmid-encoded multiple antibiotic resistance in a strain of *Listeria monocytogenes* causing endocarditis. Eur J Clin Microbiol Infect Dis 12: 928–937.
21. Flamm RK, Hinrichs DJ, Thomashow MF (1984) Introduction of pAM beta 1 into *Listeria monocytogenes* by conjugation and homology between native *L. monocytogenes* plasmids. Infect Immun 44: 157–161.
22. Nelson KE, Fouts DE, Mongodin EF, Ravel J, DeBoy RT, et al. (2004) Whole genome comparisons of serotype 4b and 1/2a strains of the food-borne pathogen *Listeria monocytogenes* reveal new insights into the core genome components of this species. Nucleic Acids Res 32: 2386–2395.
23. Canchaya C, Giubellini V, Ventura M, de Los Reyes-Gavilan CG, Margolles A (2010) Mosaic-like sequences containing transposon, phage and plasmid elements among *Listeria monocytogenes* plasmids. Appl Environ Microbiol.
24. Gilmour MW, Graham M, Van DG, Tyler S, Kent H, et al. (2010) High-throughput genome sequencing of two *Listeria monocytogenes* clinical isolates during a large foodborne outbreak. BMC Genomics 11: 120.
25. Orsi RH, Borowsky ML, Lauer P, Young SK, Nusbaum C, et al. (2008) Short-term genome evolution of *Listeria monocytogenes* in a non-controlled environment. BMC Genomics 9: 539.
26. Kuenne CT, Ghai R, Chakraborty T, Hain T (2007) GECO—linear visualization for comparative genomics. Bioinformatics 23: 125–126.
27. Weaver KE, Clewell DB, An F (1993) Identification, characterization, and nucleotide sequence of a region of *Enterococcus faecalis* pheromone-responsive plasmid pAD1 capable of autonomous replication. J Bacteriol 175: 1900–1909.
28. Wilcks A, Smidt L, Okstad OA, Kolsto AB, Mahillon J, et al. (1999) Replication mechanism and sequence analysis of the replicon of pAW63, a conjugative plasmid from *Bacillus thuringiensis*. J Bacteriol 181: 3193–3200.
29. Tinsley E, Naqvi A, Bourgogne A, Koehler TM, Khan SA (2004) Isolation of a minireplicon of the virulence plasmid pXO2 of *Bacillus anthracis* and characterization of the plasmid-encoded RepS replication protein. J Bacteriol 186: 2717–2723.
30. Francia MV, Fujimoto S, Tille P, Weaver KE, Clewell DB (2004) Replication of *Enterococcus faecalis* pheromone-responding plasmid pAD1: location of the minimal replicon and oriV site and RepA involvement in initiation of replication. J Bacteriol 186: 5003–5016.
31. Van der Auwera GA, Andrup L, Mahillon J (2005) Conjugative plasmid pAW63 brings new insights into the genesis of the *Bacillus anthracis* virulence plasmid pXO2 and of the *Bacillus thuringiensis* plasmid pBT9727. BMC Genomics 6: 103.
32. Francia MV, Weaver KE, Goicoechea P, Tille P, Clewell DB (2007) Characterization of an active partition system for the *Enterococcus faecalis* pheromone-responding plasmid pAD1. J Bacteriol 189: 8546–8555.
33. Friedberg EC, Wagner R, Radman M (2002) Specialized DNA polymerases, cellular survival, and the genesis of mutations. Science 296: 1627–1630.
34. Koskiniemi S, Andersson DI (2009) Translesion DNA polymerases are required for spontaneous deletion formation in *Salmonella typhimurium*. Proc Natl Acad Sci U S A 106: 10248–10253.
35. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673–4680.
36. Kosman DJ (2010) Multicopper oxidases: a workshop on copper coordination chemistry, electron transfer, and metallophysiology. J Biol Inorg Chem 15: 15–28.
37. Roy CR, Mukherjee S (2009) Bacterial FIC Proteins AMP Up Infection. Sci Signal 2: e14.

38. Gibson CM, Mallett TC, Claiborne A, Caparon MG (2000) Contribution of NADH oxidase to aerobic metabolism of *Streptococcus pyogenes*. J Bacteriol 182: 448–455.
39. La CS, Sauvageot N, Giard JC, Benachour A, Posteraro B, et al. (2007) Comparative study of the physiological roles of three peroxidases (NADH peroxidase, Alkyl hydroperoxide reductase and Thiol peroxidase) in oxidative stress response, survival inside macrophages and virulence of *Enterococcus faecalis*. Mol Microbiol 66: 1148–1163.
40. Ceragioli M, Mols M, Moezelaar R, Ghelardi E, Senesi S, et al. (2010) Comparative transcriptomic and phenotypic analysis of the responses of *Bacillus cereus* to various disinfectant treatments. Appl Environ Microbiol 76: 3352–3360.
41. Fang F, Flynn S, Li Y, Claesson MJ, van Pijkeren JP, et al. (2008) Characterization of endogenous plasmids from *Lactobacillus salivarius* UCC118. Appl Environ Microbiol 74: 3216–3228.
42. Madsen A, Josephsen J (2001) The *Lla*GI restriction and modification system of *Lactococcus lactis* W10 consists of only one single polypeptide. FEMS Microbiol Lett 200: 91–96.
43. Kuroda T, Tsuchiya T (2009) Multidrug efflux transporters in the MATE family. Biochim Biophys Acta 1794: 763–768.
44. Su P, Im H, Hsieh H, Kang'A S, Dunn NW (1999) *Lla*FI, a type III restriction and modification system in *Lactococcus lactis*. Appl Environ Microbiol 65: 686–693.
45. Kim MH, Kim Y, Park HJ, Lee JS, Kwak SN, et al. (2008) Structural insight into bioremediation of triphenylmethane dyes by Citrobacter sp. triphenylmethane reductase. J Biol Chem 283: 31981–31990.
46. Zhang Z, Ma C, Pornillos O, Xiu X, Chang G, et al. (2007) Functional characterization of the heterooligomeric EbrAB multidrug efflux transporter of *Bacillus subtilis*. Biochemistry 46: 5218–5225.
47. Masaoka Y, Ueno Y, Morita Y, Kuroda T, Mizushima T, et al. (2000) A two-component multidrug efflux pump, EbrAB, in *Bacillus subtilis*. J Bacteriol 182: 2307–2310.
48. Chopin MC, Chopin A, Bidnenko E (2005) Phage abortive infection in lactococci: variations on a theme. Curr Opin Microbiol 8: 473–479.
49. Ballal A, Basu B, Apte SK (2007) The Kdp-ATPase system and its regulation. J Biosci 32: 559–568.
50. Booth IR (1985) Regulation of cytoplasmic pH in bacteria. Microbiol Rev 49: 359–378.
51. Brondsted L, Kallipolitis BH, Ingmer H, Knochel S (2003) *kdpE* and a putative RsbQ homologue contribute to growth of *Listeria monocytogenes* at high osmolarity and low temperature. FEMS Microbiol Lett 219: 233–239.
52. Boylan JA, Hummel CS, Benoit S, Garcia-Lara J, Treglown-Downey J, et al. (2006) Borrelia burgdorferi bb0728 encodes a coenzyme A disulphide reductase whose function suggests a role in intracellular redox and the oxidative stress response. Mol Microbiol 59: 475–486.
53. Waters LS, Storz G (2009) Regulatory RNAs in bacteria. Cell 136: 615–628.
54. Brantl S (2007) Regulatory mechanisms employed by cis-encoded antisense RNAs. Curr Opin Microbiol 10: 102–109.
55. Toledo-Arana A, Dussurget O, Nikitas G, Sesto N, Guet-Revillet H, et al. (2009) The *Listeria* transcriptional landscape from saprophytism to virulence. Nature 459: 950–956.
56. Bruand C, Ehrlich SD (1998) Transcription-driven DNA replication of plasmid pAMbeta1 in *Bacillus subtilis*. Mol Microbiol 30: 135–145.
57. Woodward JJ, Iavarone AT, Portnoy DA (2010) c-di-AMP Secreted by Intracellular *Listeria monocytogenes* Activates a Host Type I Interferon Response. Science.
58. Meyer F, Goesmann A, McHardy AC, Bartels D, Bekel T, et al. (2003) GenDB—an open source genome annotation system for prokaryote genomes. Nucleic Acids Res 31: 2187–2195.
59. Darling AC, Mau B, Blattner FR, Perna NT (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res 14: 1394–1403.
60. Huson DH, Richter DC, Rausch C, Dezulian T, Franz M, et al. (2007) Dendroscope: An interactive viewer for large phylogenetic trees. BMC Bioinformatics 8: 460.

Genome-Wide Identification of Small RNAs in the Opportunistic Pathogen *Enterococcus faecalis* V583

Kouki Shioya[1], Charlotte Michaux[1], Carsten Kuenne[2], Torsten Hain[2], Nicolas Verneuil[1], Aurélie Budin-Verneuil[1], Thomas Hartsch[3], Axel Hartke[1], Jean-Christophe Giard[1]*

[1] Laboratoire de Microbiologie de l'Environnement, EA956-USC INRA 2017-IFR146 ICORE, University of Caen, Caen, France, [2] Institute of Medical Microbiology, Justus-Liebig-University, Giessen, Germany, [3] Genedata AG, Basel, Switzerland

Abstract

Small RNA molecules (sRNAs) are key mediators of virulence and stress inducible gene expressions in some pathogens. In this work we identify sRNAs in the Gram positive opportunistic pathogen *Enterococcus faecalis*. We characterized 11 sRNAs by tiling microarray analysis, 5′ and 3′ RACE-PCR, and Northern blot analysis. Six sRNAs were specifically expressed at exponential phase, two sRNAs were observed at stationary phase, and three were detected during both phases. Searches of putative functions revealed that three of them (EFA0080_EFA0081 and EFB0062_EFB0063 on pTF1 and pTF2 plasmids, respectively, and EF0408_EF04092 located on the chromosome) are similar to antisense RNA involved in plasmid addiction modules. Moreover, EF1097_EF1098 shares strong homologies with tmRNA (bi-functional RNA acting as both a tRNA and an mRNA) and EF2205_EF2206 appears homologous to 4.5S RNA member of the Signal Recognition Particle (SRP) ribonucleoprotein complex. In addition, proteomic analysis of the ΔEF3314_EF3315 sRNA mutant suggests that it may be involved in the turnover of some abundant proteins. The expression patterns of these transcripts were evaluated by tiling array hybridizations performed with samples from cells grown under eleven different conditions some of which may be encountered during infection. Finally, distribution of these sRNAs among genome sequences of 54 *E. faecalis* strains was assessed. This is the first experimental genome-wide identification of sRNAs in *E. faecalis* and provides impetus to the understanding of gene regulation in this important human pathogen.

Citation: Shioya K, Michaux C, Kuenne C, Hain T, Verneuil N, et al. (2011) Genome-Wide Identification of Small RNAs in the Opportunistic Pathogen *Enterococcus faecalis* V583. PLoS ONE 6(9): e23948. doi:10.1371/journal.pone.0023948

Editor: Roy Martin Roop II, East Carolina University School of Medicine, United States of America

Received April 11, 2011; **Accepted** July 28, 2011; **Published** September 2, 2011

Funding: CM is funded by a thesis grant awarded by the Ministère de l'Enseignement Supérieur et de la Recherche. This study was supported by grants of the "Agence Nationale de la Recherche" in the frame of the transnational ERA-NET PathoGenoMics program (ANR-08-PATH-008-01), and by the Region of Lower Normandy (CRBN). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: Thomas Hartsch is an employee of Genedata AG (Basel, Switzerland), and internal legal approval was granted to publish this study. It was a collaborative effort between University of Caen, University of Giessen, and Genedata company groups and is not related to any products manufactured or marketed. There are no patents to declare. The authors hereby confirm that studies presented do not alter adherence to all the PLoS ONE policies on sharing data and materials, as detailed in the PLoS ONE guide for authors.

* E-mail: jean-christophe.giard@unicaen.fr

Introduction

Some RNA molecules such as riboswitches, transfer-messenger RNA (tmRNA) and small non-cording RNAs (sRNAs) act usually as post-transcriptional regulators in bacteria [1]. sRNAs have become increasingly recognized as an emerging class of gene expression regulators for cellular processes, stress response and virulence genes and their transcription is tightly regulated and induced by distinct environmental conditions [2]. Bacterial sRNAs found on chromosomes are typically 50–400 nucleotides in length and frequently encoded in intergenic regions (IGRs). They may bind to the imperfect complementary sequence of the ribosome binding region of the target mRNA, which is often encoded at separate loci, thus inhibiting 30S ribosomal subunit association and translational initiation [1,3]. In some Gram positive and Gram negative species such as *Escherichia coli* [4] and *Listeria monocytogenes* [5], the formation of sRNA-mRNA duplex requires the RNA chaperon protein Hfq [6,7] leading to an increase of mRNA degradation by ribonucleases such as RNase E and RNase III [2]. Some sRNAs located in plasmids and phages act as antisense RNAs on *cis*-encoded mRNAs and mainly control replication initiation, conjugation efficiency and transposition [8,9]. In addition, plasmid-encoded sRNAs, called *hak/sok* system of *E. coli* plasmid R1 [10] and *par* system of *Enterococcus faecalis* pAD1 [11], stabilize their host plasmids by programming for death any cell that loses the plasmid [9,12].

In recent years, several bioinformatic approaches have been performed to identify putative sRNAs in bacterial genomes including *E. coli*, *L. monocytogenes*, *Bacillus subtilis* and *Pseudomonas aeruginosa*, and identified more than 200 sRNAs [13]. Recently, Livny *et al.* predicted *in silico* over 45,000 sRNA candidates from 932 bacterial genomes [14]. In parallel, different experimental strategies including cDNA sequencing, shotgun cloning and isolation from RNA-protein complex have been performed and sometimes lead to the discovery of new transcripts [15,16]. Tiling microarrays are powerful approaches to identify sRNAs on a genome-wide scale. Thus large numbers of sRNA candidates have been found in *Caulobacter crescentus*, *Streptococcus pyogenes*, *S. pneumoniae*, and *L. monocytogenes* genomes [17,18,19,20].

E. faecalis is a human commensal Gram-positive bacteria as well as one of the leading causes of hospital acquired infections in United States and Europe [21]. The first whole genome sequence of *E. faecalis* V583 strain (the first vancomycin resistant enterococci identified in U.S.A.) was determined in 2003 and 53 more sequences are now publically available [22]. *In silico* study performed by Livny *et al.* led to the prediction and annotation of 17 putative sRNA-encoding loci in *E. faecalis* [14]. Surprisingly, in comparison with *E. coli* and *B. subtilis*, the number of predicted sRNAs in V583 is roughly 10-fold lower, suggesting that this number is likely under-estimated. Recently, 45 sRNAs and 10 putative mRNAs have been identified in *E. faecalis* using *in silico* prediction combined with "5′tag-RACE" [23].

In this work, we developed custom-made tiling microarrays containing only IGRs of *E. faecalis* V583 chromosome and plasmids, and first performed hybridization with RNA extracted from exponential and stationary-phase cells. Fifty-three statistically significant positive signals were detected and the 12 putative sRNAs most highly expressed were selected for further characterization. Transcription of these candidates under several stress conditions was then analyzed.

Materials and Methods

Bacterial strain and growth conditions

All experiments were performed with *E. faecalis* V583 strain [24]. For our first tiling array assays, cells were grown at 37°C in M17 0.5% glucose medium and collected at exponential phase ($OD_{600} = 0.5$) and at 24 h stationary phase. Growth in BHI medium with or without aeration was tested. Cells were collected at exponential phase ($OD_{600} = 0.5$), onset of starvation ($OD_{600} = 2$) and late stationary phase (24 h). For experiments under stress conditions, bacterial cells were grown to $OD_{600} = 0.3$ in M17 medium and H_2O_2 (2 mM), lactic acid (pH 5.5), or bile salts (BS) (0.08%), were added before an additional 30 min incubation at 37°C. For the growth in urine and serum, *E. faecalis* was inoculated into human urine or horse serum (Eurobio, Courtaboeuf, Fr) during overnight. Cells were then pelleted and resuspended into fresh urine or serum for 3 hours at 37°C. Urine collected from four healthy volunteers was pooled, centrifuged and sterilized by filtration (0.22 µm-pore sizes). Written consent from all participants involved in our study was obtained. French CPP (Comité de Protection de Personnes) exempted this study from review because volunteers were informed of the goal of this study, no health information was collected and no biological analysis was performed on these samples.

RNA extraction and tiling microarray hybridization

Total RNA was extracted using Trizol reagent (Invitrogen, Carlsbad, CA) as described by Toledo-Arana *et al.* [20], with the following modifications. Bacterial cells were resuspended into 200 µl of "max bacterial enhancement reagent" (Invitrogen) and transferred into micro tubes containing glass beads and 400 µl acid phenol (Ambion, Austin, TX). Bacteria were mechanically lysed using Mixer Mill 200 (30/s, 30 min, Retsch, Haan, Germany). After centrifugation for 10 min at 14,000 g at 4°C, aqueous phase was transferred to 2 ml tubes containing 1 ml Trizol reagent, mixed and incubated for 5 min at room temperature (RT). 200 µl chloroform was added, mixed gently and incubated for 3 min at RT. Tubes were centrifuged for 15 min at 12,000 g at 4°C and aqueous phase was transferred into 2 ml tubes containing 200 µl chloroform, mixed gently and centrifuged again. RNAs contained in the aqueous phase were precipitated by addition of 500 µl isopropanol and incubated for 10 min at RT. After centrifugation, RNA pellets were washed with 75% ethanol and dried at RT. Purified RNA pellets were resuspended in DEPC-treated pure water.

To enhance detection sensitivity by enriching of sRNAs and removing non-sRNA, 10 µg RNA were fractionated using flashPAGE Fractionator (Applied Biosystems, Foster City, CA). Fractionated RNA was labelled using mirVana labelling kit (Applied Biosystems) and then hybridized onto the tiling array. 1745 "big intergenic regions (IGR)" (more than 49 nt) and 1070 "small IGR" (from 1 to 49 nt) have been deduced from *E. faecalis* V583 genome sequence. 50 nt long probes with an overlap of 15 nt were loaded on our IGR custom-made tiling arrays. rRNA and tRNA probes were used as positive control showing signal intensity of hybridization at least 10 fold the threshold level. Since the values of intensity observed in apparent untranslated regions were between 1000 and 2000, 2000 was used as threshold. For each experiment (one sample per growth condition) two chips were used; one corresponding to the forward, and one to the reverse strand. Production, hybridization and data collecting were carried out by Febit biomed GmbH Company (Heidelberg, Germany). The detection was carried out using streptavidin phycoerythrin at different exposure times. Data analyses and visualization were performed by Genedata Phylosopher Business Group (Basel, Switzerland). We have deposed the raw data at GEO/ArrayExpress under accession number GSE28741, we can confirm all details are MIAME compliant.

5′ and 3′ rapid amplification of cDNA ends (RACE) analysis

For these analysis, new RNA samples were prepared as described above. 5′ RACE was performed using 2nd Generation 5′/3′ RACE kit (Roche, Mannheim, Germany) according to the manufacturer's instructions. For polymerase chain reactions (PCR), we used Go Taq polymerase and its buffer (Promega, Madison, WI). The primers used for cDNA synthesis, and for the PCR reactions are listed in Table S1.

For 3′ RACE experiments, total RNAs were treated with poly(A) polymerase (Epicentre, Madison, WI) for 15 min at 37°C. After 3′ end RNA poly(A) tailing, cDNA was synthesized with QuantiTect Reverse Transcription kit (Qiagen, West Sussex, UK) and oligo(dT)-anchor primer supplied in 5′/3′ RACE kit. cDNA products were directly used as templates for PCR performed with the gene-specific primers (Table S1) and the respective PCR anchor primer. After sequencing, 5′ and 3′ ends sequences were determined.

Northern blotting

Northern blots were performed according to standard procedures [25]. Five µg of total RNA were separated on 1.2% formaldehyde agarose gel and transferred to Hybond N^+ membrane (Amersham, UK). 0.1–1 kb RNA Marker (Sigma, USA) was used to estimate the sizes of RNA bands. DNA oligonucleotides probes (Table S1) were labeled with α^{32}P-ATP using Terminal Deoxnucleotidyl Transferase Recombinant enzyme (Promega) as recommended by the manufactured protocol. Membranes were prehybridized for 1 h in hybridization buffer (0.25 M NaH_2PO_4, 0.25 M Na_2HPO_4, 5% SDS) at 45°C, followed by addition of labelled probes and overnight hybridization at 45°C. Membranes were washed with washing buffer (3×SSC buffer, 0.2% SDS) for 5 min at RT and were then exposed to storage phosphor screen (Packard Instrument Company, Mariden, CT) for 3 h.

In silico analysis

Rho-independent terminators were predicted with TransTerm (http://nbc11.biologie.uni-kl.de/framed/left/menu/auto/right/trans term/) [26]. Blast searches between *E. faecalis* strains were carried out using a species-level BLAST database (http://www.ncbi.nlm.nih.gov/). The Rfam database was employed to determine putative functions of sRNAs (http://rfam.sanger.ac.uk) [27]. In order to predict target genes for the identified sRNAs sRNATarget (http://ccb.bmi.ac.cn/sRNA-target/) [28] and IntaRNA (http://rna.informatik.uni-freiburg.de:8080/IntaRNA.jsp) [29] servers were used.

Construction of EF3314_EF3315 sRNA deletion mutant

For the deletion assay, a DNA fragment containing ligated upstream (869 bp) and downstream (839 bp) sequences of the EF3314_EF3315 sRNA, was cloned into plasmid pMAD [30] (see Table S1 for primers used). 1 µg of recombinant plasmid was finally used to transform competent cells. After electroporation, 300 µl of cell suspension was plated onto GM17 agar containing 50 µg ml^{-1} of erythromycin and X-Gal (100 µg ml^{-1}). Plates were incubated for 48 hours at 30°C. A few dark blue colonies were obtained and analysed for presence of the plasmid by PCR using primers *madR* and *madF* (Table S1). Some blue colonies were then cultured twice in GM17 liquid medium with erythromycin (50 µg ml^{-1}) at 45°C over-night. In the next step, the cultures were used to inoculate (0.05% v/v) GM17 liquid medium without antibiotic. The tubes were incubated for 6 hours at 30°C followed by incubation at 45°C over-night. This step was repeated 2 to 3 times. Serial dilutions of the culture were plated on GM17 agar containing 100 µg ml^{-1} of X-Gal and incubated for 48 hours at 45°C. White colonies were then isolated on GM17 agar with or without erythromycin. Antibiotic sensitive clones were analysed by PCR on the presence of a deleted sRNA.

Two-dimensional protein gel electrophoresis and protein identification

Protein samples from wild type and ΔEF3314_EF3315 mutant cells harvested in exponential growth phase were performed as described by Giard *et al.* [31]. First dimensional electrophoresis was carried out using 17 cm ReadyStrip^{TM} IPG Strips (pH 4–7) and Protean®IEF Cell apparatus (Bio-Rad Laboratories, Richmond, CA, USA) as recommended by the manufacturer. Second dimensions were performed in 14% polyacrylamide gels without stacking gel using the Millipore Investigator^{TM} 2-D electrophoresis system (Millipore, Bedford, MA, USA) as described by Giard *et al.* [31]. 2-D gels were then stained using Coomassie Blue. Spots of interest were excised from the gel, and peptides were digested by trypsin as described by Budin-Verneuil *et al.* [32]. An electrospray ion trap spectrometer (LCQ DecaXP, ThermoFinnigan, San Jose, CA, USA) coupled on line with HPLC was used for peptides analysis. Mass spectrometry were acquired in a mode that alternated a full MS scan (mass range: 400–1600) and a collision induced dissociation tandem mass spectrometry (MS/MS) of the most abundant ion. Data were analysed using the sequest algorithm incorporated with the ThermoFinnigan BioWorks software.

Results and Discussion

Tiling microarray-based identification of *E. faecalis* sRNAs

Tiling microarray has become a comprehensive approach to sRNA discovery. Identification of sRNA candidates transcribed by *E. faecalis* V583 was undertaken with two samples of cells harvested in mid-log growth phase and stationary phase after 24 h of incubation at 37°C in M17 glucose media. Analysis of IGRs tiling microarray data revealed 53 regions with intensity values of hybridization five fold higher than signals from apparent untranslated regions. Importantly, only one (see below) of these putative sRNAs identified by microarray was also predicted by bioinformatic approach as performed by Livny *et al.* [14]. This low overlap between microarray and *in silico* analysis is consistent with that observed in other bacteria [18]. These data show that computational and experimental methods are two complementary ways to identify sRNAs. As carried out for identification of sRNAs from *S. pneumoniae* using tiling arrays, we choose a stringent intensity cutoff to avoid false positives for identifying short length RNA [19]. Using a threshold of intensity of ten fold the background level led to the identification of 12 putative sRNAs (Table 1). No experimental evidence (neither sequence from RACE-PCR nor signal on Northern blot) was obtained for one of them (EF0940_EF0941). Since the IGR between EF0940 and EF0941 is only 51 bp in length, the corresponding probe putatively hybridized with the transcription product of EF0941. Thus, the candidate has been excluded from our study. The 11 other candidates that hybridized in specific intergenic regions were selected for further detailed characterization.

Experimental validation of 11 sRNAs in *E. faecalis*

One of the main goals of this study was to determine the sequence and the expression pattern of the 11 selected sRNA candidates. First, using a new RNA preparation, we performed Northern blot analysis to confirm the transcription of these RNAs during exponential growth phase and stationary phase and to determine the approximate size of each candidate. We observed a transcript for 10 out of the 11 candidates tested. Six of them (EF3314_EF3315, EF0820_EF0821, EFA0080_EFA0081, EF1368_EF1369, EF0408_EF0409 and EF0605_EF0606) were specifically expressed during exponential phase (Figure 1A–F); 1 sRNA (EF0869_EF0870) was specifically expressed after 24 h of starvation (Figure 1H); and 3 (EF1097_EF1098, EF-B0062_EFB0063 and EF2205_EF2206) were detected in comparable amounts in both phases (Figure 1G, J, K). These expression patterns were in good agreement with the results of tiling microarray except for EF1097_EF1098 which was much more expressed in stationary phase than under growing conditions on our chips. For unexplained reasons, no signal has been detected for EF0136_EF0137 (Figure 1I) by Northern blot analysis under our experimental conditions.

In order to determine the exact sequence of each sRNA candidate we identified the transcriptional start sites by 5′-RACE except for EFA0080_EFA0081 for which no result was gained. The 3′ ends of the transcripts were obtained either by 3′-RACE (Figure 1B, D, E, F, G, H, K, Table 1) or by combining transcript length data deduced from the Northern blots and computational prediction of transcriptional terminators [26] (Figure 1A, C, J, Table 1). Since neither putative terminator nor experimental data of the 3′ end of EF0136_EF0137 (Figure 1I) were obtained, the end of the sequence mentioned corresponds to the 3′ end of the tiling array probe. 5′-3′ RACE data of EF0820_EF0821 did not correlate to Northern blot results. From RACE-PCR, a 370 nt long sRNA was deduced that is larger than the predicted size (app. 100 nt) from Northern blot (using probe hybridizing on the 5′ region), suggesting that the large EF0820_EF0822 transcript was processed to short sRNA by modification of its 3′ end. Except for EF0820_EF0822, where the 99 last nucleotides correspond to the beginning sequence of EF0820, we could not identify obvious coding sequences (CdS), i.e. ORFs (open reading frames) with start codons connected to putative ribosome-binding sites in reasonable distances (around 8 nucleotides) inside the other sRNA candidates. Nevertheless, definitive exclusion of the presence of CdS in these regions needs experimental verification.

Table 1. sRNAs in *E. faecalis* V583 detected by tiling microarray.

	Intergenic Region	Left gene	sncRNA strand	Right gene			Size	Flanking genes		Expression value[a]		Expression ration (Expo/Stat)
					start	stop	(nt)			Expo	Stat	
sRNAs expressed at exponential phase												
A.	EF3314_EF3315	←	←	←	3201675	3201582[b]	94	EF3314:cell wall surface anchor family protein		65025.9	1249.6	52
						3201535[b]	141	EF3315:triphosphoribosyl-dephospho-CoA synthase				
B.	EF0820_EF0822	←	←	→	784383	784014	370	EF0820:rplY; 50S ribosomal protein L25/general stress protein Ctc		37086.5	1376.8	26.9
								EF0822:HAD (haloacid dehalogenase) superfamily hydrolase				
C.								EFA0080_EFA0081	→	←	→	63478[c]
	63423[b]	99						EFA0080:UvrC family transcriptional regulator	RNAI	37537.9	3062.9	12.3
								EF0081:hypothetical protein				
D.	EF1368_EF1369	←	→	←	1345556	1346183	628	EF1368:hypothetical protein		35465.0	3058.9	11.6
								EF1369:Cro/CI family transcriptional regulator				
								EF1370:drug resistance transporter, EmrB/QacA family protein				
E.	EF0408_EF0409	→	→	←	381297	381708	412	EF0408:PTS (phosphotransferase system) system, IIA component	RNAI	47418.0	11648.3	4.1
								EF0409:hypothetical protein				
F.	EF0605_EF0606	←	→	←	569151	569329	179	EF0605:hypothetical protein		41977.3	11288.0	3.7
								EF0606:Dps (DNA-binding protein from starved cells) family protein				
sRNAs expressed at stationary phase												
G.	EF1097_EF1098*	→	←	←	1067257	1066894	364	EF1097:hypothetical protein	tmRNA	3390.8	63399.5	0.05
								EF1098:hypothetical protein				
H.	EF0869_EF0871	←	←	→	829525	829052	474	EF0869:Cro/CI family transcriptional regulator		2655.4	47286.9	0.06
								EF0871:cation transpoter E1–E2 family ATPase				
I.	EF0136_EF0137	→	←	→	137278	137066[d]	>213	EF0136:hypothetical protein		1755.7	28560.7	0.06
								EF0137:nucleotidyl transferase domain-containing protein				
sRNAs expressed at exponential and stationary phase												
J.	EFB0062_EFB0063	→	←	→	55834	55623[b]	212	EFB0062:UvrC family transcriptional regulator	RNAI	49218.4	52343.1	0.94
								EFB0063:replication control protein PrgN				
K.	EF2205_EF2206	→	←	←	2119382	2119296	87	EF2205:hypothetical protein	4.5S	41604.0	55672.1	0.75
								EF2206:cytidine/deoxycytidylate deaminase family protein				

[a]: Intensity of hybridization from the intergenic probe showing the highest signal in exponential or stationary phase.
[b]: Computer prediction of the putative 3′ end (using TransTerm software).
[c]: 5′ end corresponding to the 5′ end of probe.
[d]: 3′end corresponding to the 3′ end of probe.
doi:10.1371/journal.pone.0023948.t001

Altogether, the length range of the identified sRNAs was 87–628 nucleotides and the deduced sequences and promoter regions of the 11 sRNAs are shown in Figure 1. In comparison with sRNAs identified by Fouquier d'Hérouel *et al.* [23] using *in silico* prediction and "5′tag-RACE" strategy, only four overlap with our sRNA candidates (EF0605_EF0606, EF1097_EF1098, EF0869_EF0871, and EF2205_EF2206 corresponding to *ref*25C, *ssr*A, *ref*19C, and *ffs*, respectively). This shows that several techniques as well as different growth conditions (see below) are necessary for more exhaustive identification of sRNAs.

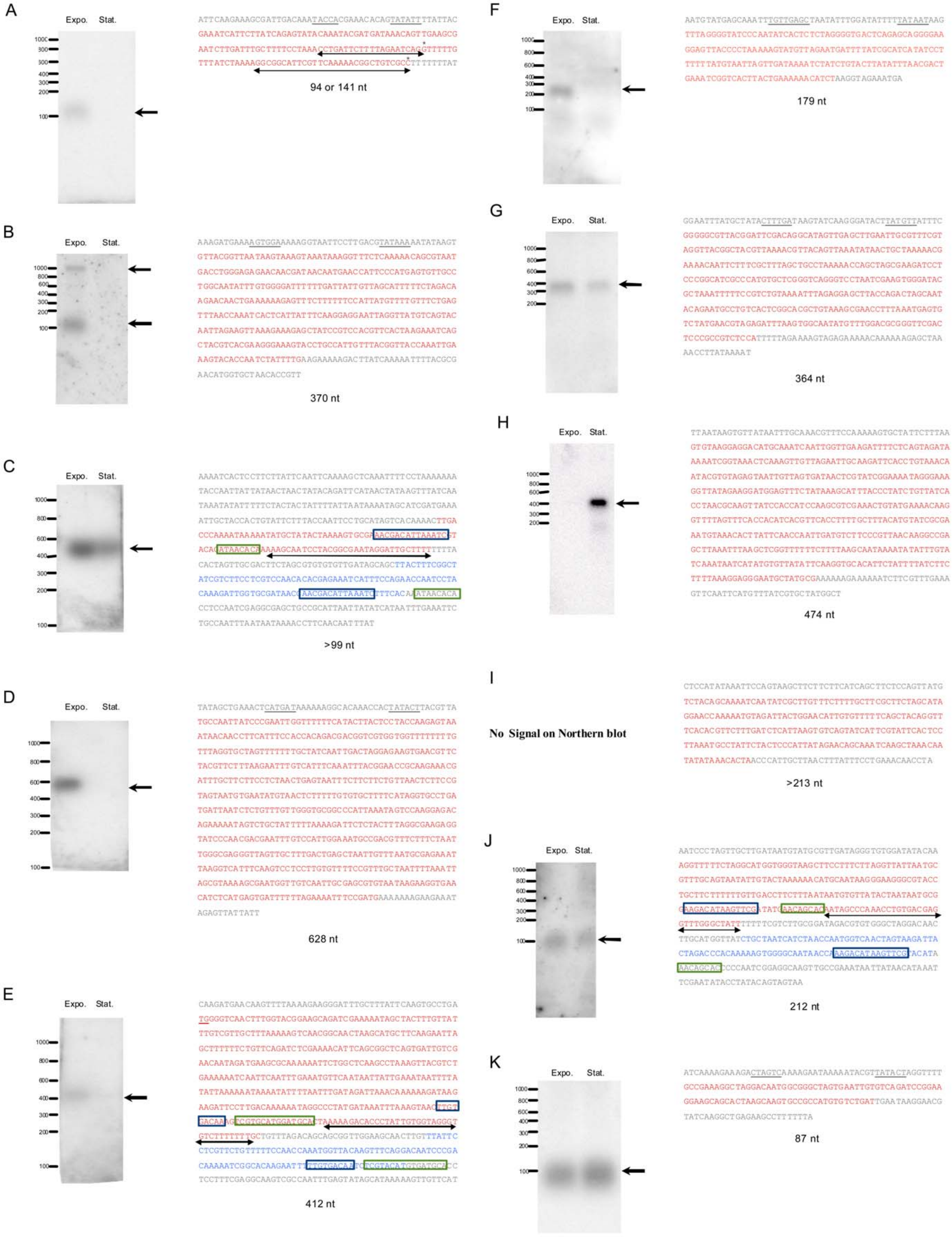

A
Expo. Stat.
94 or 141 nt
F
Expo. Stat.
179 nt
B
Expo. Stat.
370 nt
G
Expo. Stat.
364 nt
C
Expo. Stat.
>99 nt
H
Expo. Stat.
474 nt
I
No Signal on Northern blot
>213 nt
D
Expo. Stat.
628 nt
J
Expo. Stat.
212 nt
E
Expo. Stat.
412 nt
K
Expo. Stat.
87 nt
1000
800
600
400
300
200
100

Figure 1. Northern blots and sequences of sRNAs (A: EF3314_EF335, B: EF0820_EF0821, C: EFA0080_EFA0081, D: EF1368_EF1369, E: EF0408_EF0409, F: EF0605_EF0606, G: EF1097_EF1098, H: EF0869_EF0870, I: EF0136_EF0137, J: EFB0062_EFB0063 and K: EF2205_EF2206). RNA was isolated from cells at exponential (Expo) and stationary (Stat) phases. Northern blot analyses were performed using α^{32}P-labelled probes. Arrows on Northern blot picture indicate the sRNAs corresponding bands. The transcriptional start sites and terminators of sRNAs were determined by 5′ RACE and 3′ RACE or by *in silico* analysis using TransTerm software. The putative −10 and/or −35 promoter sequences are underlined, and the sRNA sequence is written in red letters. Putative 3′-ends of EF3314_EF335 sRNA (panel A) is indicated by stars (*). The 3′-end of the sequenceof EF0136_EF0137 (panel I) mentioned here corresponds to the 3′-end of the tiling array probe. Black arrows in the sequence indicate the predicted terminators. The *fst* gene is written in blue letters and direct repeats "a" and "b" (DRa and DRb) of *par* system are blue and green boxed, respectively (panels C, E, and J).
doi:10.1371/journal.pone.0023948.g001

Features of sRNAs

As previously mentioned, an antisense RNA regulated addiction module named "*par*" system was described on the *E. faecalis* plasmid pAD1 [9]. The components of this toxin-antitoxin (TA) system are antisense RNA (RNA II) and its target, RNA I encoding the peptide toxin Fst. Such systems play a crucial role in plasmid stability by killing any daughter cells that fail to inherit a copy of the plasmid. Three putative sRNAs identified in our study (EFA0080_EFA0081 in pTEF1, EFB0062_EFB0063 in pTEF2, and EF0408_EF0409 in the chromosome) corresponded to the RNAI components of the TA systems already identified in *E. faecalis* V583 by Weaver and coworkers [11]. As shown in Figure 1 (C, E, J), RNA I (including *fst* toxin gene) and RNA II homologues had two direct repeat sequences and shared the same bidirectional terminator. One interesting question concerns the role of *par* addiction module located on the bacterial chromosome. Several studies revealed various roles such as in mobile element stability or stress response [12,33,34]. As pointed out, in the case of par_{EF0409} (including EF0408_EF0409 sRNA), its association with genes encoding phosphotransferase components homologous to a mannitol transport system suggests a potential function in nutritional uptake [11].

Northern blot and tiling microarray showed that EF1097_EF1098 was expressed in both growth and stationary phases and we were able to determine the exact sequence of this sRNA (Figure 1G). EF1097_EF1098 corresponds to *E. faecali*s tmRNA (*ssrA*) that is a unique bi-functional RNA acting as both a tRNA and an mRNA. It functions as the rescue system of ribosomes stalled on aberrant mRNAs and adds a peptide tag to nascent polypeptides for directed proteolysis (named *trans*-translation) [35,36]. tmRNA is universally conserved and is one of the most abundant RNA in the cells [37]. It has not only an important role in mRNA turnover but also likely in monitoring protein folding (for review see [35]). Mutations that inactivate tmRNA are lethal for some species (ie, *Neisseria gonnorhoeae*, *Haemophilus influenzae*, *Shigella flexneri*) or, for others, affect bacterial physiology such as virulence (ie, *Salmonella enterica*, *Yersinia pseudotuberculosis*) or stress response (ie, *E. coli*, *B. subtilis*) [35,37]. Determination of the impact of tmRNA deletion in *E. faecalis* is under investigation in our laboratory.

We used the Rfam database (a collection of non-coding RNA families) to determine the putative functions of characterized sRNAs [27]. We found that EF2205_EF2206 sRNA matched with the Signal Recognition Particle (SRP) functional category. SRP is a ribonucleoprotein complex that targets proteins for secretion through co-translational process and is composed of protein Ffh and 4.5S RNA in prokaryotes. Our analysis revealed that EF_1700 gene (*ffh*) product and EF2205_EF2206 correspond to the two components of the SRP in *E. faecalis*. Interestingly, a recent study demonstrated that mutation of the gene encoding 4.5S RNA in *S. pyogenes* (phylogenetically related to *E. faecalis*) results in reduction of virulence [38].

In order to predict target genes of the other sRNAs identified in this study, we performed *in silico* analysis (Table 2, Table S2). Two different softwares were used for a more precise identification. sRNATarget server is based on the Naive Bayes probabilistic method and take RNA secondary structure profile as the feature [28]. The second, IntaRNA, predicts interactions between two RNA molecules, and the scoring is based on hybridization free energy and accessibility of the interaction sites in both molecules [29]. Numerous putative target genes were obtained by combination of these two approaches (from 9 for EF3314_EF3315 to 81 for EF0136_EF0137) (Table 2, Table S2). *In silico* prediction (Table S2) as well as sequence analysis suggested antisense activity for EF1368_EF1369 and EF0136_EF0137. Indeed, EF1369 mRNA sequence, encoding a putative transcriptional regulator, was fully complementary to EF1368_EF1369 sRNA. Likewise, the first 136 nucleotides of EF0136_EF0137 were complementary with the beginning sequence of EF0137 mRNA. The combined *in silico* data constitute hypothetical regulons for the sRNA candidates that need to be experimentally verified.

In general, sRNAs act at the post transcriptional level of regulation [1,3]. Then, in order to observe a putative influence of one sRNA in *E. faecalis*, proteomic approach was undertaken comparing profiles of the ΔEF3314_EF3315 mutant and the parental strain. Two-dimensional gel electrophoresis of proteins from growing *E. faecalis* V19 and ΔEF3314_EF3315 mutant strains are shown in Figure 2. From two distinct experiments we observed that intensity of 4 spots were reproducibly different between the two strains. Numbers 1, 2, and 4 were only present in the mutant whereas number 3 was only seen in the wild type (Figure 2). By mass spectrometry, after extraction of proteins from the gel, we identified these polypeptides. Spots 1, 2, 3, and 4 correspond to DnaK (EF_1308, 63 kDa), ribosomal protein S1 (EF_1548, 43 kDa), ribosomal protein L6 (EF_0221, 19 kDa), and translation elongation factor Tu (EF_0221, 43 kDa), respectively. However, molecular weight (MW) deduced from the gels (around 45 kDa, 30 kDa, 15 kDa, for peptides 1, 2, and 4, respectively) did not correlated with the expected sizes. Therefore, peptides indentified

Table 2. Number of putative target genes.

	Number of mRNA candidate		
sRNA	**sRNATarget (score>0.9)[a]**	**IntaRNA[b]**	**common[c]**
EF3314_EF3315	75	31[d]	9
EF0820_EF0822	176	213[d]	44
EF1368_EF1369	876	97[e]	72
EF0605_EF0606	210	85[d]	24
EF0869_EF0871	494	318[d]	62
EF0136_EF0137	1252	92[e]	81

[a]: http://ccb.bmi.ac.cn/sRNAtarget [28].
[b]: http://rna.informatik.uni-freiburg.de:8080/IntaRNA.jsp [29].
[c]: list of genes is in Table S2.
[d]: cut-off <−10 kcal/mol.
[e]: cut-off <−15 kcal/mol.
doi:10.1371/journal.pone.0023948.t002

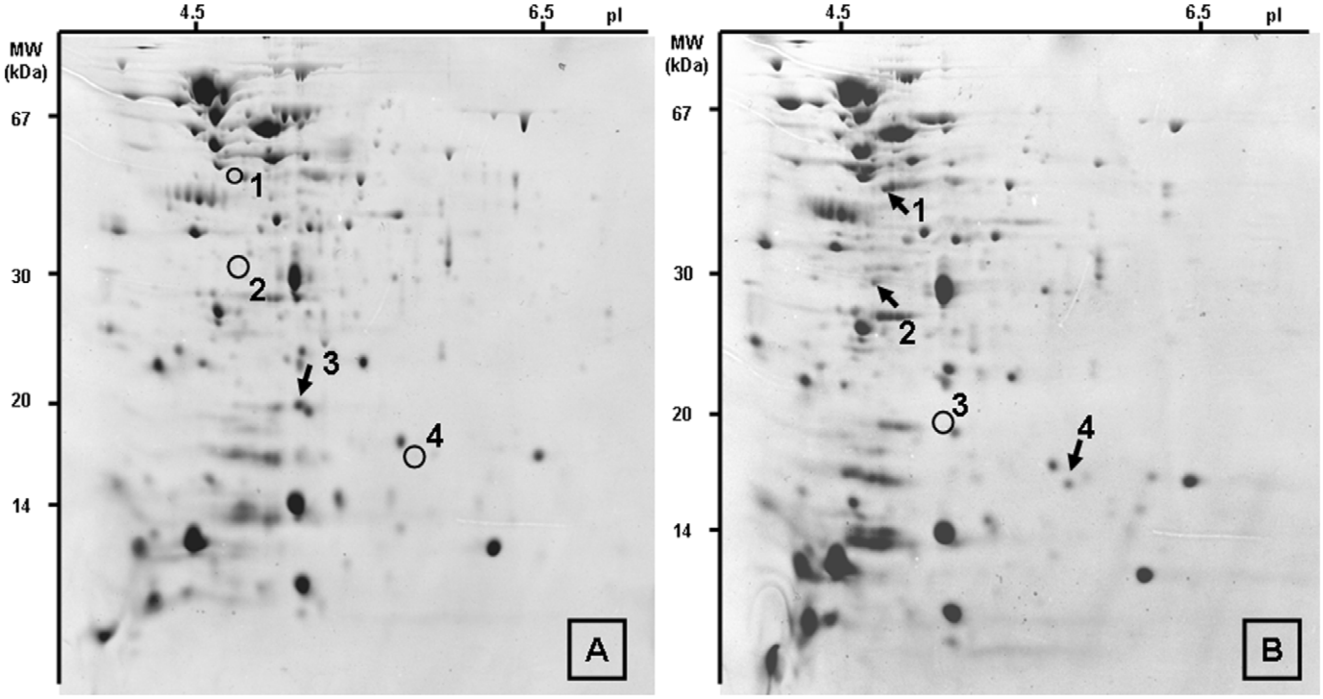


Figure 2. Two-dimensional gel electrophoresis of proteins from *E. faecalis* V19 (A) and ΔEF3314_EF335 mutant (B). Arrows indicate polypeptides that are detected in one gel but not in the other. The position of the polypeptides absent in a given gel are indicated by circles.
doi:10.1371/journal.pone.0023948.g002

from the mutant samples likely corresponded to protein degradation products. On the other hand, MW of spot number 3, which is absent in the mutant, was estimated at around 20 kDa in good accordance with the calculated size of the intact protein (19 kDa). These combined results suggested that EF3314_EF3315 might be involved in the turnover of some abundant proteins in *E. faecalis*, especially from the translational apparatus.

Expression of sRNAs in different stress conditions

Generally, the expression of sRNAs are tightly regulated and induced by specific environmental condition [2]. We then performed tiling arrays with new RNA samples in order to analyze the transcription of sRNAs previously characterized under 11 different conditions of growth some of which may correspond to stresses encountered during intestinal colonization or during the infectious process (see Material and Methods). Expression patterns of the 11 sRNAs under H_2O_2, BS, and acid stress conditions, during growth in presence or absence of O_2 and in serum and urine is presented in Table 3. EF0408_EF0409, EFA0080_EFA0081 and EFB0062_EFB0063, identified as members of TA systems were highly expressed at different stages of growth with oxygen (Table 3). Physiological significance of the induction of transcription of these

Table 3. Expression patterns of sRNAs under different growth phases and stress conditions.

sRNAs	Stress conditions										
	H_2O_2	pH (acid)	BS	Expo	Early Stat	Stat	Expo	Early Stat	Stat	Urine	Serum
				with O_2	with O_2	with O_2					
EF3314_EF3315	127	2263	596	787	1756	64	3478	659	93	191	112
EF0820_EF0822	114	150	135	370	253	47	473	605	51	59	122
EFA0080_EFA0081	102	186	314	857	43619	2364	5034	1566	3886	1756	2178
EF1368_EF1369	761	2817	874	1178	418	118	171	614	168	144	139
EF0408_EF0409	1756	2916	649	20636	1916	283	1597	1909	136	954	257
EF0605_EF0606	722	2056	246	1880	3246	214	261	113	326	802	129
EF1097_EF1098	4535	32765	106115	1835	22301	2438	1518	1492	3977	41662	11483
EF0869_EF0871	556	159	1236	196	7780	13465	374	119683	31710	5974	30293
EF0136_EF0137	59	108	11	70		27	46	11	125	144	101
EFB0062_EFB0063	125	332	194	2817	802	20636	1756	1236	5503	211	179
EF2205_EF2206	10724	21313	11296	9823	25155	10468	202452	11483	405266	29510	221227

doi:10.1371/journal.pone.0023948.t003

Table 4. Distribution of the 11 sRNAs among *E. faecalis* strains.

	sRNAs										
E. faecalis	**EF3314_**	**EF0820_**	**EFA0080_**	**EF1368_**	**EF0408_**	**EF0605_**	**EF1097_**	**EF0869_**	**EF0136_**	**EFB0062_**	**EF2205_**
strains	**EF3315**	**EF0822**	**EFA0081**	**EF1369**	**EF0409**	**EF0606**	**EF1098**	**EF0871**	**EF0137**	**EFB0063**	**EF2206**
OGR1RF	90	100		90	90		90	90			100
ARO1/DG	90	100		90	90		100	90		80–90 P	100
ATCC 29200	100	100	80–90 G	90	90		100	90		80–90 G	100
ATCC 4200	100	100		90	90		100	90			100
CH188	100	100		90	90		100	90	80–90		100
D6	100	100		90	90	100	90	90			100
DAPTP0512	90	100	80–90 G	90	90		90	90		80–90 G	90
DAPTP0516	90	100	80–90 G	90	90		90	90		80–90 G	90
DS5	100	100	100 P	90	90		90	90	80–90	80–90 P	100
E1Sol	100	90		90	90		100	90		80–90 P	100
Fly1	90	90		90	90		90	90			100
HH22	100	100	100 G	90	100		100	100		80–90 G	100
HIP11704	100	90	80–90 G	90	90		90	90		80–90 G	100
JH1	100	100	80–90 P	90	90		90	90	80–90	80–90 P	100
Merz96	90	100	80–90 P	90	90		90	90		80–90 P	90
PC1.1	100	100		90	90		90	90	80–90		100
R712	90	100	80–90 G	90	90		90	90		80–90 G	90
S613	90	100	80–90 G	90	90		90	90		80–90 G	90
T1	100	100		90	90		90	90		90 P	100
T11	100	100		90	100		100	100			100
T2	100	100	80–90 P	90	90	100	90	90	100	80–90 P	100
T3	100	100		90	90		90	90		80–90 P	100
T8	100	100	90 G	90	90		90	90		80–90 G	100
TUSoD Ef11	100	90		80–90	90		90	90			90
TX0012	100	100		90	90		90	90			90
TX0017	100	100	80–90 G	90	90		90	90		80–90 G	100
TX0027	100	100	80–90 G	90	90		90	90	80–90	80–90 G	100
TX0031	100	100		90	90		100	90			100
TX0043	100	100		90	90		100	90			90
TX0102	100	100		90	90		100	90			100
TX0104	100	100	80–90 G	90	90	90	90	90		80–90 G	100
TX0109	100	100		90	90		90	90		80–90 G	90
TX0309A	100	100		90	100	100	100	100	80–90	90 G	100
TX0309B	100	100		90	100	100	100	100	80–90	90 G	100
TX0312	100	100		90	90		100	90			90
TX0411	100	100	80–90 G	90	90		100	90		90 G	90
TX0470	100	100		80–90	90		90	100	80–90		100
TX0630	100	100	80–90 G	90	90		100	90	100	90 G	100
TX0635	100	100	90 G	90	90		100	90	80–90	80–90 G	100
TX0645	100	90		80–90	90		90	90	80–90	80–90 G	100
TX0855	100	90	80–90 G	90	90	100	90	90		90 G	100
TX0860	100	100		90	90	100	90	90		90 G	100
TX1302	100	100		90	90		90	90			100
TX1322	100	100		90	90		100	90		80–90 G	100
TX1341	100	100		90	90	100	90	100	80–90	80–90 G	100
TX1342	100	100		90	90		90	90			100
TX1346	100	90		90	90		90	90			90

Table 4. Cont.

	sRNAs										
E. faecalis	**EF3314_**	**EF0820_**	**EFA0080_**	**EF1368_**	**EF0408_**	**EF0605_**	**EF1097_**	**EF0869_**	**EF0136_**	**EFB0062_**	**EF2205_**
strains	**EF3315**	**EF0822**	**EFA0081**	**EF1369**	**EF0409**	**EF0606**	**EF1098**	**EF0871**	**EF0137**	**EFB0063**	**EF2206**
TX2134	100	100	80–90 G	90	90		100	90		90 G	100
TX2137	100	100	80–90 G	90	90	100	90	90		80–90 G	100
TX2141	100	90		80–90	90		90	90			90
TX4000	100	100		90	90		90	90			100
TX4244	100	100		90	90		90	90	80–90	80–90 G	100
TX4248	100	100	80–90 G	90	90		90	90	80–90	80–90 G	100
X98	100	100	80–90 P	90	90		100	90		80–90 P	90

100 indicates 100% identity.
90 indicates more than >90% identity.
80–90 indicates between 80 and 90% identity.
White box indicates the absence of homology.
G: on genome.
P: on plasmid.
doi:10.1371/journal.pone.0023948.t004

three homologues especially in presence of oxygen remains unclear. However, the expressions of these paralogues appeared sequential during growth phases. EF0408_EF0409 was mainly transcribed during exponential phase, EFA0080_EFA0081 during early stationary phase, and EFB0062_EFB0063 after 24 h of stationary phase (Table 3). These observations suggest that the different TA systems may have different roles according to the growth phase of the bacteria.

E. faecalis tmRNA (EF1097_EF1098) and 4.5S RNA (EF2205_EF2206) showed a high intensity of hybridization under all conditions tested but BS and late stationary phase induced the highest level of tmRNA and 4.5S RNA expression, respectively. Furthermore, EF0869_EF0871 was highly expressed in urine and serum medium (Table 3). It has been shown that transcription of some genes encoding fitness and virulence factors are affected when *E. faecalis* is incubated in these biological media [39,40]. It is then tempting to speculate that these sRNAs could play a crucial role in the cellular response triggered during the infectious process.

Surprisingly, for unexpected reason, signals corresponding to the two sRNAs EF0136_EF0137 and EF0820_EF0822 were very low in these tiling arrays experiments leading to unexploitable data. On the other hand, EF3314_EF3315, EF1368_EF1369 and EF0605_EF0606 sRNAs appeared moderately expressed but were obviously induced by acid stress (Table 3). However, exponential growth phase and early stationary phase in presence of oxygen were the most favorable conditions for EF3314_EF3315 and EF0605_EF0606 expressions, respectively (Table 3). This is in agreement with the induction of Ref25C (corresponding to EF0605_EF0606) in oxidative stress condition reported by Fouquier d'Hérouel *et al.* [23].

Our tiling arrays data using RNA samples obtained from cells incubated under 11 different growth conditions allowed us to identify 76 new IGRs with intensities of hybridization ten fold higher than signals from apparent untranslated regions. *Probe sequences and tiling array data obtained with samples from stressed cells are shown in Table S3. A more detailed analysis of these new candidates is in progress in our laboratory*. In addition, if the threshold was set to five-fold induction, 174 putative sRNAs were detected in our experiments. sRNAs are usually transcribed under specific growth conditions and it is likely that some could be expressed under stressing conditions not yet tested. Moreover, sRNAs may have been missed in our study due to experimental procedure since our chips only covered intergenic regions of the V583 genome and since fractionated RNAs have been used for the hybridizations. It has been generally predicted that genome sizes ranging from 3–4 Mbp may contain 80–300 sRNAs [14]. Taken together it is highly probable that the number of sRNA transcripts detected in *E. faecalis* will greatly increase in the near future.

Distribution of sRNAs among *E. faecalis* strains

To date, the whole genome sequence of 54 *E. faecalis* strains are available in the NCBI database. We performed standard BLAST analysis to detect the presence of the characterized sRNAs in these different *E. faecalis* strains (Table 4). Seven of them are highly conserved (90 to 100% identical) and present in all *E. faecalis* genomes (EF3314_EF3315, EF0820_EF0821, EF1368_EF1369, EF0408_EF0409, EF1097_EF1098, EF0869_EF0871 and EF2205_EF2206). The other four are not systematically observed because of their location on a mobile genetic element (EF0136_EF0137), in the pathogenicity island (PAI) (EF0605_EF0606) or on plasmids (EFA0080_EF0081 and EFB0062_EFB0063) [41]. sRNAs EF0605_EF0606, EF0136_EF0137, EFA0080_EFA0081 and EFB0062_EFB0063 homologues (at least 80% identical) are present in 9, 15, 35 and 23 strains of the 54 genomes analyzed, respectively (Table 4).

Homologues of EF0408_EF0409 (more than 90% identity) (member of TA system, see above) were systematically present in all *E. faecalis* genomes. Moreover, additional plasmidic EFA0080_EFA0081 and EFB0062_EFB0063 homologous were also observed in some chromosomes showing that most *E. faecalis* strains have several *par* systems arguing for a selective advantage for the bacterial cell.

Interestingly, EF0605_EF0606 is located in PAI between a gene encoding a Dps family protein (EF_0606) and an operon including a paralogue of *gls24* (EF_0605-EF_0604). Dps is a protein involved in the protection of DNA against oxidative stress and Gls24 corresponds to a general stress protein that is a virulence factor in *E. faecalis* [42,43,44]. In *S. pneumoniae*, two sRNAs had demonstrated *cis*-acting effects on the transcription of adjacent genes [45]. From these observations and the fact that EF0605_EF0606 sRNA is induced under aerobic growth conditions, it may be hypothesized that it has a role in the control of expression of these enzymes and hence may be implicated in stress response and virulence of *E. faecalis*.

Perspectives

In this work we have determined the sequences, locations and expression patterns of 11 sRNAs in *E. faecalis* V583. These results provide a starting point towards understanding of the complex RNA regulatory network governing *E. faecalis* physiology and virulence. Recently, comparative genome-wide analysis of putative or characterized sRNAs of five major Gram-positive pathogens (*L. monocytogenes* EGD-e, *Clostridium difficile* 630, *Staphylococcus aureus* COL, *S. pyrogenes* M1 GAS, and *E. faecalis* V583) was reported [46]. This information will help to understand the molecular mechanisms of the pathogenic process which might be useful for the development of novel microbial diagnosis tools and anti-bacterial drugs such as antisense PNAs (peptide nucleic acids) [46].

Supporting Information

Table S1 Primers and probes used in this study.
(DOC)

Table S2 List of putative target genes of EF3314_EF3315, EF0820_EF0822, EF1368_EF1369, EF0605_EF0606, EF0869_EF0871, EF0136_EF0137 sRNA candidates.
(XLS)

Table S3 Probe sequences and tiling array data obtained with samples from stressed cells of *E. faecalis*.
(XLS)

Acknowledgments

The expert technical assistance of Isabelle Rincé, Marie-Jeanne Pigny and Evelyne Marchand was greatly appreciated. We would like to thank all members of our labs for helpful discussions.

Author Contributions

Conceived and designed the experiments: KS JCG TH AH. Performed the experiments: KS CM CK AB-V. Analyzed the data: KS CM NV JCG. Contributed reagents/materials/analysis tools: CK TH NV. Wrote the paper: KS AH JCG.

References

1. Waters LS, Storz G (2009) Regulatory RNAs in bacteria. Cell 20: 615–628.
2. Repoila F, Darfeuille F (2009) Small regulatory non-coding RNAs in bacteria: physiology and mechanistic aspects. Biol Cell 101: 117–131.
3. Storz G, Opdyke JA, Zhang A (2004) Controlling mRNA stability and translation with small, noncoding RNAs. Curr Opin Microbiol 7: 140–144.
4. Massé E, Escorcia FE, Gottesman S (2003) Coupled degradation of a small regulatory RNA and its mRNA targets in *Escherichia coli*. Genes Dev 17: 2374–2383.
5. Nielsen JS, Lei LK, Ebersbach T, Olsen AS, Klitgaard JK, et al. (2010) Defining a role for Hfq in Gram-positive bacteria: evidence for Hfq-dependent antisense regulation in *Listeria monocytogenes*. Nucleic Acids Res 38: 907–919.
6. Sun X, Zhulin I, Wartell RM (2002) Predicted structure and phyletic distribution of the RNA-binding protein Hfq. Nucleic Acids Res 30: 3662–3671.
7. Valentin-Hansen P, Eriksen M, Udesen C (2004) The bacterial Sm-like protein Hfq: a key player in RNA transactions. Mol Microbiol 51: 1525–1533.
8. Brantl S (2007) Regulatory mechanisms employed by *cis*-encoded antisense RNAs. Curr Opin Microbiol 10: 102–109.
9. Weaver KE (2007) Emerging plasmid-encoded antisense RNA regulated systems. Curr Opin Microbiol 10: 110–116.
10. Gerdes K, Thisted T, Martinussen J (1990) Mechanism of post-segregational killing by the *hok/sok* system of plasmid R1: sok antisense RNA regulates formation of a *hok* mRNA species correlated with killing of plasmid-free cells. Mol Microbiol 4: 1807–1818.
11. Weaver KE, Reddy SG, Brinkman CL, Patel S, Bayles KW, et al. (2009) Identification and characterization of a family of toxin-antitoxin systems related to the *Enterococcus faecalis* plasmid pAD1 *par* addiction module. Microbiology 155: 2930–2940.
12. Gerdes K, Wagner EG (2007) RNA antitoxins. Curr Opin Microbiol 10: 117–124.
13. Livny J, Waldor MK (2007) Identification of small RNAs in diverse bacterial species. Curr Opin Microbiol 10: 96–101.
14. Livny J, Teonadi H, Livny M, Waldor MK (2008) High-throughput, kingdom-wide prediction and annotation of bacterial non-coding RNAs. PLoS One 3: e3197.
15. Sharma CM, Vogel J (2009) Experimental approaches for the discovery and characterization of regulatory small RNA. Curr Opin Microbiol 12: 536–546.
16. Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeiss S, et al. (2010) The primary transcriptome of the major human pathogen *Helicobacter pylori*. Nature 464: 250–255.
17. Landt SG, Abeliuk E, McGrath PT, Lesley JA, et al. (2008) Small non-coding RNAs in *Caulobacter crescentus*. Mol Microbiol 68: 600–614.
18. Perez N, Treviño J, Liu Z, Ho SC, Babitzke P, et al. (2009) A genome-wide analysis of small regulatory RNAs in the human pathogen group A *Streptococcus*. PLoS One 4: e7668.
19. Kumar R, Shah P, Swiatlo E, Burgess SC, Lawrence ML, et al. (2010) Identification of novel non-coding small RNAs from *Streptococcus pneumoniae* TIGR4 using high-resolution genome tiling arrays. BMC Genomics 11: 350.
20. Toledo-Arana A, Dussurget O, Nikitas G, Sesto N, Guet-Revillet H, et al. (2009) The *Listeria* transcriptional landscape from saprophytism to virulence. Nature 459: 950–956.
21. Ogier JC, Serror P (2008) Safety assessment of dairy microorganisms: *the Enterococcus genus*. Int J Food Microbiol 126: 291–301.
22. Paulsen IT, Banerjei L, Myers GS, Nelson KE, Seshadri R, et al. (2003) Role of mobile DNA in the evolution of vancomycin-resistant *Enterococcus faecalis*. Science 299: 2071–2074.
23. Fouquier d'Hérouel A, Wessner F, Halpern D, Ly-Vu J, Kennedy SP, et al. (2011) A simple and efficient method to search for selected primary transcripts: non-coding and antisense RNAs in the human pathogen *Enterococcus faecalis*. Nucleic Acids Res;[Epub ahead of print].
24. Sahm DF, Kissinger J, Gilmore MS, Murray PR, Mulder R, et al. (1989) In vitro susceptibility studies of vancomycin-resistant *Enterococcus faecalis*. Antimicrob Agents Chemother 33: 1588–1591.
25. Sambrook J, Fritsch EF, Maniatis T (1989) Molecular cloning: a laboratory manual, 2nd ed. Cold Spring Harbor Laboratory Press Cold Spring Harbor, N.Y.
26. Kingsford CL, Ayanbule K, Salzberg SL (2007) Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. Genome Biol 8: R22.
27. Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL, et al. (2009) Rfam: updates to the RNA families database. Nucleic Acids Res 37: D136–140.
28. Cao Y, Zhao Y, Cha L, Ying X, Wang L, et al. (2009) sRNATarget: a web server for prediction of bacterial sRNA targets. Bioinformation 3: 364–366.
29. Smith C, Heyne S, Richter AS, Will S, Backofen R (2010) Freiburg RNA Tools: a web server integrating IntaRNA, ExpaRNA and LocARNA. Nucleic Acids Res 38 Suppl: 373–377.
30. Arnaud M, Chastanet A, Débarbouillé M (2004) A new vector for efficient allelic replacement in naturally non transformable low GC% Gram-positive bacteria. Appl Environ Microbiol 70: 6887–6891.
31. Giard J-C, Laplace J-M, Rincé A, Pichereau V, Benachour A, et al. (2001) The Stress Proteome of *Enterococcus faecalis*. Electrophoresis 14: 2947–2954.
32. Budin-Verneuil A, Pichereau V, Auffray Y, Ehrlich DS, Maguin E (2005) Proteomic characterization of the acid tolerance response in *Lactococcus lactis* MG1363. Proteomics 5: 4794–4807.
33. Engelberg-Kulka H, Amitai S, Kolodkin-Gal I, Hazan R (2006) Bacterial programmed cell death and multicellular behavior in bacteria. PLoS Genet 2: e135.
34. Fozo EM, Hemm MR, Storz G (2008) Small toxic proteins and the antisense RNAs that repress them. Microbiol Mol Biol Rev 72: 579–589.
35. Hayes CS, Keiler KC (2010) Beyond ribosome rescue: tmRNA and co-translational processes. FEBS Lett 584: 413–419.
36. Dulebohn D, Choy J, Sundermeier T, Okan N, Karzai AW (2007) *Trans*-translation: the tmRNA-mediated surveillance mechanism for ribosome rescue, directed protein degradation, and nonstop mRNA decay. Biochemistry 46: 4681–93.
37. Keiler KC (2007) Physiology of tmRNA: what gets tagged and why? Curr Opin Microbiol 10: 169–175.
38. Treviño J, Perez N, Sumby P (2010) The 4.5S RNA component of the signal recognition particle is required for group A *Streptococcus* virulence. Microbiology 156: 1342–1350.
39. Shepard BD, Gilmore MS (2002) Differential expression of virulence-related genes in *Enterococcus faecalis* in response to biological cues in serum and urine. Infect Immun 70: 4344–4352.
40. Vebø HC, Solheim M, Snipen L, Nes IF, Brede DA (2010) Comparative genomic analysis of pathogenic and probiotic *Enterococcus faecalis* isolates, and their transcriptional responses to growth in human urine. PLoS One 5: e12489.

41. McBride SM, Coburn PS, Baghdayan AS, Willems RJ, Grande MJ, et al. (2009) Genetic variation and evolution of the pathogenicity island of *Enterococcus faecalis*. J Bacteriol 191: 3392–3402.
42. Giard JC, Verneuil N, Auffray Y, Hartke A (2002) Characterization of genes homologous to the general stress-inducible gene *gls24* in *Enterococcus faecalis* and *Lactococcus lactis*. FEMS Microbiol Lett 206: 235–239.
43. Giard JC, Rince A, Capiaux H, Auffray Y, Hartke A (2000) Inactivation of the stress- and starvation-inducible *gls24* operon has a pleiotrophic effect on cell morphology, stress sensitivity, and gene expression in *Enterococcus faecalis*. J Bacteriol 182: 4512–4520.
44. Teng F, Nannini EC, Murray BE (2005) Importance of *gls24* in virulence and stress response of *Enterococcus faecalis* and use of the Gls24 protein as a possible immunotherapy target. J Infect Dis 191: 472–480.
45. Tsui HC, Mukherjee D, Ray VA, Sham LT, Feig AL, et al. (2010) Identification and characterization of noncoding small RNAs in *Streptococcus pneumoniae* serotype 2 strain D39. J Bacteriol 192: 264–279.
46. Mraheil MA, Billion A, Kuenne C, Pischimarov J, Kreikemeyer B, et al. (2010) Comparative genome-wide analysis of small non-coding RNAs of major Gram-positive pathogens: from fundamentals to applications. Microbial Biotechnology 3: 658–676.

Complete Sequences of Plasmids from the Hemolytic-Uremic Syndrome-Associated *Escherichia coli* Strain HUSEC41

Carsten Künne,[a] Andre Billion,[a] Stephen E. Mshana,[b] Judith Schmiedel,[a] Eugen Domann,[a] Hamid Hossain,[a] Torsten Hain,[a] Can Imirzalioglu,[a] and Trinad Chakraborty[a]

Institute of Medical Microbiology, Justus-Liebig University, Giessen, Germany,[a] and Department of Microbiology, Weill Bugando Medical, College of Health Sciences, Mwanza, Tanzania[b]

The complete and annotated sequences of four plasmids from a historical enteroaggregative Shiga toxin-producing *Escherichia coli* (HUSEC) serotype O104:H4 strain, HUSEC41/01-09591, isolated in 2001 in Germany are reported.

The *Escherichia coli* serotype O104:H4 sequence type (ST) 678 strain which caused a disease outbreak in Germany in 2011 harbors three plasmids encoding a putative autotransporter serine protease, the aggregative adherence regulator *aggR*, and the extended-spectrum beta-lactamase CTX-M-15, all of which have contributed to the evolution of this virulent strain (1, 3, 6). A subsequent study of plasmids from the historical enteroaggregative Shiga toxin-producing serotype O104:H4 strain HUSEC41, isolated in 2001 from a child presenting with hemolytic-uremic syndrome, reported the detection and partial sequence of two plasmids, 95 and 75 kb (5). Our analysis of HUSEC41, however, indicated the presence of four plasmids, with sizes of 92, 74, 8, and 5 kb. This observation prompted us to determine the complete sequences of all plasmids in strain HUSEC41. These sequences provide a backdrop for the comparative analysis of the genealogy and evolution of plasmids that have contributed to the virulence properties of the HUSEC strain responsible for the recent outbreak of 2011.

Genomic DNA was isolated as described by Pitcher et al. (7). Sequencing was performed by Vertis (Germany) on a 454 GS-FLX system. Reads were assembled *de novo* with the 454 Newbler assembler, and resulting contigs were mapped against reference plasmids to determine a plasmid context. PCR-based techniques were used to close the gaps, followed by sequencing with ABI BigDye 3.0 technology (Applied Biosystems, Germany). A total of 17 contigs were assembled in four circular replicons with an average coverage of 45×. ORF calling and a first-pass automatic annotation were performed using RAST (rast.nmpdr.org) followed by manual comparative curation (4) and sequence similarity searches versus the NCBI (www.ncbi.nlm.nih.gov/BLAST), PFAM, and IS Finder (www-is.biotoul.fr) databases.

Four plasmids were detected: pHUSEC41-1, a large conjugative IncI1-type plasmid of 91,942 bp; pHUSEC41-2, a 73,564-bp nonconjugative IncF-type plasmid; and two small plasmids of 7,930 bp and 5,153 bp, designated pHUSEC41-3 and -4, respectively.

Plasmid pHUSEC41-1 displays 131 ORFs and harbors the resistance genes for streptomycin and sulfonamides. The organization comprising *trbC*, *sul2*, *strA*, *bla*$_{\text{TEM-1}}$, and *strB* is similar to p3521, an IncB plasmid (GenBank no. GU256641), and the IncQ RSF 1010 plasmid (GenBank no. M28829) (6). The transfer region of pHUSEC41-1 includes *trb*, *tra*, and *pil*, which are most related to plasmid p026vir (GenBank no. FJ386569).

The pHUSEC41-2 IncF plasmid contains 140 coding sequences (CDS) and is related to p55989 from the enteroaggregative *E. coli* strain 55989 (GenBank no. LB226692). Unlike pHUSEC41-1, it was not transferable to *E. coli* C118 by conjugation. The pHUSEC41-2 transfer region was found to exhibit deletions (e.g., *traV*) similar to those previously seen with other IncF plasmids with impaired conjugation properties (2). Plasmid pHUSEC41-3 displays 15 ORFs. Four of these are related to plasmid ColE1 mobilization proteins (MobA to -D) (GenBank no. J01566) (8).

We found 9 CDS on the smallest plasmid pHUSEC41-4, which resembles (∼70% identity) plasmid ColE1 (8) minus its mobilization module. Comobilization of pHUSEC41-3 and -1 to *E. coli* CC118 occurred with a frequency of 10^{-5} per donor cell.

Nucleotide sequence accession numbers. The plasmid sequences reported here have been deposited in the EMBL database under accession numbers HE603110 (pHUSEC41-1), HE603111 (pHUSEC41-2), HE603112 (pHUSEC41-3), and HE603113 (pHUSEC41-4).

ACKNOWLEDGEMENTS

This study was supported by grants from the Federal Ministry of Education and Research (BMBF, Germany) within the framework of the RESET research network (contract no. 01KI1013G) and the Innovation Funds of the Ministry of Science and Arts of the State of Hessen to C.I. and T.C., respectively.

We thank Christina Gerstmann and Alexandra Amend-Foerster for excellent technical assistance. Furthermore, we acknowledge the German National Reference Center for Salmonellae and Other Bacterial Enteric Pathogens for providing strain HUSEC41.

REFERENCES

1. **Bielaszewska M, et al.** 2011. Characterisation of the *Escherichia coli* strain associated with an outbreak of haemolytic uraemic syndrome in Germany, 2011: a microbiological study. Lancet Infect. Dis. **11**:671–676.
2. **Billard-Pomares T, et al.** 2011. Complete nucleotide sequence of plasmid pTN48, encoding the CTX-M-14 extended-spectrum β-lactamase from

Received 16 October 2011 Accepted 21 October 2011

Address correspondence to Can Imirzalioglu, can.imirzalioglu@mikrobio.med.uni-giessen.de, or Trinad Chakraborty, trinad.chakraborty@mikrobio.med.uni-giessen.de.

C. Künne and A. Billion contributed equally.

doi:10.1128/JB.06368-11

Escherichia coli 0102-ST405 strain. Antimicrob. Agents Chemother. **3**:1270–1273.

3. **Datta N, Hedges RW.** 1971. Compatibility groups among Fi-R factors. Nature **234**:222–223.
4. **Kuenne CT, Ghai R, Chakraborty T, Hain T.** 2007. GECO—linear visualization for comparative genomics. Bioinformatics **23**:125–126.
5. **Mellmann A, et al.** 2011. Prospective genomic characterization of the German enterohemorrhagic Escherichia coli 0104:H4 outbreak by rapid next generation sequencing technology. PLoS One **6**:e22751.
6. **Papagiannitsis CC, Tzouvelekis LS, Kotsakis SD, Tzelepi E, Miriagou V.** 2011S. Sequence of pR3521, an IncB plasmid from *Escherichia coli* encoding ACC-4, SCO-1, and TEM-1 β-lactamases. Antimicrob. Agents Chemother. **55**:376–381.
7. **Pitcher DG, Saunders NA, Owen RJ.** 1989. Rapid extraction of bacterial genomic DNA with guanidium thiocyanate. Lett. Appl. Microbiol. **8**:151–156.
8. **Tomizawa JI, Ohmori H, Bird RE.** 1977. Origin of replication of colicin E1 plasmid DNA. Proc. Natl. Acad. Sci. U. S. A. **74**:1865–1869.

BMC Genomics

DATABASE Open Access

sRNAdb: A small non-coding RNA database for gram-positive bacteria

Jordan Pischimarov[1†], Carsten Kuenne[1†], André Billion[1], Jüergen Hemberger[2], Franz Cemič[2], Trinad Chakraborty[1] and Torsten Hain[1*]

Abstract

Background: The class of small non-coding RNA molecules (sRNA) regulates gene expression by different mechanisms and enables bacteria to mount a physiological response due to adaptation to the environment or infection. Over the last decades the number of sRNAs has been increasing rapidly. Several databases like Rfam or fRNAdb were extended to include sRNAs as a class of its own. Furthermore new specialized databases like sRNAMap (gram-negative bacteria only) and sRNATarBase (target prediction) were established. To the best of the authors' knowledge no database focusing on sRNAs from gram-positive bacteria is publicly available so far.

Description: In order to understand sRNA's functional and phylogenetic relationships we have developed sRNAdb and provide tools for data analysis and visualization. The data compiled in our database is assembled from experiments as well as from bioinformatics analyses. The software enables comparison and visualization of gene loci surrounding the sRNAs of interest. To accomplish this, we use a client–server based approach. Offline versions of the database including analyses and visualization tools can easily be installed locally on the user's computer. This feature facilitates customized local addition of unpublished sRNA candidates and related information such as promoters or terminators using tab-delimited files.

Conclusion: sRNAdb allows a user-friendly and comprehensive comparative analysis of sRNAs from available sequenced gram-positive prokaryotic replicons. Offline versions including analysis and visualization tools facilitate complex user specific bioinformatics analyses.

Background

In recent years numerous small non-coding RNAs (sRNAs) were discovered in bacteria. This class of RNAs is crucial to prokaryotic life, modulating transcription or translation leading to either activation or repression of important physiological processes. sRNAs enable bacteria to trigger rapid physiological responses in order to adapt to the environment or infectious processes [1-3].

To cope with the increasing number of identified sRNAs, databases such as fRNAdb, Rfam, sRNAMap and sRNATarBase were developed [4-9]. All of these approaches have certain drawbacks. fRNAdb contains all classes of RNAs, but allows no further analysis. Rfam is one of the most informative data collections, allowing detailed analyses via a web front-end. sRNAMap is a webserver-based application for gram-negative bacteria only. sRNATarBase compiles experimental data and allows the prediction of sRNA targets. But all databases available to date limit the analysis to published data only. Therefore bioinformatics analyses of candidate sRNAs in combination with genomes, terminators and other relevant information that has not yet been published is still a very complicated task.

In an attempt to overcome some of the aforementioned drawbacks, we have developed sRNAdb. Our database is a locally installable web-suite, permitting the comparative analysis of sRNAs of gram-positive bacteria including their flanking genes. User modified files in GenBank format and gram-negative bacterial genomes, pooled sRNA candidates or further features of interest can be included in locally installed databases. Furthermore all integrated analysis tools can also be used locally.

* Correspondence: Torsten.Hain@mikrobio.med.uni-giessen.de
†Equal contributors
[1]Institute of Medical Microbiology, Justus-Liebig-University, Schubertstrasse 81, Giessen D-35392, Germany
Full list of author information is available at the end of the article

Construction and content

A database scheme of unique keys and entities, combined with corresponding relations and connections is given in Figure 1. Optional user defined extensions to locally installed versions of the database are indicated with a lighter background color than the boxes representing database entities.

Input data

To the best authors' knowledge, no general nomenclature convention for sRNAs exists to date. Therefore sRNAs imported into our database from the literature cannot always be unambiguously distinguished by name, locus or annotation only. Furthermore a large number of published sRNAs is currently annotated as predicted or putative. This leads to a myriad of sRNAs bearing indistinct names, positions or ambiguous annotations. To cope with this difficulty, sRNAdb contains a unique key composed of information about the authors, experimental conditions and sRNA properties as shown in the table termed snrax of Figure 1. Annotated sequences of organisms or plasmids downloaded from NCBI's RefSeq database [10] represent the replicons in the database. Information annotated in GenBank-formatted files such as sequences, or genes filtered from these files are automatically inserted into sRNAdb. When sRNAdb is installed locally, users can furthermore modify the local database by adding customized features such as terminators, promoters and other additional data. Terminators predicted by TransTermHP [11] serve as examples for this option, as described on the official sRNAdb server homepage.

Architecture and design

Our public sRNAdb server is implemented in Java 1.6 on a Debian Linux PC. It facilitates a client–server architecture using Java Server Pages (JSPs), Java Servlets, and Cascading Stylesheets (CSS). Apache Tomcat and MySQL serve as webserver and database, respectively.

Related sRNAs are determined using BLASTN [12], while protein homologies are established by a combination

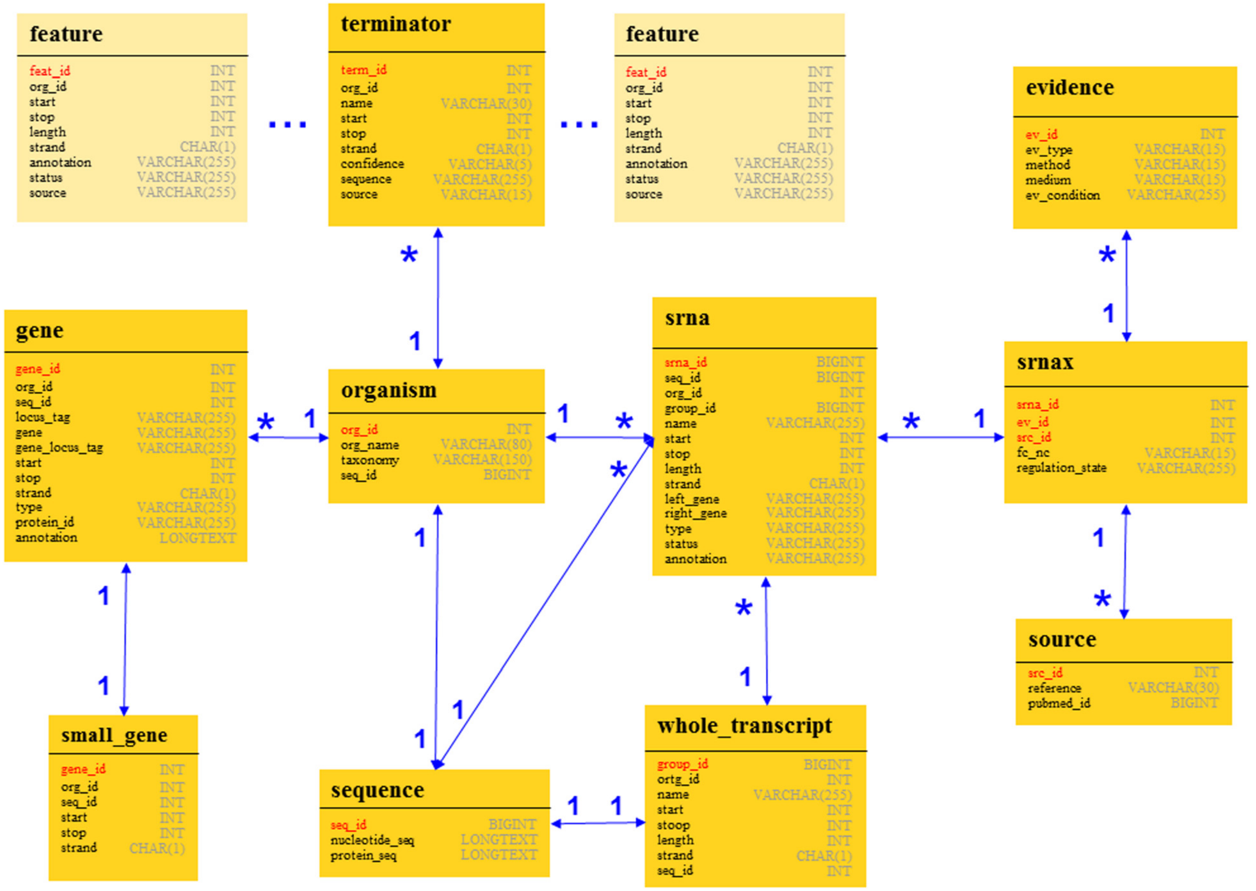


Figure 1 Database schema. The whole database with connections between tables and specific attributes are shown in UML-Notation. Unique and foreign keys of each table are given in bold letters while relations between entities are stated above the connection arrows. Optional features which can be inserted by the user into local versions of the database, are indicated using a lighter background color than employed for boxes representing entities.

of BLASTCLUST and BLASTP [12]. The addition of new data (replicons, sRNAs, terminators, promoters, RBS, etc.) to a local installation of sRNAdb is a simple process based on GenBank and tab-delimited flat-files.

Currently, the public sRNAdb server contains 558 gram-positive genomes and plasmids as well as 9993 automatically predicted and 671 experimentally verified sRNAs. An overview is given in Table 1.

Utility and discussion

The sRNAdb web-database aims to collect all published and predicted sRNAs of gram-positive bacteria for comparative analysis. sRNAs featuring an environmental condition-depending range of sizes can optionally be joined to a combined transcript. The public version of sRNAdb contains terminators predicted by TranstermHP [11]. Three web-interfaces are provided for retrieval and analysis of the data. The first module is called *search* and offers a rich query interface for the database, as shown in Figure 2A. Properties of sRNAs can be selected and filters can be defined to create task-specific queries resulting in a tabular output (Figure 2B). Related or customized data can also be collated to the query, based on the up- or downstream distance to an sRNA of interest. Furthermore, a secondary structure prediction of selected sRNA sequences by energy minimization can be performed using RNAfold (http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi).

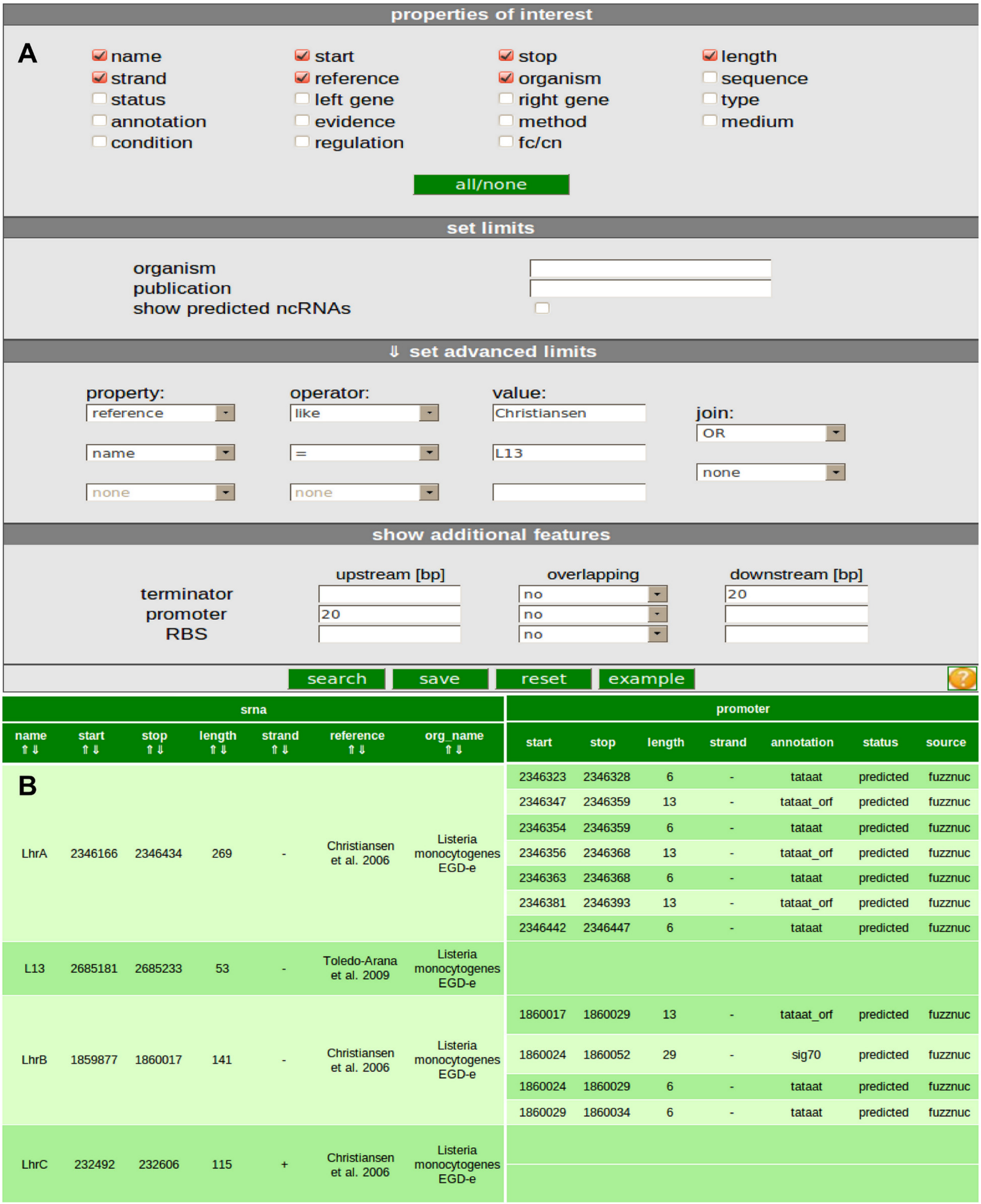
Another interface named *blast* (Figure 3A) was created to enable homology searches of sRNAs versus either public or proprietary sRNAs or whole chromosomes/plasmids using BLASTN [12]. This can be used for initial screening of potential genomic regions. Concise matrix outputs for comparative analysis purposes as shown in Figure 3B and Figure 3C, are implemented. Complete BLAST alignments are displayed in Figure 3D. Sequences from the BLAST output table can be easily selected by setting checkmarks to extract data into a multifasta-formatted file, ready to serve as input to multiple sequence alignment programs such as CLUSTALW (http://www.ebi.ac.uk/Tools/msa/clustalw2/). The resulting output can be used to predict structurally conserved and thermodynamically stable RNA secondary structures using e.g., RNAz (http://rna.tbi.univie.ac.at/cgi-bin/RNAz.cgi), facilitating screens for sRNA-homologs across genomes.

For comprehensive visual assessment the *vision* servlet (Figure 4A) was developed. This allows for a comparative analysis of multiple, related chromosome/plasmid loci of the genomic neighborhood of a single sRNA of interest (single mode) as displayed in Figure 4B. The results are translated into an image (.png-formatted) whereby homologous genes (CDS, RNA) of the sRNA locus are identified by BLASTP [12] and presented with an identical colour code. Terminators and any number of additional features previously defined can be included as desired. Each object in the image is associated with a popup-box, displaying further information and linked to corresponding database entries. The width of the resulting image can be varied to compensate for different screen resolutions. Thus one sRNA locus can be

Table 1 The table shows an overview of the current database entries. These are compiled from experiments or from bioinformatic analyses

Reference	sRNAs	Organism	Pubmed_id
Arnvig et al. 2009 [13]	9	*Mycobacterium tuberculosis* H37Rv	19555452
Bohn et al. 2010 [14]	28	*Staphylococcus aureus subsp. aureus* N315	20511587
Christiansen et al. 2006 [15]	3	*Listeria monocytogenes* EGD-e	16682563
D'Hérouel et al. 2011 [16]	22	*Enterococcus faecalis* V583	21266481
Geissmann et al. 2009 [17]	11	*Staphylococcus aureus* subsp. *aureus* N315	19786493
Irnov et al. 2010 [18]	90	*Bacillus subtilis* subsp. *subtilis* str. 168	20525796
Kumar et al. 2010 [19]	50	*Streptococcus pneumonia* TIGR4	20525227
Livny et al. 2008 [20]	9993	Gram-positive bacteria	18787707
Mandin et al. 2007 [21]	12	*Listeria monocytogenes* EGD-e	17259222
Mraheil et al. 2011 [22]	150	*Listeria monocytogenes* EGD-e	21278422
Nielsen et al. 2008 [23]	1	*Listeria monocytogenes* EGD-e	18621897
Perez et al. 2009 [24]	33	*Streptococcus pyogenes* MGAS5005	19888332
Rasmussen et al. 2009 [25]	84	*Bacillus subtilis* subsp. *subtilis* str. 168	19682248
Tezuka et al. 2009 [26]	12	*Streptomyces griseus* subsp. *griseus* NBRC 13350	19465662
Toledo-Arana et al. 2009 [27]	103	*Listeria monocytogenes* EGD-e	19448609
Vockenhuber et al. 2010 [28]	63	*Streptomyces coelicolor*	21521948

The organisms for which sRNAs are listed in the database, including references, the number of identified sRNAs for the specific organisms and their relevant pumed identification number are listed.



A

properties of interest

☑ name	☑ start	☑ stop	☑ length
☑ strand	☑ reference	☑ organism	□ sequence
□ status	□ left gene	□ right gene	□ type
□ annotation	□ evidence	□ method	□ medium
□ condition	□ regulation	□ fc/cn	

all/none

set limits

organism	
publication	
show predicted ncRNAs	□

⇓ set advanced limits

property:	operator:	value:	join:
reference	like	Christiansen	OR
name	=	L13	none
none	none		

show additional features

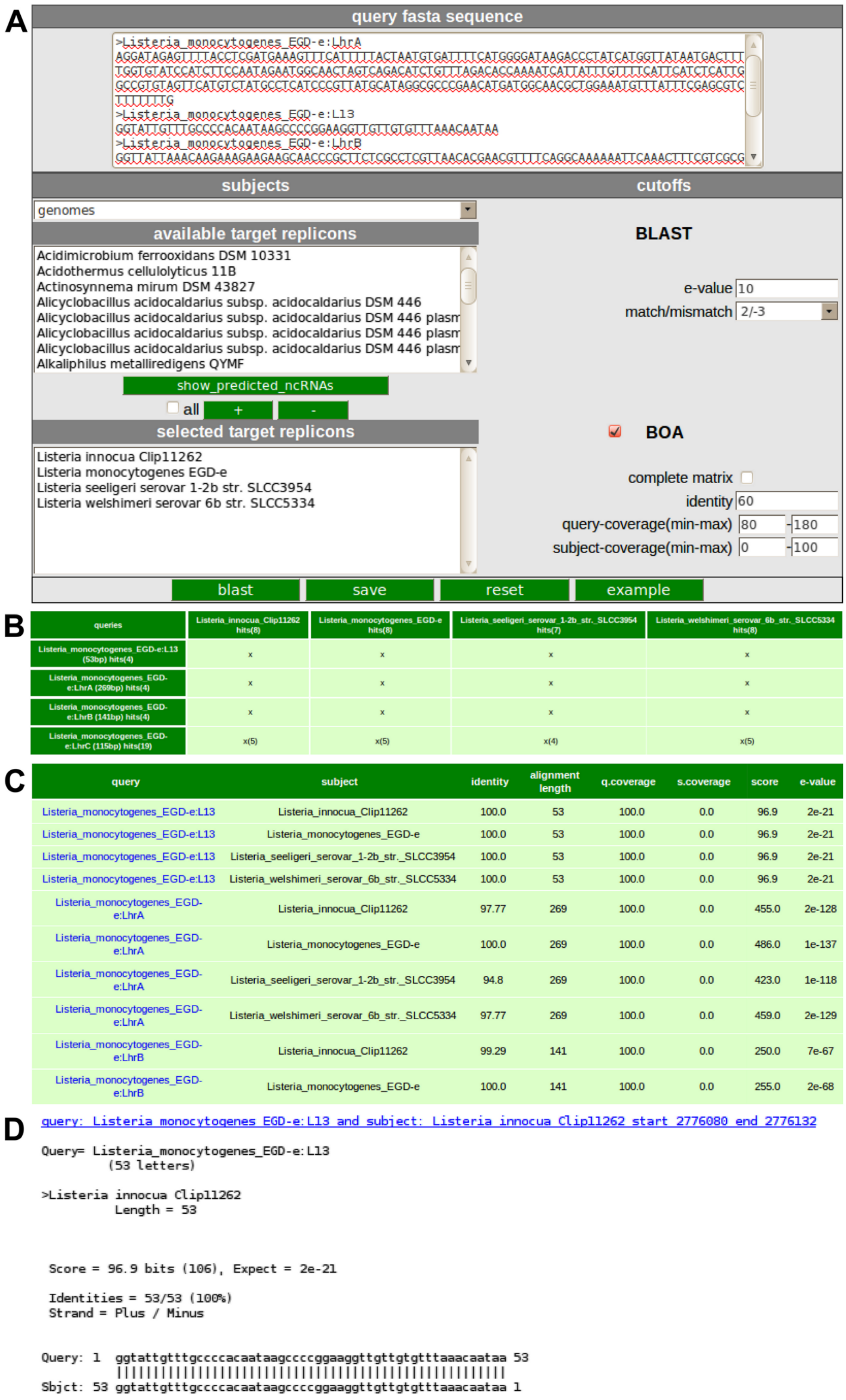
	upstream [bp]	overlapping	downstream [bp]
terminator		no	20
promoter	20	no	
RBS		no	

search save reset example ?

B

srna							promoter						
name ⇑⇓	start ⇑⇓	stop ⇑⇓	length ⇑⇓	strand ⇑⇓	reference ⇑⇓	org_name ⇑⇓	start	stop	length	strand	annotation	status	source
LhrA	2346166	2346434	269	-	Christiansen et al. 2006	Listeria monocytogenes EGD-e	2346323	2346328	6	-	tataat	predicted	fuzznuc
							2346347	2346359	13	-	tataat_orf	predicted	fuzznuc
							2346354	2346359	6	-	tataat	predicted	fuzznuc
							2346356	2346368	13	-	tataat_orf	predicted	fuzznuc
							2346363	2346368	6	-	tataat	predicted	fuzznuc
							2346381	2346393	13	-	tataat_orf	predicted	fuzznuc
							2346442	2346447	6	-	tataat	predicted	fuzznuc
L13	2685181	2685233	53	-	Toledo-Arana et al. 2009	Listeria monocytogenes EGD-e							
LhrB	1859877	1860017	141	-	Christiansen et al. 2006	Listeria monocytogenes EGD-e	1860017	1860029	13	-	tataat_orf	predicted	fuzznuc
							1860024	1860052	29	-	sig70	predicted	fuzznuc
							1860024	1860029	6	-	tataat	predicted	fuzznuc
							1860029	1860034	6	-	tataat	predicted	fuzznuc
LhrC	232492	232606	115	+	Christiansen et al. 2006	Listeria monocytogenes EGD-e							

Figure 2 Search servlet. Properties of interest for each sRNA such as name, start, stop and so forth can be selected by setting check marks in the *properties* section of the servlet form. sRNAs of specific organisms or publications can be selected according to settings defined in the *set limits* section. Furthermore advanced limits for detailed filtering are available. Additional features like promoters and terminators can be searched for in the neighborhood of sRNAs of interest. **B** An example output from the *search* servlet. The resulting table contains four sRNAs named LhrA, LhrB, LhrC and L13. The corresponding search options are shown in **A**. For each sRNA, properties as well as additional features (promoters) in the surrounding area are displayed in intervals of 20 bp. Also the properties as selected with the *search* servlet are included in the output.



query fasta sequence

```
>Listeria_monocytogenes_EGD-e:LhrA
AGGATAGAGTTTTACCTCGATGAAAGTTTCATTTTTACTAATGTGATTTTCATGGGGATAAGACCCTATCATGGTTATAATGACTTT
TGGTGTATCCATCTTCCAATAGAATGGCAACTAGTCAGACATCTGTTTAGACACCAAAATCATTATTTGTTTTCATTCATCTCATTG
GCCGTGTAGTTCATGTCTATGCCTCATCCCGTTATGCATAGGCGCCCGAACATGATGGCAACGCTGGAAATGTTTATTTCGAGCGTC
TTTTTTTG
>Listeria_monocytogenes_EGD-e:L13
GGTATTGTTTGCCCCACAATAAGCCCCGGAAGGTTGTTGTGTTTAAACAATAA
>Listeria_monocytogenes_EGD-e:LhrB
GGTTATTAAACAAGAAAGAAGAAGCAACCCGCTTCTCGCCTCGTTAACACGAACGTTTTCAGGCAAAAAATTCAAACTTTCGTCGCG
```

subjects

genomes

available target replicons

Acidimicrobium ferrooxidans DSM 10331
Acidothermus cellulolyticus 11B
Actinosynnema mirum DSM 43827
Alicyclobacillus acidocaldarius subsp. acidocaldarius DSM 446
Alicyclobacillus acidocaldarius subsp. acidocaldarius DSM 446 plasm
Alicyclobacillus acidocaldarius subsp. acidocaldarius DSM 446 plasm
Alicyclobacillus acidocaldarius subsp. acidocaldarius DSM 446 plasm
Alkaliphilus metalliredigens QYMF

□ all

selected target replicons

Listeria innocua Clip11262
Listeria monocytogenes EGD-e
Listeria seeligeri serovar 1-2b str. SLCC3954
Listeria welshimeri serovar 6b str. SLCC5334

cutoffs

BLAST

e-value 10
match/mismatch 2/-3

☑ BOA

complete matrix □
identity 60
query-coverage(min-max) 80 -180
subject-coverage(min-max) 0 -100

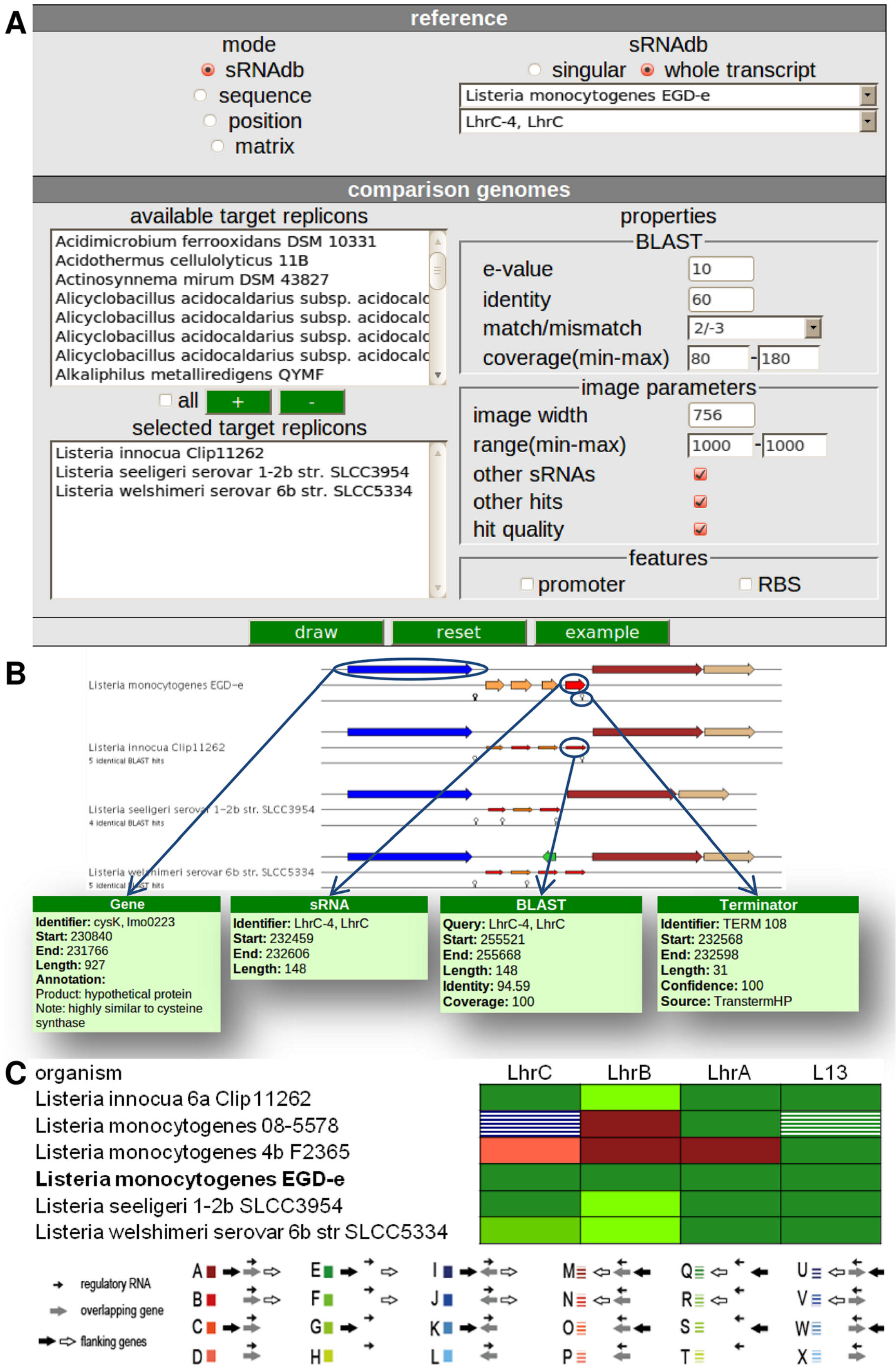
queries	Listeria_innocua_Clip11262 hits(8)	Listeria_monocytogenes_EGD-e hits(8)	Listeria_seeligeri_serovar_1-2b_str._SLCC3954 hits(7)	Listeria_welshimeri_serovar_6b_str._SLCC5334 hits(8)
Listeria_monocytogenes_EGD-e:L13 (53bp) hits(4)	x	x	x	x
Listeria_monocytogenes_EGD-e:LhrA (269bp) hits(4)	x	x	x	x
Listeria_monocytogenes_EGD-e:LhrB (141bp) hits(4)	x	x	x	x
Listeria_monocytogenes_EGD-e:LhrC (115bp) hits(19)	x(5)	x(5)	x(4)	x(5)

query	subject	identity	alignment length	q.coverage	s.coverage	score	e-value
Listeria_monocytogenes_EGD-e:L13	Listeria_innocua_Clip11262	100.0	53	100.0	0.0	96.9	2e-21
Listeria_monocytogenes_EGD-e:L13	Listeria_monocytogenes_EGD-e	100.0	53	100.0	0.0	96.9	2e-21
Listeria_monocytogenes_EGD-e:L13	Listeria_seeligeri_serovar_1-2b_str._SLCC3954	100.0	53	100.0	0.0	96.9	2e-21
Listeria_monocytogenes_EGD-e:L13	Listeria_welshimeri_serovar_6b_str._SLCC5334	100.0	53	100.0	0.0	96.9	2e-21
Listeria_monocytogenes_EGD-e:LhrA	Listeria_innocua_Clip11262	97.77	269	100.0	0.0	455.0	2e-128
Listeria_monocytogenes_EGD-e:LhrA	Listeria_monocytogenes_EGD-e	100.0	269	100.0	0.0	486.0	1e-137
Listeria_monocytogenes_EGD-e:LhrA	Listeria_seeligeri_serovar_1-2b_str._SLCC3954	94.8	269	100.0	0.0	423.0	1e-118
Listeria_monocytogenes_EGD-e:LhrA	Listeria_welshimeri_serovar_6b_str._SLCC5334	97.77	269	100.0	0.0	459.0	2e-129
Listeria_monocytogenes_EGD-e:LhrB	Listeria_innocua_Clip11262	99.29	141	100.0	0.0	250.0	7e-67
Listeria_monocytogenes_EGD-e:LhrB	Listeria_monocytogenes_EGD-e	100.0	141	100.0	0.0	255.0	2e-68

query: Listeria monocytogenes EGD-e:L13 and subject: Listeria innocua Clip11262 start 2776080 end 2776132

```
Query= Listeria_monocytogenes_EGD-e:L13
         (53 letters)

>Listeria innocua Clip11262
          Length = 53


 Score = 96.9 bits (106), Expect = 2e-21

 Identities = 53/53 (100%)
 Strand = Plus / Minus


Query: 1   ggtattgtttgccccacaataagccccggaaggttgttgtgtttaaacaataa 53
            |||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 53 ggtattgtttgccccacaataagccccggaaggttgttgtgtttaaacaataa 1
```

Figure 3 Blast servlet form and corresponding output. A FASTA formatted sRNA sequences can be inserted into the query box. Also target genomes or sRNAs have to be selected for multiple alignment using BLAST. For a detailed BLAST analysis the BLAST output analysis (BOA) options has to be selected. In this example four sRNAs resulting from a search with parameters shown in Figure 1 were selected as input. Genomes of the genus *Listeria* were set as targets and the BOA options were enabled. **B** The number of sRNAs detected in the target organism is displayed in a comparative matrix form. **C** All hits listed in a table and are linked to their corresponding alignment. **D** A detailed BLAST alignment of all results can also be plotted.



reference

mode
● sRNAdb
○ sequence
○ position
○ matrix

sRNAdb
○ singular ● whole transcript
Listeria monocytogenes EGD-e
LhrC-4, LhrC

comparison genomes

available target replicons
Acidimicrobium ferrooxidans DSM 10331
Acidothermus cellulolyticus 11B
Actinosynnema mirum DSM 43827
Alicyclobacillus acidocaldarius subsp. acidocalc
Alicyclobacillus acidocaldarius subsp. acidocalc
Alicyclobacillus acidocaldarius subsp. acidocalc
Alicyclobacillus acidocaldarius subsp. acidocalc
Alkaliphilus metalliredigens QYMF

□ all + -

selected target replicons
Listeria innocua Clip11262
Listeria seeligeri serovar 1-2b str. SLCC3954
Listeria welshimeri serovar 6b str. SLCC5334

properties

BLAST

e-value	10
identity	60
match/mismatch	2/-3
coverage(min-max)	80 - 180

image parameters

image width	756
range(min-max)	1000 - 1000
other sRNAs	☑
other hits	☑
hit quality	☑

features
□ promoter □ RBS

draw
reset
example
Listeria monocytogenes EGD-e
Listeria innocua Clip11262
5 identical BLAST hits
Listeria seeligeri serovar 1-2b str. SLCC3954
4 identical BLAST hits
Listeria welshimeri serovar 6b str. SLCC5334
5 identical BLAST hits

Gene
Identifier: cysK, lmo0223
Start: 230840
End: 231766
Length: 927
Annotation:
Product: hypothetical protein
Note: highly similar to cysteine synthase

sRNA
Identifier: LhrC-4, LhrC
Start: 232459
End: 232606
Length: 148

BLAST
Query: LhrC-4, LhrC
Start: 255521
End: 255668
Length: 148
Identity: 94.59
Coverage: 100

Terminator
Identifier: TERM 108
Start: 232568
End: 232598
Length: 31
Confidence: 100
Source: TranstermHP

organism	LhrC	LhrB	LhrA	L13
Listeria innocua 6a Clip11262	🟩	🟩	🟩	🟩
Listeria monocytogenes 08-5578	▤	🟥	🟩	▤
Listeria monocytogenes 4b F2365	🟧	🟥	🟥	🟩
Listeria monocytogenes EGD-e	🟩	🟩	🟩	🟩
Listeria seeligeri 1-2b SLCC3954	🟩	🟩	🟩	🟩
Listeria welshimeri serovar 6b str SLCC5334	🟩	🟩	🟩	🟩

Figure 4 Vision servlet forms and result of single and batch mode. Different input options are available. After selecting the sRNA of interest, replicons can be selected for visualization. Options for further analyses based on BLAST, as well as properties relating to the image output can be set. **A** An example relating to the LhrC transcript is displayed. **B** Single mode: the resulting image shows a comparative representation of a single sRNA candidate and flanking genes in selected organisms. Moving the mouse pointer over these, the corresponding properties of each object is shown in a separate popup window. **C** Batch mode: sRNAs displayed in Figure 1 are used as input in this example. The output-matrix indicates occurrence of the sRNA candidates in selected organisms and their directional relationships with respect to their surrounding genes.

compared to different chromosomes/plasmids in a concise image output.

For the genome wide analysis of multiple sRNA loci an additional batch mode is available. Results from an application of this batch mode have already been published by Mraheil and collaborators [22]. In order to permit this global analysis an option was implemented that enables export of the data to an Excel sheet. This contains a visualization matrix (Figure 4C) which indicates the occurrence of the sRNA of interest in the target organism together with its directional relationships of the flanking genes.

The software tool presented here is a valuable extension to existing solutions and will assist in the rapid analysis of large volumes of data to understand the distribution and evolution of sRNAs in bacteria. Compared to other databases the comparative batch mode of sRNAdb's *vision* servlet facilitates analyses such as *in silico* screening for phylogenetic markers, or identification of drug targets related to bacterial sRNAs. As exemplified by Mraheil and colleagues [22] a grouping of sRNAs from pathogenic, apathogenic or non-pathogenic bacterial strains based on the *vision* servlet´s result matrix, allows the user to identify sRNAs as putative phylogenetic markers. Specifically, sRNAs found exclusively in pathogenic strains can be identified as drug target candidates. Furthermore after download and local installation of sRNAdb, both the database and the dedicated software tools are available to the user. Since proprietary replicons or putative sRNAs can easily be included into locally installed versions of the database, these may be analysed making use of the full power of sRNAdb's software tools, simplifying detailed analyses of unpublished bacterial replicons or sRNA candidates. To the best of the author's knowledge, this functionality is currently not supported by any other publicly available sRNA database.

Conclusion

sRNAdb offers biologists an easy access and analysis to both proprietary and public data and allows the identification of a core set of sRNAs which can be used as putative drug targets in antimicrobial therapeutic approaches as well as specific sRNAs for potential diagnostic markers for the detection of gram-positive bacteria.

Availability and requirements

The database including documentation and tools for analysis are available free of charge at http://bioinfo.mikrobio.med.uni-giessen.de/sRNAdb.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Designed and implemented the database and related software tools: JP, CK, AB, TH. Analyzed data: JP, CK, AB, FC, TH. Wrote the paper: JP, CK, JH, FC, TC, TH. All authors read and approved the final manuscript.

Funding

This work was supported by grants from the German Federal Ministry of Education and Research (BMBF ERA-NET) Pathogenomics Network to the sncRNAomics project (62080061) to T.H. and the German Centre for Infection Research, Justus-Liebig University Giessen.

Author details

[1]Institute of Medical Microbiology, Justus-Liebig-University, Schubertstrasse 81, Giessen D-35392, Germany. [2]Institute for Biochemical Engineering and Analytics, University of Applied Sciences Giessen-Friedberg, Wiesenstrasse 14, Giessen D-35390, Germany.

Received: 3 April 2012 Accepted: 1 July 2012
Published: 10 August 2012

References

1. Frohlich KS, Vogel J: **Activation of gene expression by small RNA.** *Curr Opin Microbiol* 2009, **12**:674–682.
2. Mraheil MA, Billion A, Kuenne C, Pischimarov J, Kreikemeyer B, Engelmann S, Hartke A, Giard JC, Rupnik M, Vorwerk S, Beier M, Retey J, Hartsch T, Jacob A, Cemic F, Hemberger J, Chakraborty T, Hain T: **Comparative genome-wide analysis of small RNAs of major Gram-positive pathogens: from identification to application.** *Microb Biotechnol* 2010, **3**:658–676.
3. Waters LS, Storz G: **Regulatory RNAs in bacteria.** *Cell* 2009, **136**:615–628.
4. Cao Y, Wu J, Liu Q, Zhao Y, Ying X, Cha L, Wang L, Li W: **sRNATarBase: a comprehensive database of bacterial sRNA targets verified by experiments.** *RNA* 2010, **16**:2051–2057.
5. Kin T, Yamada K, Terai G, Okida H, Yoshinari Y, Ono Y, Kojima A, Kimura Y, Komori T, Asai K: **fRNAdb: a platform for mining/annotating functional RNA candidates from non-coding RNA sequences.** *Nucleic Acids Res* 2007, **35**:D145–D148.
6. Mituyama T, Yamada K, Hattori E, Okida H, Ono Y, Terai G, Yoshizawa A, Komori T, Asai K: **The Functional RNA Database 3.0: databases to support mining and annotation of functional RNAs.** *Nucleic Acids Res* 2009, **37**:D89–D92.
7. Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL, Lindgreen S, Wilkinson AC, Finn RD, Griffiths-Jones S, Eddy SR, Bateman A: **Rfam: updates to the RNA families database.** *Nucleic Acids Res* 2009, **37**:D136–D140.
8. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR: **Rfam: an RNA family database.** *Nucleic Acids Res* 2003, **31**:439–441.
9. Huang HY, Chang HY, Chou CH, Tseng CP, Ho SY, Yang CD, Ju YW, Huang HD: **sRNAMap: genomic maps for small non-coding RNAs, their regulators and their targets in microbial genomes.** *Nucleic Acids Res* 2009, **37**:D150–D154.
10. Pruitt KD, Katz KS, Sicotte H, Maglott DR: **Introducing RefSeq and LocusLink: curated human genome resources at the NCBI.** *Trends Genet* 2000, **16**:44–47.
11. Kingsford CL, Ayanbule K, Salzberg SL: **Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake.** *Genome Biol* 2007, **8**:R22.
12. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**:403–410.
13. Arnvig KB, Young DB: **Identification of small RNAs in *Mycobacterium tuberculosis*.** *Mol Microbiol* 2009, **73**:397–408.
14. Bohn C, Rigoulay C, Chabelskaya S, Sharma CM, Marchais A, Skorski P, Borezee-Durant E, Barbet R, Jacquet E, Jacq A, Gautheret D, Felden B, Vogel J, Bouloc P: **Experimental discovery of small RNAs in *Staphylococcus aureus* reveals a riboregulator of central metabolism.** *Nucleic Acids Res* 2010, **38**:6620–6636.
15. Christiansen JK, Nielsen JS, Ebersbach T, Valentin-Hansen P, Sogaard-Andersen L, Kallipolitis BH: **Identification of small Hfq-binding RNAs in *Listeria monocytogenes*.** *RNA* 2006, **12**:1383–1396.
16. Fouquier DA, Wessner F, Halpern D, Ly-Vu J, Kennedy SP, Serror P, Aurell E, Repoila F: **A simple and efficient method to search for selected primary transcripts: non-coding and antisense RNAs in the human pathogen *Enterococcus faecalis*.** *Nucleic Acids Res* 2011, **39**:e46.

17. Geissmann T, Chevalier C, Cros MJ, Boisset S, Fechter P, Noirot C, Schrenzel J, Francois P, Vandenesch F, Gaspin C, Romby P: A search for small noncoding RNAs in *Staphylococcus aureus* reveals a conserved sequence motif for regulation. *Nucleic Acids Res* 2009, **37**:7239–7257.
18. Irnov I, Sharma CM, Vogel J, Winkler WC: Identification of regulatory RNAs in *Bacillus subtilis*. *Nucleic Acids Res* 2010, **38**:6637–6651.
19. Kumar R, Shah P, Swiatlo E, Burgess SC, Lawrence ML, Nanduri B: Identification of novel non-coding small RNAs from *Streptococcus pneumoniae* TIGR4 using high-resolution genome tiling arrays. *BMC Genomics* 2010, **11**:350.
20. Livny J, Teonadi H, Livny M, Waldor MK: High-throughput, kingdom-wide prediction and annotation of bacterial non-coding RNAs. *PLoS One* 2008, **3**:e3197.
21. Mandin P, Repoila F, Vergassola M, Geissmann T, Cossart P: Identification of new noncoding RNAs in *Listeria monocytogenes* and prediction of mRNA targets. *Nucleic Acids Res* 2007, **35**:962–974.
22. Mraheil MA, Billion A, Mohamed W, Mukherjee K, Kuenne C, Pischimarov J, Krawitz C, Retey J, Hartsch T, Chakraborty T, Hain T: The intracellular sRNA transcriptome of *Listeria monocytogenes* during growth in macrophages. *Nucleic Acids Res* 2011, **39**:4235–4248.
23. Nielsen JS, Olsen AS, Bonde M, Valentin-Hansen P, Kallipolitis BH: Identification of a sigma B-dependent small noncoding RNA in *Listeria monocytogenes*. *J Bacteriol* 2008, **190**:6264–6270.
24. Perez N, Trevino J, Liu Z, Ho SC, Babitzke P, Sumby P: A genome-wide analysis of small regulatory RNAs in the human pathogen group A *Streptococcus*. *PLoS One* 2009, **4**:e7668.
25. Rasmussen S, Nielsen HB, Jarmer H: The transcriptionally active regions in the genome of *Bacillus subtilis*. *Mol Microbiol* 2009, **73**:1043–1057.
26. Tezuka T, Hara H, Ohnishi Y, Horinouchi S: Identification and gene disruption of small noncoding RNAs in *Streptomyces griseus*. *J Bacteriol* 2009, **191**:4896–4904.
27. Toledo-Arana A, Dussurget O, Nikitas G, Sesto N, Guet-Revillet H, Balestrino D, Loh E, Gripenland J, Tiensuu T, Vaitkevicius K, Barthelemy M, Vergassola M, Nahori MA, Soubigou G, Regnault B, Coppee JY, Lecuit M, Johansson J, Cossart P: The *Listeria* transcriptional landscape from saprophytism to virulence. *Nature* 2009, **459**:950–956.
28. Vockenhuber MP, Sharma CM, Statt MG, Schmidt D, Xu Z, Dietrich S, Liesegang H, Mathews DH, Suess B: Deep sequencing-based identification of small non-coding RNAs in *Streptomyces coelicolor*. *RNA Biol* 2011, **8**:468–477.

doi:10.1186/1471-2164-13-384
Cite this article as: Pischimarov *et al.*: sRNAdb: A small non-coding RNA database for gram-positive bacteria. *BMC Genomics* 2012 **13**:384.

BMC Genomics

RESEARCH ARTICLE Open Access

Comparative genomics and transcriptomics of lineages I, II, and III strains of *Listeria monocytogenes*

Torsten Hain[1], Rohit Ghai[1], André Billion[1], Carsten Tobias Kuenne[1], Christiane Steinweg[1], Benjamin Izar[1], Walid Mohamed[1], Mobarak Abu Mraheil[1], Eugen Domann[1], Silke Schaffrath[1], Uwe Kärst[2], Alexander Goesmann[3], Sebastian Oehm[3], Alfred Pühler[3], Rainer Merkl[4], Sonja Vorwerk[5], Philippe Glaser[6], Patricia Garrido[7], Christophe Rusniok[7], Carmen Buchrieser[7], Werner Goebel[8] and Trinad Chakraborty[1*]

Abstract

Background: *Listeria monocytogenes* is a food-borne pathogen that causes infections with a high-mortality rate and has served as an invaluable model for intracellular parasitism. Here, we report complete genome sequences for two *L. monocytogenes* strains belonging to serotype 4a (L99) and 4b (CLIP80459), and transcriptomes of representative strains from lineages I, II, and III, thereby permitting in-depth comparison of genome- and transcriptome -based data from three lineages of *L. monocytogenes*. Lineage III, represented by the 4a L99 genome is known to contain strains less virulent for humans.

Results: The genome analysis of the weakly pathogenic L99 serotype 4a provides extensive evidence of virulence gene decay, including loss of several important surface proteins. The 4b CLIP80459 genome, unlike the previously sequenced 4b F2365 genome harbours an intact *inlB* invasion gene. These lineage I strains are characterized by the lack of prophage genes, as they share only a single prophage locus with other *L. monocytogenes* genomes 1/2a EGD-e and 4a L99. Comparative transcriptome analysis during intracellular growth uncovered adaptive expression level differences in lineages I, II and III of *Listeria*, notable amongst which was a strong intracellular induction of flagellar genes in strain 4a L99 compared to the other lineages. Furthermore, extensive differences between strains are manifest at levels of metabolic flux control and phosphorylated sugar uptake. Intriguingly, prophage gene expression was found to be a hallmark of intracellular gene expression. Deletion mutants in the single shared prophage locus of lineage II strain EGD-e 1/2a, the *lma* operon, revealed severe attenuation of virulence in a murine infection model.

Conclusion: Comparative genomics and transcriptome analysis of *L. monocytogenes* strains from three lineages implicate prophage genes in intracellular adaptation and indicate that gene loss and decay may have led to the emergence of attenuated lineages.

Keywords: *Listeria monocytogenes*, Lineage, Comparative genomics, Gene decay, Comparative transcriptomics, Flagella, Prophage, Monocin, Isogenic deletion mutants, Murine infection

* Correspondence: Trinad.Chakraborty@mikrobio.med.uni-giessen.de
[1]Institute of Medical Microbiology, Justus-Liebig-University, Schubertstrasse 81, Giessen, D-35392, Germany
Full list of author information is available at the end of the article

Background

Listeria monocytogenes is a Gram-positive, motile, non-sporulating, rod shaped bacterium. It is the causative agent of listeriosis, a food-borne disease, which afflicts both humans and animals. There are only eight species in the entire genus, *L. monocytogenes*, *L. marthii*, *L. innocua*, *L. seeligeri*, *L. welshimeri*, *L. ivanovii*, *L. grayi* and *L. rocourtiae*. *L. monocytogenes* and *L. ivanovii* are the pathogenic species while the others are apathogenic [1,2]. In the genus *Listeria*, non-pathogenic species have been hypothesized to have evolved through genome reduction from pathogenic progenitor strains [3]. *L. monocytogenes* is able to invade and replicate in both phagocytic and non-phagocytic cells. The infectious life cycle has been elucidated in detail, and several virulence factors, essential for each stage of infection have been identified [4,5]. Pathogenic listeriae encode several virulence factors that are localized in a virulence gene cluster (*vgc*) or *Listeria* pathogenicity island-1 (LIPI-1) in the genome. However, a number of genes required for virulence are not localized in this cluster, including the two internalins *inlA* and *inlB*. These encode proteins that are expressed on the surface of the bacterium and facilitate the entry of the bacterium into the eukaryotic cell and their incorporation into a membrane-bound vacuole [6,7]. Further pathogenicity islands present in the genus *Listeria* code for multiple internalins and additional hemolysin genes in species *L. ivanovii* (LIPI-2) [8] and a subset of strains of lineage I (LIPI-3) [9].

Within the four lineages of *L. monocytogenes*, strains are generally classified by serotyping or MLST [10,11], of which 1/2a, 1/2b and 4b are most commonly associated with human listerial infections [2,12]. The first outbreak of *L. monocytogenes* was described for the strain EGD-e, a serotype 1/2a strain of lineage II, following an epidemic in rabbits and guinea pigs in 1926 by E.G.D. Murray [13]. This strain has become a model *Listeria* strain, and was the first listerial strain to be completely sequenced, along with the non-pathogenic *Listeria innocua* 6a CLIP11262 [14]. Subsequently, the first genome of a 4b serotype strain (F2365) of lineage I was completely sequenced [14,15]. It was isolated from Jalisco cheese during a listeriosis outbreak in California in 1985 and mainly associated with pregnancy-related cases. However, it has been recently shown that this strain contains nonsense and frameshift mutations in several genes. Owing to a frameshift in *inlB*, F2365 is severely compromised in Caco-2 invasion assays [16].

Here we report thus the genome sequence of a clinical isolate of the 4b serotype of lineage I, the *L. monocytogenes* 4b strain CLIP80459 that was isolated in a clinical outbreak of listeriosis in France affecting 42 persons [17]. We also present the complete genome sequencing of *L. monocytogenes* strain 4a L99 of lineage III. L99 was originally isolated from food by Kampelmacher in 1950s in the Netherlands. This strain is attenuated in its virulence properties and exhibits a restricted ability to grow within the liver and spleen of infected mice [18]. The availability of the complete genome of *L. monocytogenes* EGD-e serotype 1/2a has permitted analysis of the intracellular gene expression profile of this strain [19-21].

The genome sequences of strains 4a L99 and 4b CLIP80459 presented in this work provide a unique opportunity to delineate specific adaptations of these lineage representives both at the genomic and at the transcriptional level.

Results

General features of complete genomes of three lineages of *L. Monocytogenes*

The overall features of the completely sequenced circular genomes of *L. monocytogenes* 4a L99, *L. monocytogenes* 4b CLIP80459, *L. monocytogenes* 1/2a EGD-e, *L. monocytogenes* 4b F2365 and *L. innocua* 6a CLIP11262 are given in Table 1. Computational multi-virulence-locus sequence typing (MVLST) [22] analysis showed that strain 4b CLIP80459 belongs to epidemic clone ECII and strain 4b F2365 to epidemic clone ECI as previously reported by Nelson and colleagues [15], respectively. The *L. monocytogenes* genomes are remarkably syntenic: genome size, G + C content, percentage coding and average length of protein-coding genes are similar among all four strains (which was previously reported for other listerial genomes) [14,15]. All four *L. monocytogenes* genomes harbour 67 tRNA genes and contain six complete copies of rRNA operons (16 S-23 S-5 S), of which two are located on the right and four on the left replichore. The chromosomes of 4a L99 and 4b CLIP80459 are devoid of mobile genetic elements and harbour no plasmid.

We observed four different prophage regions in the genome of the 4a L99 and only one in the 4b CLIP80459 strain (see prophage region II). *L. monocytogenes* 4a L99 prophage I is located at position 71438 bp (*lmo4a_0064-lmo4a_0115*), prophage II at (*lmo4a_0148-lmo4a 0153*, prophage-remnant: *lmaDC*; 4b ClIP80459 Lm4b_00117b-Lm4b00134 or monocin region), prophage III at 1224779 bp (*lmo4a_1221-lmo4a_1293*) and prophage IV at 2668913 bp (*lmo4a_2599-lmo4a_2658*). Two prophage regions, I and III, are located adjacent to tRNAs. Prophage region I is flanked by *tRNA*^{Lys} and prophage region III is inserted within the region between the gene for *tRNA*^{Arg} and *ydeI* compared to *L. monocytogenes* 1/2a EGD-e. At this very chromosomal location in *L. welshimeri* 6b SLCC5334 there is an insertion of a prophage [3,23,24], while *L. ivanovii* harbours the species-specific *Listeria* pathogenicity island 2 (LIPI-2), which contains a sphingomyelinase C (SmcL) and also a cluster of internalin genes [8]. These findings confirm previous observations [3]

Table 1 General features of *L. monocytogenes* 1/2a EGD-e, *L. monocytogenes* 4a L99, *L. monocytogenes* 4b CLIP80459, *L. monocytogenes* 4b F2365 and *L. innocua* 6a CLIP11262

	L. monocytogenes 4a L99	*L. monocytogenes* 4b CLIP80459	*L. monocytogenes* 4b F2365	*L. monocytogenes* 1/2a EGD-e	*L. innocua*6a CLIP11262
Size of chromosome [bp]	2979198	2912690	2905187	2944528	3011208
G + C content [%]	38.2	38.1	38.0	38.0	37.4
G + C content of protein-coding genes [%]	38.7	38.5	38.5	38.4	37.8
Protein-coding genes (pseudogenes)	2925 (1)	2790 (24)	2821 (26)	2855 (9)	2981 (13)
Average length of protein-coding genes [aa]	301	311	303	306	300
Number of rRNA operons (16 S-23 S-5 S)	6	6	6	6	6
Number of tRNA genes	67	67	67	67	66
Percentage coding	88.9	89.4	88.4	89.2	89.2
Number of prophages (genes)	4 (191)	1 (16)	1 (16)	2 (79)	6 (322)
Plasmid	0	0	0	0	1
Number of strain-specific genes*	111	49	105	120	89
Number of orthologous genes*	2623	2725	2699	2656	2570
Number of transposons	0	0	0	1	0
***Prophage genes excepted.**					

Core and specific genes were analyzed using orthologous pairs excluding prophage genes as described previously [3].

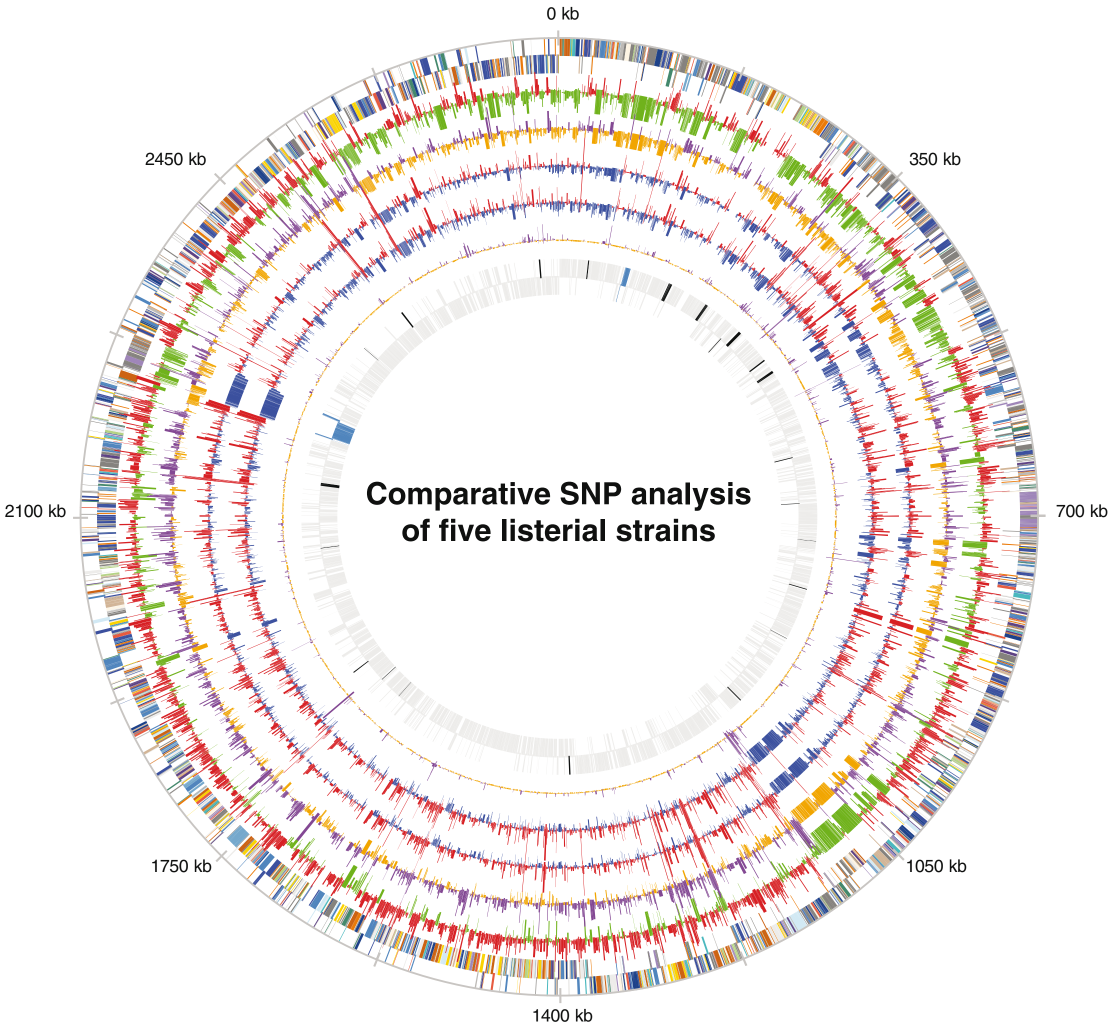
indicating that tRNAs represent genetic “anchoring elements” for the uptake of listerial prophage DNA by transduction processes and thus contributing to evolutionary genome diversity of listeriae. Pseudogenes were detected for both 4b F2365 (24 pseudogenes) and 4b CLIP80459 (26 pseudogenes) genomes respectively, which is a higher number compared to that seen in *L. monocytogenes* 1/2a EGD-e (9 pseudogenes), *L. monocytogenes* 4a L99 (one pseudogene) and *L. innocua* (13 pseudogenes).

When comparing the two *L. monocytogenes* 4b genomes (CLIP80459 and F2365) 115 genes are specific for strain 4b CLIP80459 with respect to strain 4b F2365. The dominant functions encoded by these genes are related to sugar metabolism as they comprise five PTS systems and five sugar permeases or sugar transporters. Furthermore, four transcriptional regulators and four surface anchored proteins are specific to 4b CLIP80459 indicating differences in regulation, sugar metabolism and surface characteristics between the two strains. Of the 146 genes found to be specific for strain 4b F2365, the majority were of unknown function, apart from a PTS system and a specific surface protein. Most interestingly, *inlB* although it is reported to be important for virulence of *L. monocytogenes* has a frameshift mutation in this strain [15].

When comparing the genomes of different lineages at the nucleotide sequence level a number of genomic differences were revealed (Figure 1). Surface proteins showed the highest number of single nucleotide polymorphisms (SNPs). Even in the comparison of the two closely related 4b genomes, two LPXTG-motif containing proteins were identified as encoding a large number of SNPs. One of these, *lm4b_01142* shares substantial similarity to internalins. Comparison of the 4a L99 and the 1/2a EGD-e genomes reflected larger evolutionary divergence, but once again involved surface proteins, such as the LPXTG-motif containing protein *lmo1799*, internalin *lmo0409* (*inlF*), autolysin *lmo1215*, as well as proteins involved in surface antigen biosynthesis like *lmo2552* (*murZ*) and *lmo2549* (*gtcA*). Further analysis identified genes that are most divergent in the three lineages and classification of the most divergent orthologous gene groups was performed (Additional file 1: Table S1). Thus, distribution of SNPs in *Listeria* suggests considerable evolutionary adaptation among surface-associated genes.

Comparison of the virulence genes cluster of lineage I, II and III

All genes of the virulence gene cluster are present in the four studied strains [27]. We performed a nucleotide sequence alignment of the entire virulence genes cluster, using the EGD-e sequence as a reference. As shown in Figure 2 we identified a truncation in the *actA* sequence of the 4b and the 4a genomes. In addition, a small truncation upstream the *mpl* gene and a truncation of a short repeat region distal to the PrfA binding box of *mpl* was present in the 4a genome. However, the PrfA binding site was not affected. Moreover, the alignment



	L. monocytogenes **1/2a EGD-e vs.** ***L. innocua*** **6a CLIP11262**	***L. monocytogenes*** **1/2a EGD-e vs.** ***L. monocytogenes*** **4a L99**	***L. monocytogenes*** **1/2a EGD-e vs.** ***L. monocytogenes*** **4b CLIP80459**	***L. monocytogenes*** **1/2a EGD-e vs.** ***L. monocytogenes*** **4b F2365**	***L. monocytogenes*** **4b F2365 vs.** ***L. monocytogenes*** **4b CLIP80459**
% nucleotide divergence (surface-associated CDS)	**11.14**	**7.00**	**5.16**	**5.09**	**0.59**
% nucleotide divergence (non-surface-associated CDS)	**10.77**	**6.38**	**5.01**	**4.99**	**0.57**
% nucleotide divergence ratio	**3.46**	**9.79**	**2.96**	**2.02**	**4.55**

Figure 1 Comparative SNP analysis of five listerial strains From outside to inside: genome of *L. monocytogenes* 1/2a EGD-e colored according to COG categories (two strands shown separately). Number of SNPs normalized by gene length in the comparison of 1/2a EGD-e and *L. innocua* 6a CLIP11262, 1/2a EGD-e and 4a L99, 1/2a EGD-e and 4b CLIP80459, 1/2a EGD-e and 4b F2365, and the two 4b strains (4b F2365 and 4b CLIP80459). The innermost circle shows the location of phage genes (blue) and virulence genes (black) in the 1/2a EGD-e genome. Line graphs indicate the number of SNPs/gene length reflecting loci in the genome having a disproportionate number of SNPs. However, if a gene is specific to a certain genome, this will also be shown as a peak indicating a region of divergence within the two genomes under comparison. This analysis was performed using the MUMmer package [25] and SNPs were mapped to coding regions using PERL scripts. Data were visualized by GenomeViz [26]. For each pairwise comparison of strains, percentage of SNPs per gene length of surface- and non-surface-associated genes, as well as the ratio of these values is given in the table. The latter was named "nucleotide divergence ratio" and denotes the relative amount of difference between those two classes of genes, in order to identify more (positive value) or less (negative value) abundant mutation in surface-associated than in non-surface-associated genes.

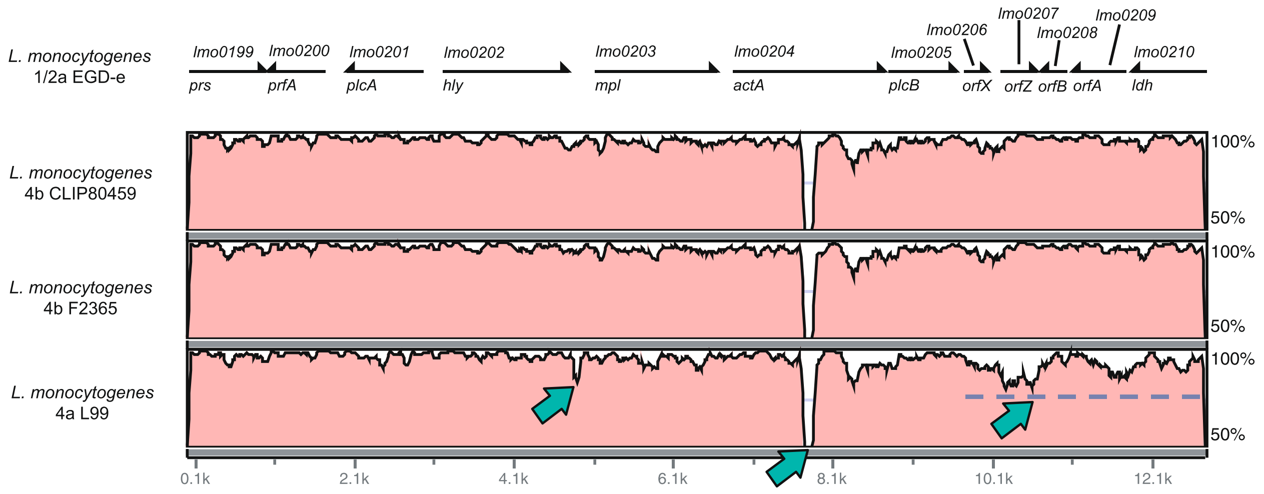


Figure 2 Alignment of the virulence gene cluster of representatives of three *L. monocytogenes* lineages *L. monocytogenes* 1/2a EGD-e was used as reference genome. Nucleotide sequence identity of compared genomes is visualized. The top panel indicates location and direction of virulence genes.

identity decreased slightly in the latter half of the cluster, with differences most prominently visible in the regions containing *lmo0207* and *lmo0209*. *lmo0207* encodes a lipoprotein and was identified as one of the most divergent genes of the LIPI-1 when comparing three lineages.

Interestingly, both the *L. monocytogenes* 4b strains (CLIP80459 and F2365), and the *L. monocytogenes* 4a L99 strain, have an identical repeat truncation in the ActA protein compared to ActA of the 1/2a EGD-e (Additional file 2: Table S2 Additional file 3: Table S3). Such truncations in *actA* have been reported previously for strain 4a L99 and affect the speed of movement of intracellular bacteria [28]. We surveyed sequenced *act*A alleles present in GenBank and discovered that the truncation in the ActA protein is far more frequent in 1/2b and 4b strains (77% and 51% respectively) than in 1/2a strains (7.5%).

Loss of surface proteins in lineage III

Several genes encoding internalin-like proteins are absent in the *L. monocytogenes* 4a L99 genome in comparison to the 1/2a EGD-e and the 4b strains (Additional file 4: Table S4) as previously reported for lineage III strains [27,29]. The entire *inlGHE* cluster [30] is absent in the 4a L99 genome (Additional file 5: Table S5) [27,30]. The corresponding loci in both 4b genomes are identical to each other, but different to strain 1/2a EGD-e. Another PrfA-independent internalin (InlJ) that has been shown to be specifically expressed only in vivo [31] is also absent from the 4a L99 genome. Similarly, Internalin C [27], involved in cell-to-cell spread and innate immune response in the vertebrate host [32-35], is absent in 4a L99 but is conserved in both 4b strains and 1/2a EGD-e. A comparable situation was identified for internalin F [27], however deletion mutants have not been shown to be reduced in invasion into non-phagocytic cells [36]. Apart from the absence of these characterized internalin genes, several other internalin-like genes (*lmo1666*, *lmo2470* and *lmo2821*, Additional file 4: Table S4) are present in the 1/2a EGD-e and 4b genomes, but are absent from the 4a L99 genome. In addition, we analysed the repertoire of genes encoding surface proteins for recently published 4a genomes of strain HCC23 [37] and M7 [38] as well as 4c FSL J2-071 (*Listeria monocytogenes* Sequencing Project, Broad Institute of Harvard and MIT; http://www.broad.mit.edu) (Additional files 4: Table S4, Additional file 6: Table S6 Additional file 7: Table S7 Additional file 8: Table S8 Additional file 9: Table S9 Additional file 10: Table S10 and Additional file 11: Figure S1). We confirmed by comparative genomics that these 4a genomes lack a similar number of surface proteins (Additional files 4: Table S4, Additional file 6: Table S6 Additional file 7: Table S7 Additional file 8: Table S8 Additional file 9: Table S9 Additional file 10: Table S10 and Additional file 11: Figure S1). These findings were independently verified by additional PCR analysis to confirm the absence of genes encoding surface proteins for four 4a strains and three 4c strains, respectively. Half of the inspected chromosomal loci differed by PCR analysis among 4a and 4c genomes (Additional file 11: Figure S1). Some non-internalin like cell-wall proteins that have been shown to be important for invasion are also absent, e.g. *auto* a GW-motif containing (Additional file 6: Table S6), PrfA-independent, surface autolysin. Previous studies revealed an essential role for *auto* in the entry into non-phagocytic eukaryotic cells [39].

The *vip* gene product, a PrfA-dependent LPXTG protein (Additional file 7: Table S7), described as a receptor for the eukaryotic Gp96 surface protein and important for late stages of infection [40], is also absent from the 4a L99 genome. In addition to these missing genes, InlI is slightly truncated. However Ami (Additional file 6: Table S6), an important listerial adhesion protein seems to be present in a shorter version in both 4b strains [41,42], whereas the number of lipoproteins (Additional file 8: Table S8), LysM- and (Additional file 9: Table S9) NLPC/P60-motif containing proteins (Additional file 10: Table S10) was comparable among the four strains under study.

Overall, in comparison to 1/2a EGD-e and the two 4b genomes, 4a L99 strain has lost a number of crucial determinants required for listerial invasion. The selective loss of genes primarily responsible for the first steps of infection may contribute to the poor invasion ability and the attenuated nature of the 4a L99 strain.

Decay of phage genes in the *L. Monocytogenes* 4a L99 strain

The 1/2a EGD-e genome contains 79 prophage genes in two different loci, the 4a L99 genome includes 193 phage genes at four loci, while the 4b genomes encode with 16, for the smallest number of prophage genes limited to a single locus (also called the monocin-locus) at the same position in the chromosomes.

This monocin locus, a cryptic prophage region, is conserved in all *L. monocytogenes* lineages and includes the *lma* genes [43]. Although previously thought to be specific to *L. monocytogenes*, it was shown that *lmaDCBA* is also present in several apathogenic *L. innocua* strains. However, not all genes of the operon are present in all *L. monocytogenes* strains. The 4a L99 genome lacks *lmaA* and *lmaB* (Additional file 12: Figure S2). The entire locus in 1/2a EGD-e and the two 4b genomes has 16 genes, but only five of these genes are present in the 4a L99 genome. *lmaA* and *lmaB* are absent in *L. welshimeri*. Interestingly, the structure of this prophage locus in strain 4a L99 and other lineage III strains is more similar to *L. welshimeri* than to other pathogenic listeriae (Additional file 12: Figure S2).

The CRISPR system of *Listeria*

The *L. monocytogenes* 4a L99 genome was found to contain two adjacent CRISPR loci (I and II) with CRISPR repeats (Figure 3A and 3B). Both loci contain sequences of length 35 bp separated by repeat sequences of length 29 bp. However, they differ considerably in the number of repeat copies (6 in locus I, and 29 in locus II, respectively). While locus I is highly conserved in the 4b strains, 1/2a EGD-e and *L. innocua*, locus II was exclusively present in 4a genomes of L99, HCC23, M7, but not in another lineage III genome of 4c FSL J2-071 (Figure 3 A-C). It is not known whether the CRISPR system is functional in the 4a L99 genome. However, by sequence similarity searches using the spacers to detect possible prophage DNA traces, we were able to identify the PSA prophage that is known to infect serotype 4 strains. Assuming a functional CRISPR system in 4a L99 suggests a resistance to the PSA bacteriophage (Additional file 13: Figure S3).

Gene duplications in the *Listeria* genomes expand metabolic systems

We found substantial evidence for a minimum of 231 to a maximum of 296 gene duplications in the *Listeria* genomes (Additional file 14: Figure S4 and Additional file 15: Figure S5). It is evident that the majority of these duplications are ancient events as they are shared among all species and the number of gene pairs with a very high percentage identity is very low (1-12% per strain). Functional classification of the duplicated genes revealed that many of these have important implications in metabolic pathways, like the pentose phosphate pathway, fructose and mannose metabolism, carbon fixation, glycolysis and pyruvate metabolism.

While several duplicated genes could be mapped to central metabolic pathways from the KEGG database, this was not possible for horizontally transferred genes (Additional file 16 Figure S6 and Additional file 17: Figure S7). However, not all duplicated genes seem to have arisen from true duplications, but some may have been transferred horizontally, like some PTS system genes that are *L. monocytogenes* EGD-e strain-specific genes. The number of genes classified into known metabolic pathways or systems was significantly higher for duplicated genes, while several horizontally transferred genes could not be mapped.

Comparative intracellular transcriptomics of four *L. Monocytogenes* strains of the three major lineages

Comparative transcriptome analysis of *Listeria monocytogenes* strains of the two major lineages revealed differences in virulence, cell wall, and stress response [44]. Here we performed intracellular gene expression analyses using whole genome microarrays between four *L. monocytogenes* strains belonging to the three major lineages to investigate eventual differences. P388D1 murine macrophages were infected and total RNA was isolated four hours post infection and hybridized to bioarrays.

In order to determine the core intracellular response of *L. monocytogenes* we created a dataset of core-syntenic homologous genes for all four genomes and the expression data for these genes were compared. We found that in all strains studied the entire virulence genes cluster, (*prfA*, *plcA*, *hly*, *mpl*, *actA*, *plcB* and *orfX)* was highly

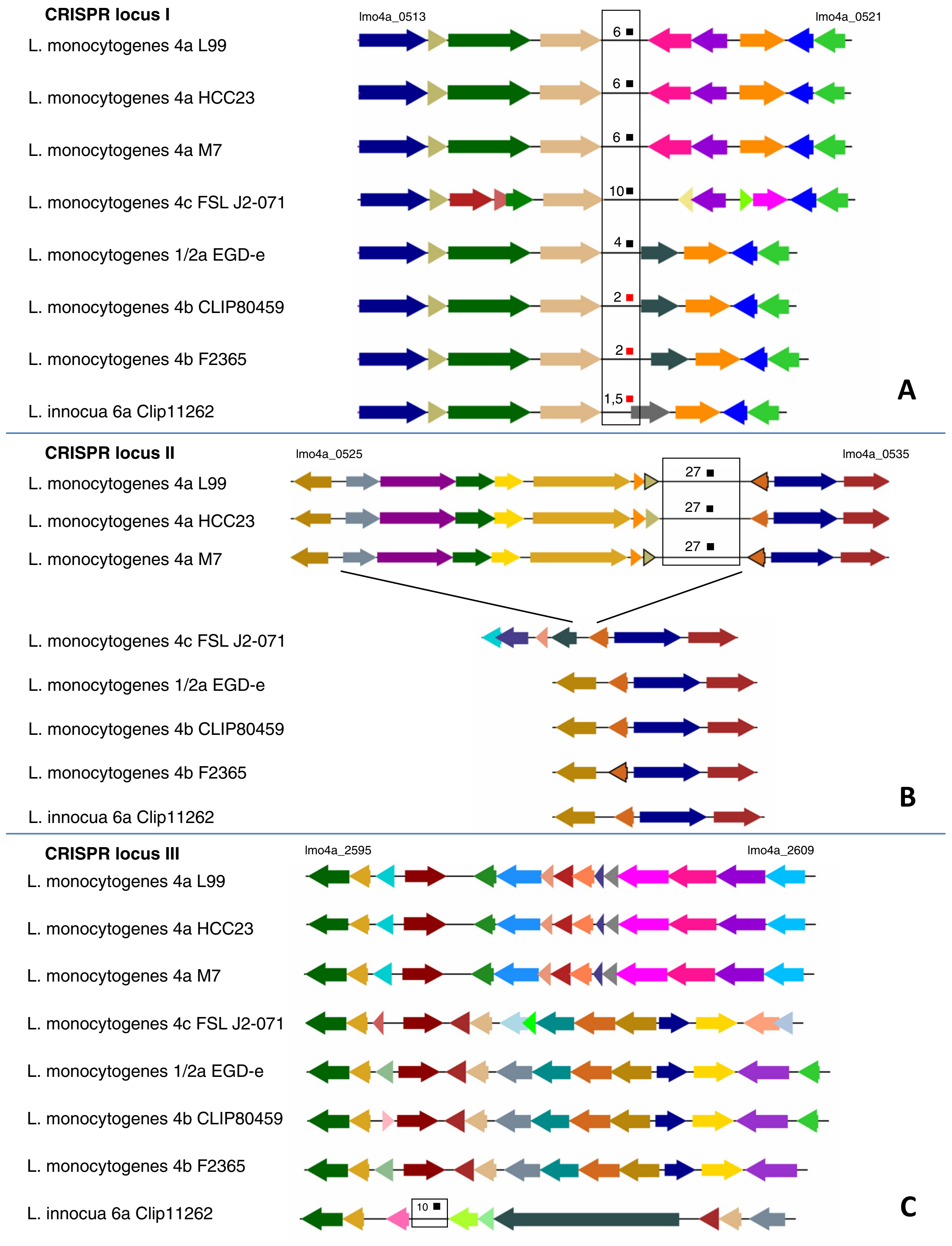

CRISPR locus I
L. monocytogenes 4a L99
L. monocytogenes 4a HCC23
L. monocytogenes 4a M7
L. monocytogenes 4c FSL J2-071
L. monocytogenes 1/2a EGD-e
L. monocytogenes 4b CLIP80459
L. monocytogenes 4b F2365
L. innocua 6a Clip11262
lmo4a_0513
lmo4a_0521
6 ■
6 ■
6 ■
10■
4 ■
2 ■
2 ■
1,5 ■
A
CRISPR locus II
L. monocytogenes 4a L99
L. monocytogenes 4a HCC23
L. monocytogenes 4a M7
L. monocytogenes 4c FSL J2-071
L. monocytogenes 1/2a EGD-e
L. monocytogenes 4b CLIP80459
L. monocytogenes 4b F2365
L. innocua 6a Clip11262
lmo4a_0525
lmo4a_0535
27 ■
27 ■
27 ■
B
CRISPR locus III
L. monocytogenes 4a L99
L. monocytogenes 4a HCC23
L. monocytogenes 4a M7
L. monocytogenes 4c FSL J2-071
L. monocytogenes 1/2a EGD-e
L. monocytogenes 4b CLIP80459
L. monocytogenes 4b F2365
L. innocua 6a Clip11262
lmo4a_2595
lmo4a_2609
10 ■
C

Figure 3 (See legend on next page.)

(See figure on previous page.)
Figure 3 Overview of CRISPR (clustered regularly interspaced short palindromic repeats) loci in *L* *monocytogenes* 1/2a EGD-e, *L. monocytogenes* 4a L99, *L. monocytogenes* 4a HCC23, *L. monocytogenes* 4a M7, *L. monocytogenes* 4c FSL J2-071, *L. monocytogenes* 4b CLIP80459, *L. monocytogenes* 4b F2365 and *L. innocua* 6a CLIP11262. (A): CRISPR locus I is shown for all five listeriae, black boxes indicate complete CRISPR repeats, red boxes represent incomplete or truncated (*) CRISPR repeats. No *cas* genes were found to be associated with this locus. Flanking genes are conserved in 1/2a EGD-e and both 4b genomes. Comparison of the intergenic sequences with the 4a L99 genome revealed a sequence footprint of decaying repeat elements (2 repeat copies in both 4b genomes, and 1 copy in *L. innocua 6a* CLIP11262), indicating loss of the CRISPR repeats. (B): Locus II shows 29 copies of repeats and is associated with several *cas* genes (*cas2*, *cas3*, *cas5* and *cas6*. *cas1* is partially detectable, but seems to be truncated. (C): *L. innocua* 6a CLIP11262 harbours the CRISPR locus III at position 2.77 Mb in the genome, which is neighboured by a single *cas2* gene. No other CRISPR repeats nor any *cas* gene homologs were found in the 4b genomes.

induced within the infected host cells. Furthermore genes known to be important for bacterial survival, such as *hpt*, *clpE*, *bilEA* and two LRR domain-containing proteins (*lmo0514* and *lmo2445*) were upregulated in all strains.

Interestingly, three mannose transporting PTS systems (*lmo0021-lmo0024*, *lmo0781-lmo0784*, *lmo1997-lmo2002*), two fructose specific systems (*lmo2335* and *lmo2733*), two galacitol specific systems (*lmo0503*, *lmo0507*, *lmo0508* and *lmo2665-lmo2667*), two beta-glucoside systems (the partial system *lmo0373-lmo0374* and *lmo0874-lmo0876*), and two cellobiose specific systems (the partial system *lmo0901* and *lmo0914-lmo0916*) were commonly upregulated in all strains. These possibly represent the most frequently used substrates of listeriae in the cytosol. Only one mannose specific PTS system, (*lmo0096-lmo0098*) is downregulated by all studied strains (Additional file 18 Figure S8 and Additional file 19: Text S1).

Most surprisingly, all *Listeria* strains studied expressed the genes of the *lma* operon and surrounding prophage genes of the monocin locus, including a conserved holin (*lmo0112*, *lmo0113*, *lmo0115*, *lmo0116*, *lmo0128*) during intracellular growth. However, the functions of several of these genes are not defined. The only locus that is conserved in all three lineages (albeit with some deletions in 4a L99) is the monocin *lma* locus. The *lmaA* gene product has been shown to provoke a delayed type hypersensitivity reaction in mice immune to *L. monocytogenes*. It is also secreted at 20°C but much less [45] at 37° C. The *lma* operon produces two transcripts, a 2100 bp *lmaDCBA* transcript expressed both at 20°C and 37°C, and a 1050 bp *lmaBA* transcript induced at lower temperatures [43]. Additional prophage genes were highly expressed in the individual strains (Figure 4). Taken together, high intracellular prophage gene expression, despite several differences in prophage gene content, is one of the most striking observations across all *Listeria* lineages.

All strains showed induction of the *eut* operon suggesting that ethanolamine may be used as a carbon and nitrogen source in intracellular conditions. The zinc transporters were also commonly upregulated indicating a role of zinc in intracellular survival as well as the spermidine/putrescine ABC transporters (*potB, potC* and *potD*). Furthermore, the non-oxidative branch of the pentose phosphate pathway was utilized by all listeriae, possibly to generate NADPH for countering oxidative stress in intracellular conditions. The upregulation of genes of the pentose phosphate pathway has been shown previously [19,20,46] and it has been speculated that it is important for generation of erythrose-4-phosphate for aromatic amino acid biosynthesis or for generation of pentose sugars. Accordingly; we observed a downregulation of several genes involved in pyrimidine and purine biosynthesis from pentose sugars (e.g. *lmo1463*, *lmo1497*, *lmo1565*, *lmo1832*, *lmo1836*, *lmo1856*, *lmo1929*, *lmo2154*, *lmo2155*, *lmo2390* and *lmo2559*).

Downregulated genes included the *agr* locus (*lmo0048-lmo0051*) as demonstrated previously [20,46] and several genes of the tryptophan biosynthesis operon (*trpA*, *trpB*, *trpF* and *trpD*), and some tRNA synthetase genes (*ileS*, *valS*, *glyS* and *glyQ*). Diminished energy generation was indicated by decreased expression of the cytochrome genes cluster *cytABCD*. With respect to the pentose phosphate pathway, we detected downregulation of the phosphoribosyl pyrophosphate synthetase (*prs, lmo0199*) gene, which is required for the production of PRPP (phosphoribosyl pyrophosphate) that links the pentose phosphate pathway to the biosynthesis of purines and pyrimidines. While several genes of the glycolytic operon, and several individual genes were downregulated by 1/2a EGD-e, 4b CLIP80459 strain or 4a L99, the 4b F2365 strain showed increased expression (Additional file 20: Text S2).

Differences in flagellin expression are the most prominent differences among strains

To address the observation that strain 4b CLIP80459 grows more efficiently inside the host than strain 4b F2365, we performed a direct comparison of the transcriptome data derived from these two strains. Most important differences were found in the regulation of flagellar genes. While intracellular bacteria of strain 4b F2365 upregulated a substantial number of flagellar genes, including *fliS*, *fliI*, *flhA*, *fliF*, *filE*, *flgB*, *flgC*, *flgG*, *fliD* as well as the transcriptional regulator *degU* (*lmo2515*), in the 4b strain CLIP80459 only *fliR* was upregulated. When comparing the intracellular transcriptome of strain 4a L99 to the 1/2a and 4b strains the

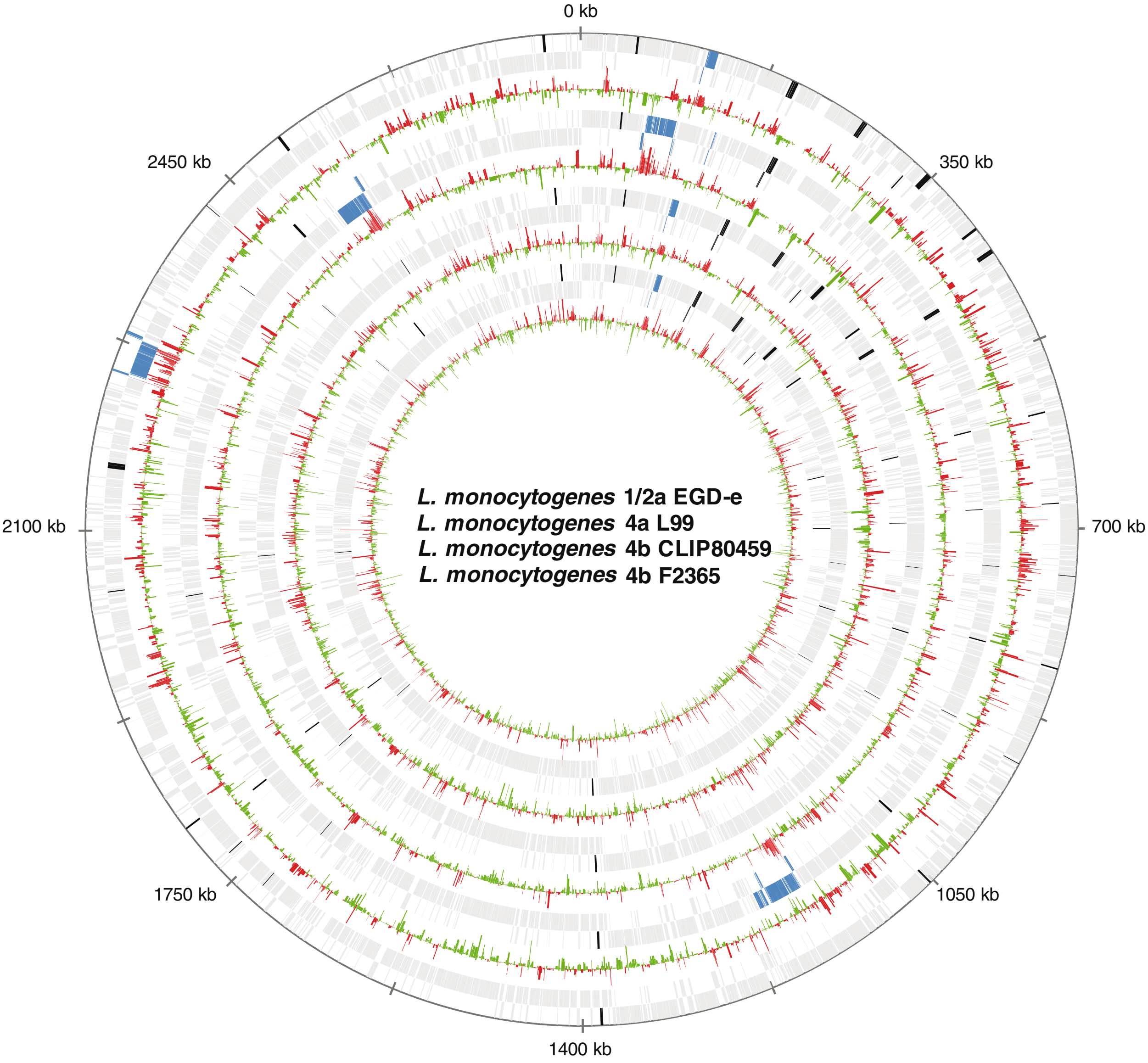


Figure 4 Comparative transcriptomics of four *L. monocytogenes* genomes: *L. monocytogenes* 1/2a EGD-e, *L. monocytogenes* 4a L99, *L. monocytogenes* 4b CLIP80459, *L. monocytogenes* 4b F2365 (from outside to inside). There are two tracks per strain: the first one shows the coding sequences (gray), phage genes (blue) and virulence genes (black). The second one visualizes increase (red) or decrease (green) of intracellular gene expression (log fold changes). Phage and virulence genes are clearly upregulated intracellularly. Data were illustrated using GenomeViz [26].

most striking difference was again the expression of the flagellar operon. We observed a strong induction of nearly all flagellar genes in the operon, including flagellin (Additional file 21: Text S3) (homologues of *lmo0675*, *lmo0676*, *lmo0681*, *lmo0685*, *lmo0686*, *lmo0690*-*lmo0696*, *lmo0698*-*lmo0701*, *lmo0703*-*lmo0706*, *lmo0708*, *lmo0709*, *lmo0712*, *lmo0714* and *lmo0715*) in strain 4a L99. Strong expression of these genes is counterproductive within infected cells, because it probably enables the host to efficiently detect bacterial presence and the formation of an inflammasome.

Apart from genes that are important for pathogen recognition mechanisms by the host, a concerted expression profile (Additional file 22: Figure S9) involving genes of cell wall synthesis, host cell invasion, response to oxidative stress, utilization of host carbohydrates and propanediol, which are crucial for intracellular survival as well as virulence and surface proteins were identified.

Differential growth of the three lineages and Δ*lmaB* and Δ*lmaD* isogenic mutants in a mouse infection and cell infection models

We observed a severe deficiency in entry of strain L99 in HeLa and Caco-2 cells as well as poor cell-to-cell transmission with macrophages and L929 fibroblasts when compared to 1/2a EGD-e (data not shown). Impaired invasion ability of host cells may be due to lack of several internalin genes in the genome of strain 4a L99. It is likely that both, decreased invasive ability and strong intracellular expression of flagellar genes contribute towards the rapid clearance of the 4a L99 strain in in vivo experiments in mice. Upregulation of several DNA repair genes was also seen in strain 4a L99 compared to the other strains, e.g. (*recF*, *recN*, *radA* and *mutL*), suggesting genomic damage during the infection process.

To further assess the virulence potential of the three lineages, we performed mouse infection experiments with each of the four strains (1600 cfu/mouse), and measured bacterial loads in spleens and livers at different time points (Figure 5A and 5B). The 4a L99 strain was cleared rapidly from the mice and was not detectable after five days of infection, in accordance with previous results [18], indicating that the 4a L99 strain is attenuated in its pathogenicity. However, the other three strains were able to survive in both spleens and livers of infected mice. Interestingly, while they could comparably replicate in the spleen, the 1/2a EGD-e and the 4b F2365

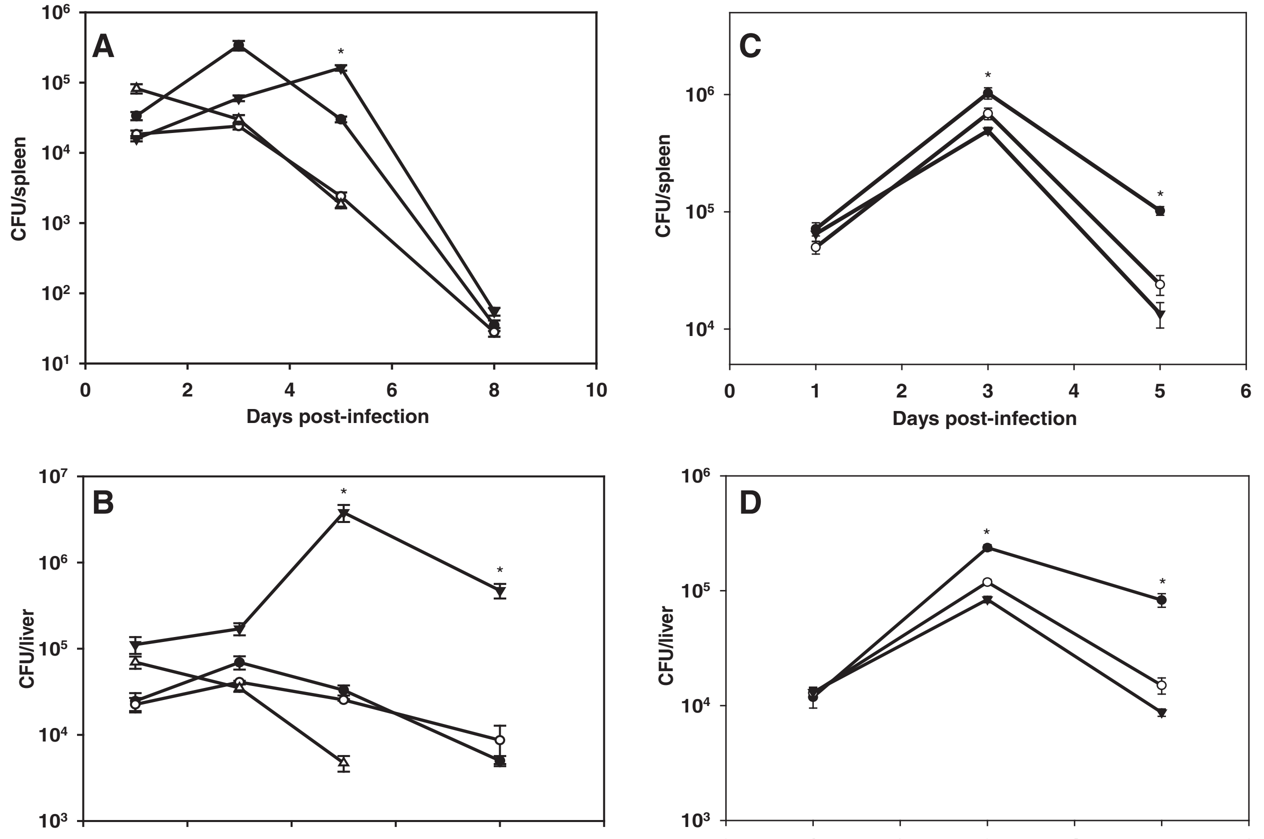


Figure 5 Murine infection studies with three different *Listeria* serotypes and two chromosomal deletion mutants of Δ*lmaB* and Δ*lmaD* of *L. monocytogenes* 1/2a EGD-e Mice were infected i.v. with 2000 cfu of *L. monocytogenes* serotypes 1/2a EGD-e (filled circles), 4b F2365 (open circles), 4b CLIP80459 (filled triangles), and 4a L99 (open triangles). On days 1, 3, 5, and 8 after infection, the numbers of viable bacteria in spleens (A) and livers (B) of three animals per group were determined (P ≤0,05 and P ≤0,001 of 4b CLIP80459 vs. 1/2a EGD-e and 4b F2365 vs. 1/2a EGD-e in spleen and liver respectively). Bacterial load in mice organs were also determined following i.v. infection with 2000 cfu of *L. monocytogenes* 1/2a EGD-e wild type strain (filled circles) as well as its isogenic mutants Δ*lmaB* (open circles), and Δ*lmaD* (filled triangles). On days 1, 3, and 5 after infection, the numbers of viable bacteria in spleens (C) and livers (D) of three animals per group were determined (P ≤0,05 and P ≤0,01 of 1/2a EGD-e versus Δ*lmaB* and Δ*lmaD* in spleen and liver respectively). Data presented are representative of three independent experiments. An asterisk indicates means that are significantly different from the wild type. Significance analysis was performed with student *t*-test.

bacterial loads in liver were significantly lower than the 4b CLIP80459 strain whose counts remained significantly higher even on days five and eight post-infection. Isogenic mutants of Δ*lmaB* and Δ*lmaD* showed similar counts in mice spleens and livers. However, both mutants have shown a significantly lower level of growth than 1/2a EGD-e on days 3 and 5 post-infection (Figure 5C and 5D).

Discussion

We sequenced and analysed the genomes of representatives of three major lineages of species *L. monocytogenes* to correlate gene content with (i) its wide spectrum of pathogenic abilities, (ii) its differing properties for survival in the hosts, and (iii) its adaptive properties during growth under extracellular conditions.

Decay of surface proteins in the virulence attenuated *L. Monocytogenes* 4a strain

Analysis of the 4a L99 genome revealed extensive loss of a large number of internalins, internalin-like proteins and other surface proteins important for invasive ability. For strain 4a L99, which was isolated from contaminated food in the 1950's, it might be possible that mutations have taken place over this lengthy time of storage under in vitro conditions. Surprisingly, a previously known *act*A truncation in the 4a genomes of L99, HCC23 and M7, was also found in a higher number of lineages I strains compared to lineage II, but not in the *act*A gene of another lineage III strain of 4c FSL J2-071 indicating a serotype-specific heterogeneity of ActA sequences within the genus *Listeria*. The loss of this proline-repeat in ActA is correlated with lowered actin-based motility in the cytosol. In addition, comparative nucleotide analysis indicated that the latter half of the LIPI-I pathogenicity island in strain 4a L99 has diverged significantly from that of the 4b and 1/2a strain leading to a loss of the open reading frames *lmo0206* to *lmo0209*. Loss of *lmo0206* (*orfX*) has been shown to confer a severe growth effect on survival in macrophages, [20] while loss of *lmo0207* has a small effect on growth in macrophages and no data are presently available for *lmo0208* and *lmo0209* and their role in virulence.

Differential regulation of intracellular flagella gene expression by strains of different lineages

Highly sensitive and widely distributed host microbe-associated microbial pattern receptors (TLRs and NLRs) continuously patrol the cell surface, endosomes and the cytosol for signs of microbial presence by sensing cell wall components, bacterial DNA, lipoproteins and flagellin. Ligands may be shared between the surface and the cytosolic receptors, e.g. cell wall components and flagellin may be sensed both by TLRs and also by cytosolic receptors. We detected the intracellular expression of the flagellin gene in 1/2a EGD-e [20]. Recently, it has been shown that cytosolic flagellin, expressed by *L. monocytogenes* strain 10403 S (serotype 1/2a) is detected by multiple Nod-like receptors, including IPAF and NALP3, and also by a pathway involving the adaptor protein ASC and the cytosolic DNA sensor AIM2, which is required for the formation of the inflammasome [47-49]. Detection of flagellin in the cytosol via these pathways leads to caspase-1 mediated cleavage of pro-IL-1B and release of active IL-1B. Mice lacking caspase-1 or ASC are unable to mount active IL-1B response to intracellular pathogens such as *Shigella flexneri* and *Francisella tularensis* [50,51]. All strains investigated in this study were found to express flagellar genes in the cytosol, except for strain 4b CLIP80459. The ability to successfully downregulate flagellar (*flaA*) gene expression is probably critical for evading host detection and promoting bacterial intracellular growth. In line with this observation, a 1/2a EGD-e chromosomal deletion mutant of the gene displayed increased survival in mouse infection assays [52].

In keeping with this finding, both strains 4b F2365 and 4a L99 displayed strong induction of several flagellar genes during intracellular growth and were more readily cleared from the host. This suggests strain-specific differences in the ability to avoid host recognition can lead to large differences in virulence manifestation, despite several commonalities in the adaptations of the lineages to the intracellular lifestyle. Although all the strains investigated in this study were able to induce all genes of the virulence genes cluster intracellularly, it is likely that there are a multitude of effects including differences in virulence gene expression, uptake of carbohydrates, membrane protein expression and flagellar biosynthesis, all of which contribute to the observed phenotypic properties.

Effects of gene duplication events on metabolic adaptation and survival within the host

The processes of gene duplications, horizontal gene transfer and gene loss influence the short- and long-term evolution of prokaryotic genomes. The benefits of gene duplications in the short term can be seen clearly in conditions of antibiotic treatment [53,54], toxin exposure [55], heavy metal stress [56,57], extreme temperatures [58], nutrient limitation [59,60] and even parasitic and symbiotic lifestyles [54,61]. Duplications found in all *Listeria* genomes seem to have been ancient i.e. precede species differentiation, with only the exception of the recent prophage duplication in *L. innocua* 6a CLIP11262. Classification of duplicated genes revealed several paralogous genes in metabolic pathways, while very few horizontally transferred genes could be classified at all.

The highest numbers of gene duplications were identified in the following categories: ABC transporters, PTS systems, pentose phosphate pathway, starch and sucrose metabolism, fructose and mannose metabolism, and carbon fixation. Surprisingly, we found a high number of duplicated gene paralogues involved in the regulation of the non-oxidative branch of the pentose phosphate pathway and in the generation of ribose-5-phospate from ribulose-5-phosphate. Under conditions of intracellular growth, we observed differences in the ability of the lineages to express horizontally transferred genes. 1/2a EGD-e was most successful in this regard (17 genes), followed by 4a L99 (10 genes), 4b F2365 (6 genes) and 4b CLIP80459 (2 genes). Apart from the horizontally transferred genes, differences in the expression of strain-specific genes in the cytosol were apparent (1/2a EGD-e: 45; 4a L99: 49; 4b F2365 11; 4b CLIP80459: 3).

PTS systems enable listeriae to utilize host carbohydrates, a mechanism that is essential for the intracellular survival. PTS systems (EII) for the utilization of fructose and beta-glucosides, mannose and cellobiose were most frequently observed in the investigated *Listeria* genomes. Although the numbers of PTS systems are comparable among the investigated genomes (Additional file 18: Figure S8), even a slight difference in presence/absence of a PTS system available as an additional carbohydrate utilization mechanism may have dramatic effects on listerial survival inside the host cytosol [61-63], specifically on the master regulator PrfA [61,62,64,65]. For instance, the pentitol PTS system in 1/2a EGD-e is not present in either the 4b or the 4a L99 genomes. A transposon insertion mutant of this system (*lmo1971*) has been shown to have significantly attenuated growth in epithelial cells [46]. Several partial PTS systems are also present in the genome (Additional file 19: Text S1). These are independently expressed intracellularly, and represent broadly shared and commonly regulated systems. In accordance, the pathogenic strain 4b CLIP80459 was found to upregulate more PTS systems than strain 4b F2365, which may contribute to better intracellular survival of 4b CLIP80459.

In addition to phosphorylated sugars, there are other nitrogen and carbon sources available to intracellular bacteria, such as ethanolamine. Ethanolamine is used as substrate and an energy supply by *Salmonella enterica* grown under anaerobic conditions and is suggested to be used by other bacteria [66]. A locus homologous to that of the ethanolamine operon of *S. enterica* has also been described in *Listeria* [67]. The gene organization of the locus is not identical to the *Salmonella* cluster, but all the genes of the cluster have homologous sequences in *Listeria* (Additional file 23: Figure S10). Previous studies identified genes of the locus to be upregulated intracellularly during infection and were shown to play a critical role for intracellular survival [46]. Our data support this observation and further demonstrate upregulation of several genes of this locus across all three pathogenic lineages of *Listeria*, suggesting that the functions of the locus are conserved. However, since the locus is also present in the apathogenic *L. innocua* strain 6a CLIP11262, it may exemplify a general requirement of *Listeria* to cope with nutrient rather than a specific virulence adaptation. Furthermore, degradation of the phagosomal membrane that traps intracellular listeriae, results in the release of ethanolamine as a byproduct and may serve an energy source in the host cytosol.

Not only the efficient recruitment of carbohydrate substrates, but also the differential channeling through different pathways represents an important adaption within the host cytosol. It has been shown that an essential mechanism to counteract oxidative stress is to reroute carbohydrate flux via the pentose phosphate pathway, which is required for the biosynthesis of reductive substrates rather than through glycolysis pathway [68]. Indeed, we observed that all lineages prefer to channel carbohydrate flux via the pentose phosphate pathway, rather than glycolysis. In contrast to the other strains, only strain 4b F2365 was unable to downregulate glycolysis, suggesting that the inability to route sugars efficiently via pentose phosphate contributes to the poor intracellular growth of this strain.

The CRISPR system in *Listeria* reveals expansion and atrophy

A CRISPR (Clustered, regularly interspaced short palindromic repeats) locus, associated with several *cas* genes was identified in the 4a L99 genome. CRISPRs are highly divergent loci found in genomes of all archaea and several bacteria [69]. A CRISPR system is composed of the *cas* (CRISPR-associated) genes, a leader sequence and arrays of direct repeats separated by non-repetitive spacer sequences resulting in a RNA-interference like innate phage-resistance mechanism [70]. A recent study in *Streptococcus thermophilus* demonstrated how bacteria are able to integrate new spacer sequences derived from infecting phages, directly into the CRISPR arrays, and that this ability confers phage-resistance [71]. The mechanism of resistance has also been elucidated [70]. Among the genomes compared in this study, only the 4a L99 genomes of L99, HCC23 and M7 possesses *cas* genes and several CRISPR repeats. There are only two repeats in each 4b genome, five in 1/2a EGD-e a single one in *L. innocua* 6a CLIP11262, but none of these strains harbour identifiable *cas* genes. In addition, a small sRNA *rliB* is located in the repeat region of 1/2a EGD-e and contributes to virulence in mice [72]. We were also able to detect a DNA sequence of a potential prophage (PSA) using the spacers from the 4a genome.

As prophages evolve quite rapidly, it is likely that this acquisition is a recent event.

Distinct role of intracellularly upregulated phage genes in virulence of listerial strains

The four *L. monocytogenes* strains have different numbers of prophage genes (1/2a EGD-e: 79; 4a L99: 191; 4b CLIP80459: 16 and 4b F2365: 16) distributed in different loci. Regardless of location and lineage, all strains expressed several prophage genes within the infected host cell. However, only a single locus, the *lma* locus is conserved across the three lineages and is also induced during infection. The role of prophage genes in the virulence of *Listeria* has not been examined in detail. We show that chromosomal deletion mutants of two genes in this locus (*lmaB* and *lmaD*) resulted in growth reduction of 1/2a EGD-e in a murine infection model. Although the underlying mechanisms leading to the attenuated phenotypes remain unclear, a recent study revealed that prophage diversification represents an essential mechanism for short-term genome evolution within the species *L. monocytogenes* [73,74] and is subject of further investigation.

Conclusion

Listeria monocytogenes is arguably one of the best characterized pathogens and has been established as an unparalleled model microorganism in infection biology. Detailed understanding of differences in virulence of the three major lineages of *Listeria* provides us with invaluable information about evolutionary adaptation of this pathogen. Here we used comparative genomics and whole-genome based transcriptome analysis of strains from all lineages to obtain a comprehensive view as to how these strains have evolutionarily diverged. This approach suggests that (i) reductive evolution of strains of serotype 4a such as L99, HCC23 and M7 is the major force driving the attenuated phenotype, (ii) acquisition and adaptation of prophage genes and metabolic systems, respectively, identify novel virulence-associated factors of listeriae and (iii) listeriae avoid detection and subsequent immune response of the host via downregulation of surface structures and by differences in intracellular expression of flagellar genes.

Methods

Strains and growth conditions

Four *L. monocytogenes* strains were used in the study, *L. monocytogenes* 1/2a EGD-e [14], *L. monocytogenes* 4a L99 [18], *L. monocytogenes* 4b CLIP80459 [17], *L. monocytogenes* 4b F2365 [15] and chromosomal deletion mutants of *L. monocytogenes* 1/2a EGD-e Δ*lmaB* and Δ*lmaD*. Bacteria were grown in brain heart infusion (BHI) broth (Difco) at 37°C with shaking. For further comparative genomic analysis *L. monocytogenes* 4a HCC23 [37] *L. monocytogenes* 4a M7 [38] and *L. monocytogenes* 4c FSL J2-071) (*Listeria monocytogenes* Sequencing Project, Broad Institute of Harvard and MIT; http://www.broad.mit.edu) was used.

Genome sequencing and annotation

In brief, genome sequencing *L. monocytogenes* 4a L99 was performed on ABI PRISM 3100 or 3730xl Genetic Analyzers (Applied Biosystems). Whole genome shotgun sequencing was performed by LGC (Berlin, Germany). Sequence data were analysed and assembled using Phred/Phrap/Consed [75,76]. A total number of 27,637 sequences of shotgun libraries, 1684 fosmid and 671 PCR gap closure sequences were assembled by the Phrap software resulting in a ~6.7-fold coverage. Genome annotation was performed as previously described [3].

Genome sequencing of *L. monocytogenes* 4b CLIP80459 was performed using the conventional whole genome shotgun strategy [77,78]. One library (2–3 kb inserts) was generated by random mechanical shearing of genomic DNA and cloning into pcDNA-2.1 (Life technologies) and recombinant plasmids were used as templates for cycle sequencing reactions. Samples were loaded on capillary automatic 3700 and 3730 DNA sequencers (Applied Biosystems). In an initial step 35,610 sequences were assembled into 361 contigs using the Phred/Phrap/Consed software [75,76]. CAAT-Box [79] was used to predict links between contigs. 379 PCR products amplified from *L. monocytogenes* CLIP80459 chromosomal DNA as template were used to fill gaps and to re-sequence low quality regions. Final assembly resulted in a ~7.8-fold coverage. Genome annotation was performed as previously described [14].

Alignment of the virulence gene cluster

The alignment was performed using MAVID [80] after extracting the virulence gene cluster of all genomes. The plot was created using VISTA [81].

ActA repeat analysis

Available ActA protein sequences for all *L. monocytogenes* strains were retrieved from GenBank (http://www.ncbi.nlm.nih.gov/Genbank/). Only sequences that contained at least 500 amino acids (reference strain 1/2a EGD-e ActA: 639 amino acids) were downloaded (774 sequences). It was possible to assign a lineage to only 386 ActA sequences. Duplicates with identical length, strain and sequence were also removed, leaving a total of 218 sequences for the analysis. These were aligned using ClustalW and the alignment of repeat regions was examined manually.

Single nucleotide polymorphisms

Single nucleotide Polymorphisms (SNPs) were detected by the MUMmer [25] and SNPs were mapped to coding regions using PERL scripts. The SNP-density per gene normalized by gene length was calculated and the data were visualized in GenomeViz [26].

CRISPR repeats analysis

Comparative visualization of the CRISPR related genome loci was performed by GECO [82]. CRISPR repeats were identified using the PILER-CR software [83]. Subsequent analysis and visualization of repeat footprints was performed using BLAST and ACT [84].

Horizontal gene transfer and gene duplications

Horizontally transferred genes were detected using SIGI [85] and SIGI-HMM [86]. Duplicated genes were identified using BLAST cut-offs of at least 40% identity and 80% coverage considering both sequences.

Cell culture and infection model

All cell culture experiments were performed as described by Chatterjee and colleges [20].

Microarrays

For each of the four strains of the study, a genome-wide custom microarray chip was designed and implemented using the Geniom One platform from Febit Biomed GmbH, Germany. All transcriptome studies were performed with this platform. Complete details of the protocols are provided in the ArrayExpress database (http://www.ebi.ac.uk/microarray-as/ae/). Data were background corrected and then normalized using quantile normalization [87]. Pearson's correlation coefficients were used to assess reproducibility within at least two technical and three biological replicates ($r^2 > =0.94$ in all cases). The significance analysis of microarrays (SAM) program was used to analyze the data [88] as an unpaired response.

Construction of the deletion mutants Δ*lmaB* and Δ*lmaD*

Chromosomal in frame deletion mutants of *L. monocytogenes* 1/2a EGD-e Δ*lmaB* and Δ*lmaD* were constructed by generating the 5′ (with primers P1 and P2) and the 3′ (with primers P3 and P4) flanking region of the gene concerned. Primers used to generate the flanking regions are shown (Additional file 24: Table S11). The purified PCR fragments of 5′ and 3′ flanking regions were amplified using primer P1 and P4, ligated into pCRII (Life technologies) and transformed into *E. coli* InvαF' electrocompetent cells (Life technologies). Subsequently, the vector was digested with restriction enzyme *Eco*RI and ligated into the temperature sensitive suicide vector pAUL-A which was digested with the same enzymes and transformed into *E. coli* InvαF' electrocompetent cells. Plasmid DNA of pAUL-A bearing the fragment was isolated from the recombinants and used to transform *L. monocytogenes* EGD-e to generate the chromosomal deletion mutants as described in detail by Schaeferkordt et al. [89]. The deletion in the gene concerned was identified by PCR and confirmed by sequencing the PCR fragment using primers P1 and P4.

Murine infection assay

Primary infection with *L. monocytogenes* serotypes and mutants was performed by intravenous injection of viable bacteria in a volume of 0.2 ml of PBS. Bacterial growth in spleens and livers was determined by plating 10-fold serial dilutions of organ homogenates on BHI after several days. The detection limit of this procedure was 10^2 CFU per organ. Colonies were counted after 24 h of incubation at 37°C. Six- to eight-week-old female BALB/c mice, purchased from Harlan Winkelmann (Borchen, Germany), were used in all experiments.

Ethics statement

This study was carried out in strict accordance with the regulation of the National Protection Animal Act (§7-9a Tierschutzgesetz). The protocol was approved by the local Committee on the Ethics of Animal Experiments (Regierungsbezirk Mittelhessen) and permission was given by the local authority (Regierungspraesidium Giessen, Permit Number: GI 15/5-Nr.63/2007).

Statistical data analysis of infection experiments

All infection experiments were performed a minimum of three times. Significant differences between two values were compared with a paired Student's *t*-test. Values were considered significantly different when the *p* value was less than 0.05 ($p < 0.05$).

Nucleotide sequence and microarray accession number

The genome sequences have been deposited in the EMBL database with accession numbers FM211688 for *L. monocytogenes* 4a L99 and FM242711 for *L. monocytogenes* 4b CLIP80459 respectively. The microarray data have been submitted to ArrayExpress with the accession number E-MEXP-1947.

Additional files

Additional file 1: Table S1. Nucleotide analysis of *actA* repeats of *Listeria.*

Additionaf file 2: Table S2. Prediction of LRR region containing proteins by Augur [90].

Additional file 3: Table S2. x Prediction of proteins containing GW modules by Augur [90].

Additional file 4: Table S4. Prediction of LPXTG motif harbouring proteins by Augur [90].

Additional file 5: Table S5. Prediction of lipoproteins by Augur [90].

Additional file 6: Table S6. Prediction of LysM domain containing proteins by Augur [90].

Additional file 7: Table S7. Prediction of NLPC/P60 domain containing proteins by Augur [90].

Additional file 8: Table S8. Comparative CRISPR analysis table.

Additional file 9: Table S9. Metabolosome of *L. monocytogenes* 1/2a EGD-e.

Additional file 10: Table S10. Primers used in this study.

Additional file 11: Figure S1. Comparative analysis of *L. monocytogenes* ActA protein sequences.

Additional file 12: Figure S2. Comparison of the *inlGHE* locus in the three listerial lineages. All three genes in this cluster have been absent in the *L. monocytogenes* 4a L99 genome.

Additional file 13: Figure S3. Genome analysis of *lmaDCBA* region of six listeriae. Comparative analysis was performed using GECO [82] applying bidirectional pairs.

Additional file 14: Figure S4. Frequency of distributions of the percentage identity between all duplicated gene pairs in the *Listeria* genomes.

Additional file 15: Figure S5. Gene duplication and horizontal gene transfer in *Listeria* genomes.

Additional file 16: Figure S6. Duplication vs. HGT classifiable genes in listeriae.

Additional file 17: Figure S7. Complete PTS Systems in *L. monocytogenes* strains.

Additional file 18: Figure S8. Partial PTS Systems in *L. monocytogenes* strains.

Additional file 19: Text S1. SNP analysis of three listerial lineages.

Additional file 20: Text S2. Differential regulation of glycolysis in *L. monocytogenes* 4b F2365.

Additional file 21: Text S3. Comparison of two *L. monocytogenes* 4b strains CLIP80459 and F2365.

Additional file 22: Figure S9. Confirmation of lacking genes encoding surface proteins in four *L. monocytogenes* 4a strains and three *L. monocytogenes* 4c strain generated by PCR analysis.

Additional file 23: Figure S10. Intracellular flagellin expression data of *L. monocytogenes* 1/2a EGD-e, *L. monocytogenes* 4a L99, *L. monocytogenes* 4b CLIP80459 and *L. monocytogenes* 4b F2365 generated by qRT-PCR analysis.

Additional file 24: Figure S11. List of gene duplication in *Listeria* genomes.

Competing interests

The authors declare that they have no competing interests.

Acknowledgments

We thank Alexandra Amend, Claudia Zörb and Nelli Schklarenko for excellent technical assistance. This work was supported by the fund obtained from the BMBF through the Competence Network PathoGenoMik (031U213B) to T.H. and T.C and through the ERA-NET Pathogenomics Network (62080061) to T.H.

Author details

[1]Institute of Medical Microbiology, Justus-Liebig-University, Schubertstrasse 81, Giessen, D-35392, Germany. [2]Helmholtz-Zentrum für Infektionsforschung GmbH, Inhoffenstraße 7, Braunschweig, 38124, Germany. [3]Center for Biotechnology, University of Bielefeld, Bielefeld, D-33594, Germany. [4]Institut für Biophysik und physikalische Biochemie, Universität Regensburg, Universitätstrasse 31, Regensburg, D-93053, Germany. [5]Febit Biomed GmbH, Im Neuenheimer Feld 519, Heidelberg, D-69120, Germany. [6]Institut Pasteur, Laboratoire Evolution et Génomique Bactériennes and CNRS URA 2171, Paris, 75724, France. [7]Institut Pasteur, Biologie des Bactéries Intracellulaires and CNRS URA 2171, Paris, 75724, France. [8]Max von Pettenkofer-Institut for Hygiene and Medical Microbiology, Ludwig Maximilians-University München, Pettenkoferstrasse 9a, Munich, D-80336, Germany.

Author' contributions

TH and TC designed the study. CB, PG, CS and TH performed genome sequencing, MAM isolated total RNA. SV performed microarray samples preparation and hybridization. TH, AB, CTK UK, ED, BI, RG, SS, PG, CR, CB, AG, SO and WM performed the genome annotation and analysis work. RG, AB, CTK, BI, TH and TC drafted and wrote the manuscript. All authors contributed to and approved the final manuscript.

Received: 25 August 2011 Accepted: 12 April 2012
Published: 24 April 2012

References

1. Hain T, Chatterjee SS, Ghai R, Kuenne CT, Billion A, Steinweg C, Domann E, Karst U, Jansch L, Wehland J, Eisenreich W, Bacher A, Joseph B, Schar J, Kreft J, Klumpp J, Loessner MJ, Dorscht J, Neuhaus K, Fuchs TM, Scherer S, Doumith M, Jacquet C, Martin P, Cossart P, Rusniock C, Glaser P, Buchrieser C, Goebel W, Chakraborty T: **Pathogenomics of *Listeria* spp.** *Int J Med Microbiol* 2007, **297:**541–557.
2. Vazquez-Boland JA, Kuhn M, Berche P, Chakraborty T, Dominguez-Bernal G, Goebel W, Gonzalez-Zorn B, Wehland J, Kreft J: ***Listeria* pathogenesis and molecular virulence determinants.** *Clin Microbiol Rev* 2001, **14:**584–640.
3. Hain T, Steinweg C, Kuenne CT, Billion A, Ghai R, Chatterjee SS, Domann E, Karst U, Goesmann A, Bekel T, Bartels D, Kaiser O, Meyer F, Puhler A, Weisshaar B, Wehland J, Liang C, Dandekar T, Lampidis R, Kreft J, Goebel W, Chakraborty T: **Whole-genome sequence of *Listeria welshimeri* reveals common steps in genome reduction with *Listeria innocua* as compared to *Listeria monocytogenes*.** *J Bacteriol* 2006, **188:**7405–7415.
4. Cossart P, Vicente MF, Mengaud J, Baquero F, Perez-Diaz JC, Berche P: **Listeriolysin O is essential for virulence of *Listeria monocytogenes*: direct evidence obtained by gene complementation.** *Infect Immun* 1989, **57:**3629–3636.
5. Cossart P, Toledo-Arana A: ***Listeria monocytogenes*, a unique model in infection biology: an overview.** *Microbes Infect* 2008, **10:**1041–1050.
6. Gaillard JL, Berche P, Frehel C, Gouin E, Cossart P: **Entry of *L. monocytogenes* into cells is mediated by internalin, a repeat protein reminiscent of surface antigens from gram-positive cocci.** *Cell* 1991, **65:**1127–1141.
7. Lecuit M, Ohayon H, Braun L, Mengaud J, Cossart P: **Internalin of *Listeria monocytogenes* with an intact leucine-rich repeat region is sufficient to promote internalization.** *Infect Immun* 1997, **65:**5309–5319.
8. Dominguez-Bernal G, Muller-Altrock S, Gonzalez-Zorn B, Scortti M, Herrmann P, Monzo HJ, Lacharme L, Kreft J, Vazquez-Boland JA: **A spontaneous genomic deletion in *Listeria ivanovii* identifies LIPI-2, a species-specific pathogenicity island encoding sphingomyelinase and numerous internalins.** *Mol Microbiol* 2006, **59:**415–432.
9. Cotter PD, Draper LA, Lawton EM, Daly KM, Groeger DS, Casey PG, Ross RP, Hill C: **Listeriolysin S, a novel peptide haemolysin associated with a subset of lineage I Listeria monocytogenes.** *PLoS Pathog* 2008, **4:**e1000144.
10. Orsi RH, den Bakker HC, Wiedmann M: ***Listeria monocytogenes* lineages: Genomics, evolution, ecology, and phenotypic characteristics.** *Int J Med Microbiol* 2011, **301:**79–96.
11. Ward TJ, Ducey TF, Usgaard T, Dunn KA, Bielawski JP: **Multilocus genotyping assays for single nucleotide polymorphism-based subtyping of *Listeria monocytogenes* isolates.** *Appl Environ Microbiol* 2008, **74:**7629–7642.
12. Jacquet C, Doumith M, Gordon JI, Martin PM, Cossart P, Lecuit M: **A molecular marker for evaluating the pathogenic potential of foodborne *Listeria monocytogenes*.** *J Infect Dis* 2004, **189:**2094–2100.
13. Murray EGD, Webb RA, Swann HBR: **A disease of rabbits characterized by a large mononuclear leucocytosis caused by a hitherto undescribed bacillus *Bacterium monocytogenes* (n.sp.).** *J Pathol Bacteriol* 1926, **29:**407–439.
14. Glaser P, Frangeul L, Buchrieser C, Rusniok C, Amend A, Baquero F, Berche P, Bloecker H, Brandt P, Chakraborty T, Charbit A, Chetouani F, Couve E, de Daruvar A, Dehoux P, Domann E, Dominguez-Bernal G, Duchaud E, Durant L, Dussurget O, Entian KD, Fsihi H, Garcia-del Portillo F, Garrido P, Gautier L, Goebel W, Gomez-Lopez N, Hain T, Hauf J, Jackson D: **Comparative genomics of *Listeria* species.** *Science* 2001, **294:**849–852.
15. Nelson KE, Fouts DE, Mongodin EF, Ravel J, DeBoy RT, Kolonay JF, Rasko DA, Angiuoli SV, Gill SR, Paulsen IT, Peterson J, White O, Nelson WC, Nierman W,

Beanan MJ, Brinkac LM, Daugherty SC, Dodson RJ, Durkin AS, Madupu R, Haft DH, Selengut J, Van Aken S, Khouri H, Fedorova N, Forberger H, Tran B, Kathariou S, Wonderling LD, Uhlich GA, Bayles DO, Luchansky JB, Fraser CM: **Whole genome comparisons of serotype 4b and 1/2a strains of the food-borne pathogen *Listeria monocytogenes* reveal new insights into the core genome components of this species.** *Nucleic Acids Res* 2004, **32:**2386–2395.

16. Nightingale KK, Milillo SR, Ivy RA, Ho AJ, Oliver HF, Wiedmann M: ***Listeria monocytogenes* F2365 carries several authentic mutations potentially leading to truncated gene products, including *inlB*, and demonstrates atypical phenotypic characteristics.** *J Food Prot* 2007, **70:**482–488.
17. de Valk H, Vaillant V, Jacquet C, Rocourt J, Le Querrec F, Stainer F, Quelquejeu N, Pierre O, Pierre V, Desenclos JC, Goulet V: **Two consecutive nationwide outbreaks of Listeriosis in France, October 1999-February 2000.** *Am J Epidemiol* 2001, **154:**944–950.
18. Chakraborty T, Ebel F, Wehland J, Dufrenne J, Notermans S: **Naturally occurring virulence-attenuated isolates of *Listeria monocytogenes* capable of inducing long term protection against infection by virulent strains of homologous and heterologous serotypes.** *FEMS Immunol Med Microbiol* 1994, **10:**1–9.
19. Camejo A, Buchrieser C, Couve E, Carvalho F, Reis O, Ferreira P, Sousa S, Cossart P, Cabanes D: **In vivo transcriptional profiling of *Listeria monocytogenes* and mutagenesis identify new virulence factors involved in infection.** *PLoS Pathog* 2009, **5:**e1000449.
20. Chatterjee SS, Hossain H, Otten S, Kuenne C, Kuchmina K, Machata S, Domann E, Chakraborty T, Hain T: **Intracellular gene expression profile of *Listeria monocytogenes*.** *Infect Immun* 2006, **74:**1323–1338.
21. Joseph B, Goebel W: **Life of *Listeria monocytogenes* in the host cells' cytosol.** *Microbes Infect* 2007, **9:**1188–1195.
22. Chen Y, Zhang W, Knabel SJ: **Multi-virulence-locus sequence typing identifies single nucleotide polymorphisms which differentiate epidemic clones and outbreak strains of *Listeria monocytogenes*.** *J Clin Microbiol* 2007, **45:**835–846.
23. Canchaya C, Fournous G, Brussow H: **The impact of prophages on bacterial chromosomes.** *Mol Microbiol* 2004, **53:**9–18.
24. Dorscht J, Klumpp J, Bielmann R, Schmelcher M, Born Y, Zimmer M, Calendar R, Loessner MJ: **Comparative genome analysis of *Listeria* bacteriophages reveals extensive mosaicism, programmed translational frameshifting, and a novel prophage insertion site.** *J Bacteriol* 2009, **191:**7206–7215.
25. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL: **Versatile and open software for comparing large genomes.** *Genome Biol* 2004, **5:**R12.
26. Ghai R, Hain T, Chakraborty T: **GenomeViz: visualizing microbial genomes.** *BMC Bioinforma* 2004, **5:**198.
27. Doumith M, Cazalet C, Simoes N, Frangeul L, Jacquet C, Kunst F, Martin P, Cossart P, Glaser P, Buchrieser C: **New aspects regarding evolution and virulence of *Listeria monocytogenes* revealed by comparative genomics and DNA arrays.** *Infect Immun* 2004, **72:**1072–1083.
28. Sokolovic Z, Schuller S, Bohne J, Baur A, Rdest U, Dickneite C, Nichterlein T, Goebel W: **Differences in virulence and in expression of PrfA and PrfA-regulated virulence genes of *Listeria monocytogenes* strains belonging to serogroup 4.** *Infect Immun* 1996, **64:**4008–4019.
29. Jia Y, Nightingale KK, Boor KJ, Ho A, Wiedmann M, McGann P: **Distribution of internalin gene profiles of *Listeria monocytogenes* isolates from different sources associated with phylogenetic lineages.** *Foodborne Pathog Dis* 2007, **4:**222–232.
30. Raffelsbauer D, Bubert A, Engelbrecht F, Scheinpflug J, Simm A, Hess J, Kaufmann SH, Goebel W: **The gene cluster i*nlC2DE* of *Listeria monocytogenes* contains additional new internalin genes and is important for virulence in mice.** *Mol Gen Genet* 1998, **260:**144–158.
31. Linden SK, Bierne H, Sabet C, Png CW, Florin TH, McGuckin MA, Cossart P: ***Listeria monocytogenes* internalins bind to the human intestinal mucin MUC2.** *Arch Microbiol* 2008, **190:**101–104.
32. Domann E, Zechel S, Lingnau A, Hain T, Darji A, Nichterlein T, Wehland J, Chakraborty T: **Identification and characterization of a novel PrfA-regulated gene in *Listeria monocytogenes* whose product, IrpA, is highly homologous to internalin proteins, which contain leucine-rich repeats.** *Infect Immun* 1997, **65:**101–109.
33. Engelbrecht F, Chun SK, Ochs C, Hess J, Lottspeich F, Goebel W, Sokolovic Z: **A new PrfA-regulated gene of *Listeria monocytogenes* encoding a small, secreted protein which belongs to the family of internalins.** *Mol Microbiol* 1996, **21:**823–837.
34. Gouin E, dib-Conquy M, Balestrino D, Nahori MA, Villiers V, Colland F, Dramsi S, Dussurget O, Cossart P: **The *Listeria monocytogenes* InlC protein interferes with innate immune responses by targeting the I{kappa}B kinase subunit IKK{alpha}.** *Proc Natl Acad Sci U S A* 2010, **107:**17333–17338.
35. Rajabian T, Gavicherla B, Heisig M, Muller-Altrock S, Goebel W, Gray-Owen SD, Ireton K: **The bacterial virulence factor InlC perturbs apical cell junctions and promotes cell-to-cell spread of *Listeria*.** *Nat Cell Biol* 2009, **11:**1212–1218.
36. Dramsi S, Dehoux P, Lebrun M, Goossens PL, Cossart P: **Identification of four new members of the internalin multigene family of *Listeria monocytogenes* EGD.** *Infect Immun* 1997, **65:**1615–1625.
37. Steele CL, Donaldson JR, Paul D, Banes MM, Arick T, Bridges SM, Lawrence ML: **Genome sequence of lineage III *Listeria monocytogenes* strain HCC23.** *J Bacteriol* 2011, **193:**3679–3680.
38. Chen J, Xia Y, Cheng C, Fang C, Shan Y, Jin G, Fang W: **Genome sequence of the nonpathogenic *Listeria monocytogenes* serovar 4a strain M7.** *J Bacteriol* 2011, **193:**5019–5020.
39. Cabanes D, Dussurget O, Dehoux P, Cossart P: **Auto, a surface associated autolysin of *Listeria monocytogenes* required for entry into eukaryotic cells and virulence.** *Mol Microbiol* 2004, **51:**1601–1614.
40. Cabanes D, Sousa S, Cebria A, Lecuit M, Garcia-del Portillo F, Cossart P: **Gp96 is a receptor for a novel *Listeria monocytogenes* virulence factor, Vip, a surface protein.** *EMBO J* 2005, **24:**2827–2838.
41. Milohanic E, Jonquieres R, Glaser P, Dehoux P, Jacquet C, Berche P, Cossart P, Gaillard JL: **Sequence and binding activity of the autolysin-adhesin Ami from epidemic *Listeria monocytogenes* 4b.** *Infect Immun* 2004, **72:**4401–4409.
42. Milohanic E, Jonquieres R, Cossart P, Berche P, Gaillard JL: **The autolysin Ami contributes to the adhesion of *Listeria monocytogenes* to eukaryotic cells via its cell wall anchor.** *Mol Microbiol* 2001, **39:**1212–1224.
43. Schaferkordt S, Chakraborty T: **Identification, cloning, and characterization of the *lma* operon, whose gene products are unique to *Listeria monocytogenes*.** *J Bacteriol* 1997, **179:**2707–2716.
44. Severino P, Dussurget O, Vencio RZ, Dumas E, Garrido P, Padilla G, Piveteau P, Lemaitre JP, Kunst F, Glaser P, Buchrieser C: **Comparative transcriptome analysis of *Listeria monocytogenes* strains of the two major lineages reveals differences in virulence, cell wall, and stress response.** *Appl Environ Microbiol* 2007, **73:**6078–6088.
45. Gohmann S, Leimeister-Wachter M, Schiltz E, Goebel W, Chakraborty T: **Characterization of a *Listeria monocytogenes*-specific protein capable of inducing delayed hypersensitivity in *Listeria*-immune mice.** *Mol Microbiol* 1990, **4:**1091–1099.
46. Joseph B, Przybilla K, Stuhler C, Schauer K, Slaghuis J, Fuchs TM, Goebel W: **Identification of *Listeria monocytogenes* genes contributing to intracellular replication by expression profiling and mutant screening.** *J Bacteriol* 2006, **188:**556–568.
47. Sauer JD, Witte CE, Zemansky J, Hanson B, Lauer P, Portnoy DA: ***Listeria monocytogenes* triggers AIM2-mediated pyroptosis upon infrequent bacteriolysis in the macrophage cytosol.** *Cell Host Microbe* 2010, **7:**412–419.
48. Warren SE, Mao DP, Rodriguez AE, Miao EA, Aderem A: **Multiple Nod-like receptors activate caspase 1 during *Listeria monocytogenes* infection.** *J Immunol* 2008, **180:**7558–7564.
49. Wu J, Fernandes-Alnemri T, Alnemri ES: **Involvement of the AIM2, NLRC4, and NLRP3 inflammasomes in caspase-1 activation by *Listeria monocytogenes*.** *J Clin Immunol* 2010, **30:**693–702.
50. Mariathasan S, Weiss DS, Dixit VM, Monack DM: **Innate immunity against *Francisella tularensis* is dependent on the ASC/caspase-1 axis.** *J Exp Med* 2005, **202:**1043–1049.
51. Sansonetti PJ, Phalipon A, Arondel J, Thirumalai K, Banerjee S, Akira S, Takeda K, Zychlinsky A: **Caspase-1 activation of IL-1beta and IL-18 are essential for *Shigella flexneri*-induced inflammation.** *Immunity* 2000, **12:**581–590.
52. Bigot A, Pagniez H, Botton E, Frehel C, Dubail I, Jacquet C, Charbit A, Raynaud C: **Role of FliF and FliI of *Listeria monocytogenes* in flagellar assembly and pathogenicity.** *Infect Immun* 2005, **73:**5530–5539.
53. Koch AL: **Evolution of antibiotic resistance gene function.** *Microbiol Rev* 1981, **45:**355–378.
54. Romero D, Palacios R: **Gene amplification and genomic plasticity in prokaryotes.** *Annu Rev Genet* 1997, **31:**91–111.
55. Reinbothe S, Ortel B, Parthier B: **Overproduction by gene amplification of the multifunctional arom protein confers glyphosate tolerance to a plastid-free mutant of Euglena gracilis.** *Mol Gen Genet* 1993, **239:**416–424.

56. Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV: **Selection in the evolution of gene duplications.** *Genome Biol* 2002, **3**:RESEARCH0008.
57. van Hoof NA, Hassinen VH, Hakvoort HW, Ballintijn KF, Schat H, Verkleij JA, Ernst WH, Karenlampi SO, Tervahauta AI: **Enhanced copper tolerance in *Silene vulgaris* (Moench) Garcke populations from copper mines is associated with increased transcript levels of a 2b-type metallothionein gene.** *Plant Physiol* 2001, **126**:1519–1526.
58. Riehle MM, Bennett AF, Long AD: **Genetic architecture of thermal adaptation in *Escherichia coli*.** *Proc Natl Acad Sci U S A* 2001, **98**:525–530.
59. Brown CJ, Todd KM, Rosenzweig RF: **Multiple duplications of yeast hexose transport genes in response to selection in a glucose-limited environment.** *Mol Biol Evol* 1998, **15**:931–942.
60. Sonti RV, Roth JR: **Role of gene duplications in the adaptation of *Salmonella typhimurium* to growth on limiting carbon sources.** *Genetics* 1989, **123**:19–28.
61. Lai CY, Baumann L, Baumann P: **Amplification of *trpEG*: adaptation of *Buchnera aphidicola* to an endosymbiotic association with aphids.** *Proc Natl Acad Sci U S A* 1994, **91**:3819–3823.
62. Stoll R, Mertins S, Joseph B, Muller-Altrock S, Goebel W: **Modulation of PrfA activity in *Listeria monocytogenes* upon growth in different culture media.** *Microbiology* 2008, **154**:3856–3876.
63. Stoll R, Goebel W: **The major PEP-phosphotransferase systems (PTSs) for glucose, mannose and cellobiose of *Listeria monocytogenes*, and their significance for extra- and intracellular growth.** *Microbiology* 2010, **156**:1069–1083.
64. Ake FM, Joyet P, Deutscher J, Milohanic E: **Mutational analysis of glucose transport regulation and glucose-mediated virulence gene repression in *Listeria monocytogenes*.** *Mol Microbiol* 2011, .
65. Joseph B, Mertins S, Stoll R, Schar J, Umesha KR, Luo Q, Muller-Altrock S, Goebel W: **Glycerol metabolism and PrfA activity in *Listeria monocytogenes*.** *J Bacteriol* 2008, **190**:5412–5430.
66. Cannon GC, Bradburne CE, Aldrich HC, Baker SH, Heinhorst S, Shively JM: **Microcompartments in prokaryotes: carboxysomes and related polyhedra.** *Appl Environ Microbiol* 2001, **67**:5351–5361.
67. Buchrieser C, Rusniok C, Kunst F, Cossart P, Glaser P: **Comparison of the genome sequences of *Listeria monocytogenes* and *Listeria innocua*: clues for evolution and pathogenicity.** *FEMS Immunol Med Microbiol* 2003, **35**:207–213.
68. Ralser M, Wamelink MM, Kowald A, Gerisch B, Heeren G, Struys EA, Klipp E, Jakobs C, Breitenbach M, Lehrach H, Krobitsch S: **Dynamic rerouting of the carbohydrate flux is key to counteracting oxidative stress.** *J Biol* 2007, **6**:10.
69. Makarova KS, Haft DH, Barrangou R, Brouns SJ, Charpentier E, Horvath P, Moineau S, Mojica FJ, Wolf YI, Yakunin AF, van der OJ, Koonin EV: **Evolution and classification of the CRISPR-Cas systems.** *Nat Rev Microbiol* 2011, **9**:467–477.
70. Tang TH, Polacek N, Zywicki M, Huber H, Brugger K, Garrett R, Bachellerie JP, Huttenhofer A: **Identification of novel non-coding RNAs as potential antisense regulators in the archaeon *Sulfolobus solfataricus*.** *Mol Microbiol* 2005, **55**:469–481.
71. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P: **CRISPR provides acquired resistance against viruses in prokaryotes.** *Science* 2007, **315**:1709–1712.
72. Toledo-Arana A, Dussurget O, Nikitas G, Sesto N, Guet-Revillet H, Balestrino D, Loh E, Gripenland J, Tiensuu T, Vaitkevicius K, Barthelemy M, Vergassola M, Nahori MA, Soubigou G, Regnault B, Coppee JY, Lecuit M, Johansson J, Cossart P: **The *Listeria* transcriptional landscape from saprophytism to virulence.** *Nature* 2009, **459**:950–956.
73. Orsi RH, Borowsky ML, Lauer P, Young SK, Nusbaum C, Galagan JE, Birren BW, Ivy RA, Sun Q, Graves LM, Swaminathan B, Wiedmann M: **Short-term genome evolution of *Listeria monocytogenes* in a non-controlled environment.** *BMC Genomics* 2008, **9**:539.
74. Verghese B, Lok M, Wen J, Alessandria V, Chen Y, Kathariou S, Knabel S: **comK prophage junction fragments as markers for *Listeria monocytogenes* genotypes unique to individual meat and poultry processing plants and a model for rapid niche-specific adaptation, biofilm formation, and persistence.** *Appl Environ Microbiol* 2011, **77**:3279–3292.
75. Ewing B, Hillier L, Wendl MC, Green P: **Base-calling of automated sequencer traces using phred.** *I. Accuracy assessment. Genome Res* 1998, **8**:175–185.
76. Gordon D, Abajian C, Green P: **Consed: a graphical tool for sequence finishing.** *Genome Res* 1998, **8**:195–202.
77. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM: **Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd.** *Science* 1995, **269**:496–512.
78. Frangeul L, Nelson KE, Buchrieser C, Danchin A, Glaser P, Kunst F: **Cloning and assembly strategies in microbial genome projects.** *Microbiology* 1999, **145(Pt 10)**:2625–2634.
79. Frangeul L, Glaser P, Rusniok C, Buchrieser C, Duchaud E, Dehoux P, Kunst F: **CAAT-Box, Contigs-Assembly and Annotation Tool-Box for genome sequencing projects.** *Bioinformatics* 2004, **20**:790–797.
80. Bray N, Pachter L: **MAVID: constrained ancestral alignment of multiple sequences.** *Genome Res* 2004, **14**:693–699.
81. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I: **VISTA: computational tools for comparative genomics.** *Nucleic Acids Res* 2004, **32**:W273–W279.
82. Kuenne CT, Ghai R, Chakraborty T, Hain T: **GECO–linear visualization for comparative genomics.** *Bioinformatics* 2007, **23**:125–126.
83. Edgar RC: **PILER-CR: fast and accurate identification of CRISPR repeats.** *BMC Bioinforma* 2007, **8**:18.
84. Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG, Parkhill J: **ACT: the Artemis Comparison Tool.** *Bioinformatics* 2005, **21**:3422–3423.
85. Merkl R: **SIGI: score-based identification of genomic islands.** *BMC Bioinforma* 2004, **5**:22.
86. Waack S, Keller O, Asper R, Brodag T, Damm C, Fricke WF, Surovcik K, Meinicke P, Merkl R: **Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models.** *BMC Bioinforma* 2006, **7**:142.
87. Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.** *Bioinformatics* 2003, **19**:185–193.
88. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response.** *Proc Natl Acad Sci U S A* 2001, **98**:5116–5121.
89. Schaferkordt S, Chakraborty T: **Vector plasmid for insertional mutagenesis and directional cloning in *Listeria* spp.** *Biotechniques* 1995, **19**:720–725.
90. Billion A, Ghai R, Chakraborty T, Hain T: **Augur–a computational pipeline for whole genome microbial surface protein prediction and classification.** *Bioinformatics* 2006, **22**:2819–2820.

doi:10.1186/1471-2164-13-144
Cite this article as: Hain *et al.*: Comparative genomics and transcriptomics of lineages I, II, and III strains of *Listeria monocytogenes*. *BMC Genomics* 2012 **13**:144.

RESEARCH ARTICLE Open Access

Reassessment of the *Listeria monocytogenes* pan-genome reveals dynamic integration hotspots and mobile genetic elements as major components of the accessory genome

Carsten Kuenne[1], André Billion[1], Mobarak Abu Mraheil[1], Axel Strittmatter[2], Rolf Daniel[2], Alexander Goesmann[3], Sukhadeo Barbuddhe[4], Torsten Hain[1*] and Trinad Chakraborty[1*]

Abstract

Background: *Listeria monocytogenes* is an important food-borne pathogen and model organism for host-pathogen interaction, thus representing an invaluable target considering research on the forces governing the evolution of such microbes. The diversity of this species has not been exhaustively explored yet, as previous efforts have focused on analyses of serotypes primarily implicated in human listeriosis. We conducted complete genome sequencing of 11 strains employing 454 GS FLX technology, thereby achieving full coverage of all serotypes including the first complete strains of serotypes 1/2b, 3c, 3b, 4c, 4d, and 4e. These were comparatively analyzed in conjunction with publicly available data and assessed for pathogenicity in the *Galleria mellonella* insect model.

Results: The species pan-genome of *L. monocytogenes* is highly stable but open, suggesting an ability to adapt to new niches by generating or including new genetic information. The majority of gene-scale differences represented by the accessory genome resulted from nine hyper variable hotspots, a similar number of different prophages, three transposons (Tn916, Tn554, IS3-like), and two mobilizable islands. Only a subset of strains showed CRISPR/Cas bacteriophage resistance systems of different subtypes, suggesting a supplementary function in maintenance of chromosomal stability. Multiple phylogenetic branches of the genus *Listeria* imply long common histories of strains of each lineage as revealed by a SNP-based core genome tree highlighting the impact of small mutations for the evolution of species *L. monocytogenes*. Frequent loss or truncation of genes described to be vital for virulence or pathogenicity was confirmed as a recurring pattern, especially for strains belonging to lineages III and II. New candidate genes implicated in virulence function were predicted based on functional domains and phylogenetic distribution. A comparative analysis of small regulatory RNA candidates supports observations of a differential distribution of *trans*-encoded RNA, hinting at a diverse range of adaptations and regulatory impact.

Conclusions: This study determined commonly occurring hyper variable hotspots and mobile elements as primary effectors of quantitative gene-scale evolution of species *L. monocytogenes*, while gene decay and SNPs seem to represent major factors influencing long-term evolution. The discovery of common and disparately distributed genes considering lineages, serogroups, serotypes and strains of species *L. monocytogenes* will assist in diagnostic, phylogenetic and functional research, supported by the comparative genomic GECO-LisDB analysis server (http://bioinfo.mikrobio.med.uni-giessen.de/geco2lisdb).

* Correspondence: Torsten.Hain@mikrobio.med.uni-giessen.de; Trinad.Chakraborty@mikrobio.med.uni-giessen.de
[1]Institute of Medical Microbiology, German Centre for Infection Research, Justus-Liebig-University, D-35392, Giessen, Germany
Full list of author information is available at the end of the article

Background

The genus *Listeria* consists of eight species being *L. monocytogenes*, *L. innocua*, *L. welshimeri*, *L. seeligeri*, *L. ivanovii*, *L. grayi*, *L. marthii* and *L. rocourtiae* [1-3]. *Listeria* are saprotrophic with *L. monocytogenes* and *L. ivanovii* considered facultative pathogens, the latter predominantly causing infections in ruminants [4]. *L. monocytogenes* represents the species most commonly associated with listeriosis in humans which primarily affects immunocompromised individuals [5]. The majority of infections are thought to be foodborne and results in high mortality rates [6].

Strains of *L. monocytogenes* can be grouped into four evolutionary lineages and 12 serotypes representing distinct phylogenetic, ecologic and phenotypic characteristics [7-9]. Lineage I was found to be overrepresented among human clinical isolates and epidemic outbreaks in most studies while lineage II is typically sporadically isolated from both humans and animals. Lineage III and IV are rare and predominantly identified in animals. These associations show frequent regional differences, thus rendering the definition of a natural environment difficult. Lineages II, III and IV show higher recombination rates and a lower degree of sequence similarity than lineage I. This observation was proposed to result from less diverse lifestyles for the latter and may denote strains of lineage I as descendants of a recently emerged highly virulent clone [10,11]. Plasmids are more prevalent in lineage II and include a multitude of resistance genes dealing with toxic metals, horizontal gene transfer, oxidative stress and small toxic peptides [12]. Furthermore, strains of this lineage often show virulence attenuated phenotypes due to deletions inside important virulence genes [13]. About 98% of human cases of listeriosis are caused by strains of serotypes 4b, 1/2a, 1/2b and 1/2c [14].

Virulence of the bacterium is heavily dependent on the virulence gene cluster (VGC, LIPI-1) which promotes cytosolic replication as well as intra- and intercellular movement [15]. A second cluster required for virulence contains an operon of two genes (*inlA/B*) that encode internalins necessary for the attachment to and invasion of non-phagocytic host cells [16]. The species *L. ivanovii* displays a specific island with virulence factors called LIPI-2, comprising of multiple internalins and *smcL* sphingomyelinase hemolysis gene [17]. A subset of strains of lineage I carry an additional hemolysin called listeriolysin S (LIPI-3) which contributes to virulence *in vitro* [18]. Other genes involved in the infectious process modulate the bacterial metabolism and stress response [19,20]. Interestingly, prophage genes may also have a function in virulence as identified by transcriptomic analyses of intracellular regulation of genes of three major lineages [21].

A variety of cell wall components are important for the survival of strains of species *L. monocytogenes* in the environment and the infected host, which are frequently encoded by genes harboring domains involved in cell-wall anchoring or protein-protein interactions (e.g. LPXTG, GW, P60, LysM, lipo-box, LRR) [9,22-26].

To protect from bacteriophage activity, some Archaea and bacteria have developed an adaptive immune system (CRISPR: clustered regularly interspaced short palindromic repeats) based on a variable module of repeats, spacers and protein coding genes (Cas: CRISPR associated) [27]. Recently it was shown that CRISPR spacers can bear sequences homologous to chromosomal genes which may represent a form of autoimmunity or regulatory mechanism [28,29]. Some CRISPR/Cas subtypes lacking endoribonucleases necessary for the maturation of crRNAs were shown to appropriate a *trans*-encoded small RNA (*tracrRNA*) in combination with a host factor (RNase III) in order to facilitate the silencing of foreign nucleic acids [30]. CRISPR/Cas systems were previously identified inside a number of strains of genus *Listeria* but never discussed in detail [21,30-32].

Small non-coding regulatory RNAs have emerged as a further layer of gene expression regulation in prokaryotes [33]. They regulate transcription by pairing with other RNAs, forming parts of RNA-protein complexes, or adopting regulatory secondary structures [34]. Small non-coding RNAs were previously identified in species *L. monocytogenes* based on microarrays or deep sequencing approaches and have been implicated in responses to iron limitation, oxidative stress, low temperature and intracellular growth [35-41].

The pan-genome concept has recently been introduced to explore the diversity of a number of bacterial species and found varying degrees of conservation reflecting differences in habitat, evolutionary pressure and gene pool [42-46]. Analyses of the pan-genome of genus *Listeria* showed that gene loss played an important role in the development of modern *Listeria* species from a putatively pathogenic ancestor [31]. Previous attempts to study the pan-genome of *L. monocytogenes* were focused on the identification of genes present in lineage I/II while being absent in lineage III and based on microarrays containing mostly draft quality genomes missing several serotypes, thus limiting the possible resolution [9,47].

This study is the first one to base its evolutionary analyses on a set of 16 completely sequenced genomes of species *L. monocytogenes* including strains of all serotypes, arguably bearing the most diverse pan-genome to be assessed for this species. These include five previously sequenced and extensively studied strains of three major lineages (I-III) being 4a L99, 4b F2365, 1/2a EGD-e, 1/2a 08–5578 and 1/2a 08–5923 as well as the eleven newly sequenced genomes [21,32,48,49]. Efficient invasion into epithelial cells was described for strains 1/2a EGD-e, 1/2b SLCC2755, and 4b L312 while strains 4c SLCC2376, 3a SLCC7179, 3c SLCC2479, 1/2c SLCC2372, 7 SLCC2482, 3b SLCC2540, 4e SLCC2378, and 4d ATCC19117 displayed attenuation or absence of this ability [50]. An association with human illness was previously established for strains 1/2c SLCC2372, 1/2a 08-5923/08-5578, 3b SLCC2540, and 4b F2365 [32,49,51,52]. We determined common and distinct genetic elements to understand the diversity of forces shaping the species down to the level of strains. Most of the analyses were conducted using the GECO comparative genomics software, which was heavily extended in relation to the previously published version in order to satisfy the needs of this study [53]. This work focuses on major molecular aspects relating to evolutionary adaptation of species *L. monocytogenes*, and is intended to serve as a framework to support future analyses for the *Listeria* research community.

Results and discussion

Basic features of strains selected among known serotypes of *L. monocytogenes*

In order to analyze the evolution and pan-genomic potential of the species, strains of *L. monocytogenes* spanning all known serotypes originating from various sources were selected for comparison (Table 1). The chromosome of *L. monocytogenes* 7 SLCC2482 contains one gap located at 2125011 bp and estimated to have a size of approximately 10000 bp. Four strains harbored plasmids which were described previously [12]. All strains were classified according to known sequence types and chromosomal complexes using the BIGSdb software [7,54].

The chromosomes compared show a similar size, G+C content, average length of protein coding genes and percentage of protein coding DNA (Table 2). The number of coding sequences ranged from 2755 (SLCC2376) to 3010 (08–5578). We identified six 16S-23S-5S-rRNA operons in most strains with the exception of 1/2a 08–5578 and 1/2a 08–5923 which lack one rRNA module and several tRNAs.

Pan-genome model predicts a highly conserved species

The pan-genome of 16 chromosomes of *L. monocytogenes* was found to contain 4387 genes including 114 paralogues based on a similarity cutoff of 60% amino acid identity and 80% coverage of protein alignments (Figure 1). Approximately 78% of coding sequences per strain consist of mutually conserved core genes (2354 / species) indicating a highly stable species backbone with relatively few accessory genes (2033 / species) (Additional file 1). More than half of the species accessory genes (1161) furthermore displayed homologues in only one or two strains implying relatively recent insertions that are rarely fixed in the population. A power law regression analysis predicting a future pan-genomic distribution after further sequencing resulted in a mean power law fitting for new genes of n=397.4N$^{-0.7279}$ (α=0.7279). This indicates a conserved but open pan-genome that permits limited integration of foreign DNA or generation of genetic diversity by other evolutionary forces such as mutation, duplication and recombination as previously described [55]. Regression

Table 1 Origin of compared strains of species *L. monocytogenes*

Serotype	Strain	Lineage	Chromosome accession	Plasmid accession	ST*	CC*	Source of isolate	Year of isolation	Country of isolation	Reference
4c	SLCC 2376	III	FR733651		71		poultry			SLCC: Haase et al. (2011)
4a	L99	III	FM211688		201		cheese	1950	Netherlands	Hain et al. (2012)
3a	SLCC 7179	II	FR733650		91		cheese	1986	Austria	SLCC: Haase et al. (2011)
3c	SLCC 2479	II	FR733649		9	9		1966		SLCC: Haase et al. (2011)
1/2c	SLCC 2372	II	FR733648	FR667691	122	9	human	1935	UK	SLCC: Haase et al. (2011)
1/2a	08-5923	II	NC_013768		120		human	2008	Canada	Gilmour et al. (2010)
1/2a	08-5578	II	NC_013766	CP001603			human	2008	Canada	Gilmour et al. (2010)
1/2a	SLCC 5850	II	FR733647		12	7	rabbit	1924	UK	SLCC: Haase et al. (2011)
1/2a	EGD-e	II	NC_003210		35	9	rabbit	1926	UK	Glaser et al. (2001)
7	SLCC 2482	I	FR720325	FR667690	3	3	human	1966		SLCC: Haase et al. (2011)
1/2b	SLCC 2755	I	FR733646	FR667692	66	3	chinchilla	1967		SLCC: Haase et al. (2011)
3b	SLCC 2540	I	FR733645				human	1956	USA	SLCC: Haase et al. (2011)
4e	SLCC 2378	I	FR733644		73	1	poultry			SLCC: Haase et al. (2011)
4d	ATCC 19117	I	FR733643		2	2	sheep			SLCC: Haase et al. (2011)
4b	L312	I	FR733642		4	4	cheese			Chatterjee et al. (2006)
4b	F2365	I	NC_002973		1	1	cheese	1985	USA	Nelson et al. (2004)

*Sequence Type.
**Clonal Complex.

Table 2 General features of the chromosomes of compared strains

Strain	Gaps	Length of chromosome [bp]	G+C content [%]	Number of CDS	Protein coding DNA [%]	Number of rRNA genes	Number of tRNA genes
SLCC 2376	closed	2840185	38.3	2755	89.3	18	67
L99	closed	2979198	38.2	2925	88.9	18	67
SLCC 7179	closed	2882234	38.0	2826	89.3	18	67
SLCC 2479	closed	2972172	38.0	2935	89.3	18	65
SLCC 2372	closed	2972810	38.0	2936	89.3	18	67
08-5923	closed	2999054	38.0	2966	89.3	15	58
08-5578	closed	3032288	38.0	3010	89.3	15	58
SLCC 5850	closed	2907142	38.0	2866	89.2	18	67
EGD-e	closed	2944528	38.0	2855	89.2	18	67
SLCC 2482	1	2936689*	38.0	2874	89.1	18	67
SLCC 2755	closed	2966146	38.1	2877	89.3	18	67
SLCC 2540	closed	2976958	37.9	2907	89.4	18	67
SLCC 2378	closed	2941360	38.0	2874	89.1	18	66
ATCC 19117	closed	2951805	38.0	2868	89.3	18	67
L312	closed	2912346	38.1	2821	89.3	18	67
F2365	closed	2905187	38.0	2847	88.4	18	67

*including 100 N gap spacer.

curves predict the presence of ca. 6000 different genes in the pan-genome of *L. monocytogenes* after 100 strains have been completely sequenced.

Other studies relying on the hybridization of eight lineage III strains on a microarray based on 20 strains (two complete, 18 draft chromosomes) found a closed species pan-genome [47,56]. These likely represent an underestimation of true sequence diversity of the species because they lack multiple serotypes (e.g. 3a, 3b, 3c, 4e, 7), less stringent similarity cutoffs and a lower number of fully sequenced strains. The pan-genome of genus *Listeria* based on chromosomes of 13 strains (six complete, seven draft) was determined to be open [31].

Summarily, our research shows a conserved species, which tolerates low levels of horizontal gene transfer.

Hyper variable hotspots contain one fourth of the accessory genes and permitted the insertion of major pathogenicity determinants

The accessory gene content of compared strains is not scattered evenly across the chromosomes, but accumulates in nine defined chromosomal regions supporting previous observations considering the clustered distribution of strain-specific genes [57] (Additional file 2, Additional file 3). These hotspots were defined by the localization of at least three non-homologous insertions between mutually conserved core genes. The latter showed no over representation among any particular functional or genetic category. Nearly every fourth of the accessory genes (454 = 22%) was found to be located in such a highly variable region. Interestingly, strains of lineage III displayed an average of 56 genes inside these loci, while strains of lineage I and II contained nearly twice as many (80–90), indicating either stronger deleterious forces in the former or an increased number of insertions in the latter. One third of these genes were accounted for by strain-specific insertions leading to a low average conservation of hotspot genes in only three strains. The majority of these genes have no known function (298), 35 are part of restriction modification systems, and 13 are involved in genetic mobilization.

Only a small number of genes could be identified inside hotspot loci which exhibit an obvious adaptive value for the host genome, including the previously described pathogenicity determinants *inlA/B* and LIPI-3 [16,18]. Transposon Tn916 introduced additional cadmium resistance genes into its host strain 1/2a EGD-e [19]. Two variants of an IS3-like transposon were inserted in different hotspot integration sites of the epidemic lineage I and found to bear multiple surface-associated proteins. The latter are implied in attachment, invasion, and other interactions with the environment and were identified in most hotspots resulting in the presence of a total of 40 genes of this category.

These hyper variable hotspots have previously been suggested to be the result of a founder effect resulting from a primary insertion that did not reduce the fitness of the respective strain, which now offers a larger target for neutral insertions, thus increasing their likelihood [44]. It is tempting to speculate, that these regions represent evolutionary test areas attracting new genetic information by frequent insertions, deletions and other differentiating forces, rarely leading to fixation of genes

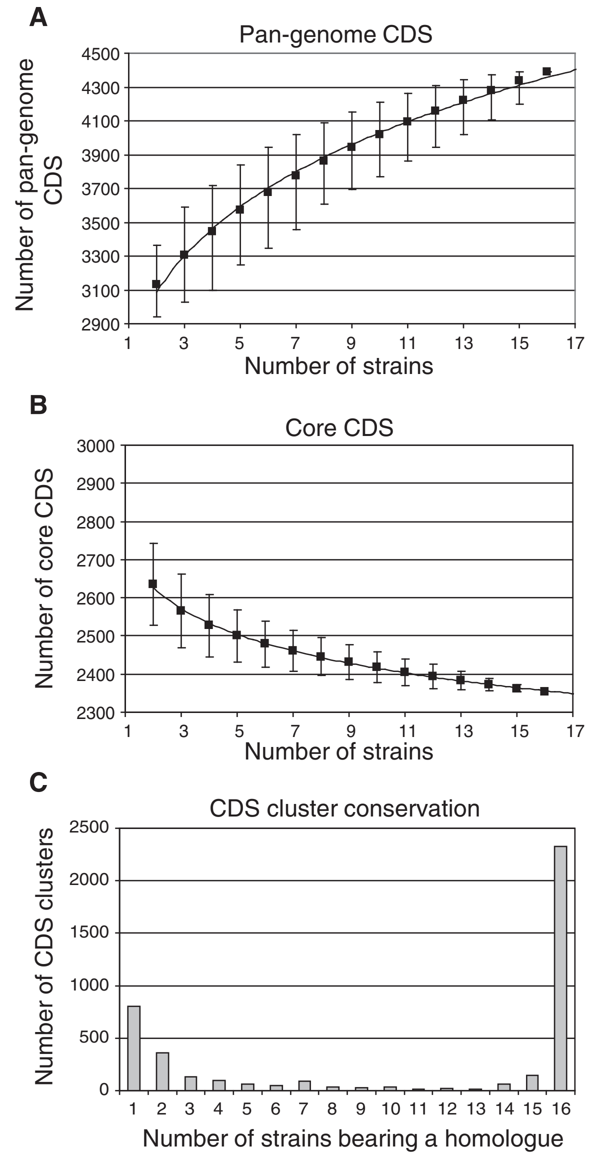


Figure 1 Pan-genomic distribution. Distribution of CDS based on a homology measure of 60% amino acid identity and 80% coverage. Chromosomes were added 10000 times without replacement in a randomized order and the number of core (mutually conserved) and accessory (found in at least one but not all strains) genes was noted. Since mean and median values for each step showed only little variation the mean numbers of gene classes were plotted. In order to predict a possible future pan-genomic distribution for this species we performed a power law fitting. **A**) Pan-genomic CDS after each consecutive addition of a strain, **B**) mutually conserved CDS, **C**) conservation of CDS and homology clusters.

in the population. Interestingly, all but one of these hotspots are located on the right replichore, which thus represents an area of increased genomic plasticity. Only half of the variable regions displayed identifiable mobilization genes indicating either unidentified mobilization genes, decay or other means to facilitate insertions putatively including also mechanisms for homologous recombination.

Chromosomal mobile genetic elements are major sources of diversity – prophages, transposons and genetic islands

In order to find large insertions in the chromosomes of the respective strains we plotted all coding sequences, which were not conserved in all strains, resulting in the identification of between one and five mobile genetic elements (MGE) such as prophages, transposons, insertion sequences and genomic islands per chromosome (Figure 2). These introduced 6 to 235 protein coding genes per strain included in 15 different MGE insertions into 13 distinct chromosomal loci (Additional file 4). This translates into 703 genes of the pan-genome (15%) or one third of the accessory genes.

Among these are 8 different prophages which are typically inserted by site-specific recombination into chromosomal loci adjacent to tRNA genes as previously observed [58]. We also found two different bacteriophages (A006 and A118) which targeted the *comK* gene [59]. Most prophages belong to the class of listeriaphages (B025, A118, and A006) or show a high similarity to unnamed prophages also found in the genera *Bacillus, Enterococcus, Clostridium* and *Staphylococcus*. It should be noted that the only strains without apparently complete prophages are both strains of serotype 4b as well as 4c SLCC2376. Rarity of prophages in serogroup 4 was previously proposed to result from differences in teichoic acid composition, which is supported by strains of this study due to the absence of 12 out of 16 genes of an operon encoding a rhamnose pathway for teichoic acid biosynthesis conserved in all other compared strains (*lmo1076-lmo1091*) as well as several missing glycosyl transferases (*lmo0497, lmo0933, lmo2550*) [60].

Three putative transposons were identified in the strains studied. Two of them are located between homologues of genes *lmo1096-lmo1115* in strain EGD-e (ICELm1, TN916-like) and *lmo2676-lmo2677* in 3c SLCC2479 and 1/2c SLCC2372 (TN554-like), respectively (Additional file 5) [61]. ICELm1 contains two genes involved in cadmium resistance and a fibrinogen-binding protein with an LPXTG domain which is implied in host cell attachment in *Staphylococcus epidermidis* [62]. The Tn554-like transposon introduced an arsenate resistance operon (*arsCBADR*) also found in *Enterococcus faecalis* (ca. 70% amino acid identity) into its host chromosomes. The third putative transposon consists of 15 genes including two insertion elements bearing two IS3-type transposases as found in its complete form in strain 3b SLCC2540 (Additional file 6). It contains a module consisting of a transcriptional regulator and four homologues of a lipoprotein. The latter was predicted by previous studies to furthermore contain an

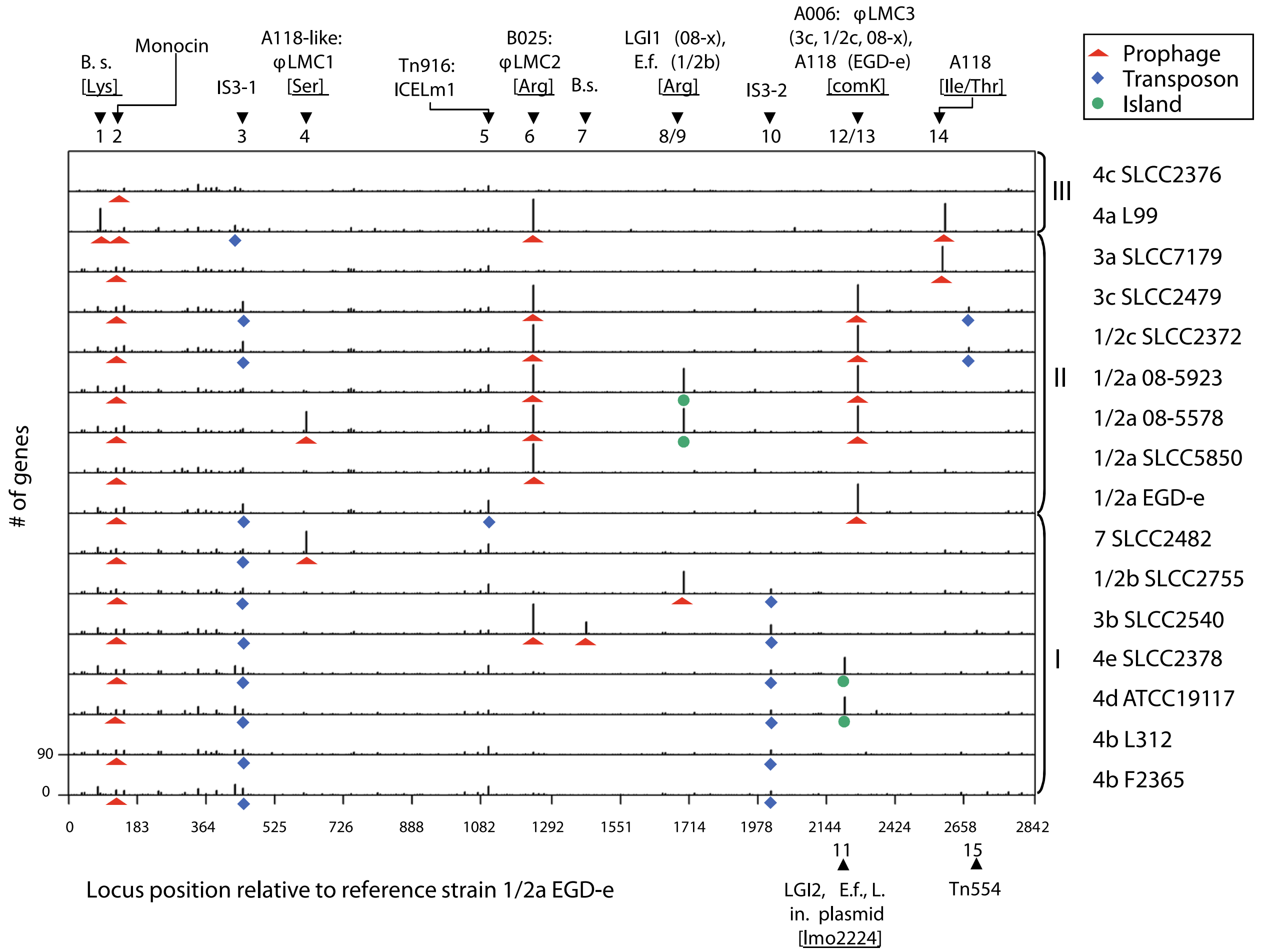


Figure 2 Insertions between syntenic core genes. Bar chart of CDS inserted between syntenic core-CDS existing in all strains depicted relative to reference strain 1/2a EGD-e. The oriC inversion of strains 08–5923 and 08–5578 was removed for this analysis. Mobile genetic elements (MGE) are classified as prophage (red triangle), transposon/IS element (blue square), genetic island (green circle). The MGEs were numbered according to their relative position in strain 1/2a EGD-e. Putative anchor genes in the chromosome (ex.: tRNA, comK) are included in square brackets. If different elements inserted at the same chromosomal locus, the strains involved are denoted in round brackets. Multiple designations per element are delimited by a colon. If an element was not described yet, the genus bearing the highest overall nucleotide similarity to the respective region was included instead (e.g. *Bacillus subtilis, Enterococcus faecalis*). Lineages are denoted with roman numbers.

RGD motif implied in integrin binding and a weak homology to leucine-rich-repeat domains, indicating a putative function in host-pathogen interaction [23,48]. Deletion versions of this transposon, which have lost one insertion element, can be found at the same relative position at approximately 2.1 Mb (e.g. *LMOf2365_2051-9*) in all strains of lineage I and another variant at ca. 0.5 Mb (e.g. *LMOf2365_0493-500)* in a subset of strains of all lineages. Interestingly, indels of the complete transposon and the lipoprotein itself have led to a distribution of 4–7 instances of the lipoprotein in epidemic lineage I in comparison to 0–1 in lineages II and III, which further indicates these two modules as potential targets for research regarding virulence determinants. All but one transposon were found in a hyper variable hotspot suggesting either relaxed deleterious forces in these areas or an enrichment of repeats targeted by the respective mobilization genes.

Another type of MGE is designated genomic island and denotes a module of genes inserted by horizontal gene transfer which frequently encodes fitness conferring genes and typically contains at least one integrase gene employed for mobility. One of these was called Listeria genomic island 1 (LGI1) and putatively introduced by serine recombinases into 1/2a 08–5923 and 08–5578 [49]. It was described to include genes involved in secretion, protein-protein interaction, adhesion, multidrug efflux, signal transduction and restriction modification. We identified a second genomic island named LGI2, which has not yet been described in the literature. It spans approximately 35000 bp in strains 4e SLCC2378 and 4d ATCC19117 and integrated into genes orthologous to *lmo2224* (1/2a EGD-e). This mobile element consists of 36 genes and putatively inserted by means of a bacteriophage integrase (*LMOSLCC2378_2256*) distantly related to temperate *Lactococcus lactis* bacteriophage phiLC3 [63]. Additionally, a putative operon of eight genes coding for arsenate resistance proteins (*LMOSLCC2378_2263-70*) was found to be homologous to a region of *Listeria innocua* Clip11262 plasmid pLI100, indicating recombination between phages, plasmids and chromosomes which resulted in the formation of this mobile element. Other genes of this locus code for ATP transporters, a putative anti-restriction protein, a secreted and a cell wall surface anchor protein.

In summary, nearly one third of the accessory genes of the species have been introduced by identifiable MGEs, representing a large proportion of gene-scale diversity [64]. The distribution of most MGEs is heterogeneous indicating either recent insertions and/or frequent deletion of these sequences. Prophage-related genes of species *L. monocytogenes* represent major chromosomal disparities, have been described to assist intracellular survival, and were found to serve as genetic switches in order to modulate the virulence of its host [21,64-67]. The general rarity of mobile genetic elements in the compared strains nonetheless supposes mechanisms to limit inclusion of foreign DNA as previously proposed [31].

CRISPR/Cas systems represent supplementary bacteriophage defense mechanisms for the species *L. monocytogenes*

Chromosomes of *L. monocytogenes* contain parts of a CRISPR/Cas-system implied in defense versus bacteriophages at three different loci (Additional file 7). These were identified by a combination of PILER-CR 1.02, CRT 1.1 and manual correction using BLASTN leading to slightly higher counts of repeat/spacer modules than previously published for strains 4a L99 and 1/2a EGD-e [21,68,69] (Additional file 8).

All strains bear a putative remnant of a CRISPR-system at ca. 0.5 Mb in strain 1/2a EGD-e which is not associated with any *cas* genes [37]. The distribution of spacers indicates, that ancestors of lineage I and II have lost the *cas* genes necessary to create new spacers inside this locus, leading to a relatively homogenous distribution, while strains of lineage III maintained this ability for a period long enough to completely differentiate their spacer sequences.

Locus 2 is located ca. 10kb adjacent to locus 1 and resembles the *Thermotoga neapolitana* (Tneap) subtype which consists of *cas6*, *cst1*, *cst2*, *cas5t*, *cas3* and *cas2* [70]. Homologues of this system exist in 4a L99, 7 SLCC2482 and 1/2b SLCC2755 at the same relative chromosomal position and no sequence remnants could be identified in other chromosomes, suggesting the insertion of this locus in a common ancestor of these strains. Spacers are identical in strains 7 SLCC2482 and 1/2b SLCC2755, while 4a L99 shows a completely different content.

Locus 3 is inserted into homologues of a lipoprotein gene (*lmo2595*) located at ~2.7 Mb relative to the chromosome of reference strain 1/2a EGD-e. It was found to be present in 1/2a SLCC5850, 7 SLCC2482, 1/2b SLCC2755 and 3b SLCC2540 without any local sequence homologies in other strains, implying insertion into a common ancestor of the former strains. This locus was found to contain *csn2*, *cas2*, *cas1* and *csn1* and thus classified as subtype *Neisseria meningitidis* (Nmeni). Spacer content of locus 3 is clonal for strains 7 SLCC2482 and 1/2b SLCC2755 while 1/2a SLCC5850 and 3b SLCC3540 display mostly unique spacers, including a number of duplicates versus listeriaphages A500 and A118. Locus 3 belongs to subtype Nmeni which was previously described to rely on a *trans*-encoded sRNA (tracrRNA) located upstream of *csn1* and host factor RNase III in order to compensate for a missing endoribonuclease gene [30]. We could exclusively identify perfect matches of the 94 bp tracrRNA variant as expressed by *L. innocua* Clip11262 in all compared strains of *L. monocytogenes* bearing locus 3 at a position upstream of *csn1*. We thus hypothesize, that this locus functions according to the former principles and may only be able to silence foreign nucleic acids inside a host which is able to supply an RNase III enzyme.

All identifiable spacers (81/276) are directed versus known listeriaphages or related composite prophages. We also encountered multiple different spacers homologous to sequences of the same phage in the same array, as well as identical duplications of one spacer. It is tempting to speculate that inclusion of redundant spacer sequences increases the likelihood of a successful defense against the respective bacteriophage (ex.: A118, A500, B025). We never observed identical spacers to be present in multiple arrays, indicating a clear separation of all loci. No spacer was found to target chromosomal or plasmid sequences of species *L. monocytogenes* apart from integrated prophages, indicating that CRISPR/Cas does not serve further regulatory roles facilitated by direct base-pairing with target sequences [28,29].

In conclusion, we propose that an ancestor of genus *Listeria* contained a functional CRISPR locus 1 (*lmo0519-lmo0520*) that lost its associated *cas* genes during early evolutionary events. Interestingly, this locus was previously described as trans-acting small non-coding RNA RliB in strain 1/2a EGD-e indicated in control of virulence [35,37]. Thus, this remnant CRISPR array may have been adapted for regulation in 1/2a EGD-e and possibly other strains of the species. Five of 16 strains compared in this work still contain at least one of two types of putatively functional CRISPR/Cas systems indicating an ongoing selective pressure by bacteriophages. On the other hand, presence or lack of such a system does not correlate with number or type of prophages identified per strain and 11 strains neither bear a functional CRISPR/Cas system nor an increase of other defense mechanisms such as restriction modification systems (data not shown). We suggest that CRISPR/Cas represents an additional line of defense directed against bacteriophage attacks that can be gained by horizontal gene transfer and seems to be effective only for a subset of strains of genus *Listeria*. The variable nature of CRISPR-arrays suggests their future use in differentiating strains or lineages by typing procedures. Further research will now be necessary to determine the operational capability of locus 2 and 3 in the environment or host.

Phylogenies compared – relationships between lineages, serogroups, serotypes and strains according to genomic and genetic content

This analysis used the complete genomic sequences of 19 strains of genus *Listeria* including those of related species being *L. innocua* 6a Clip11262, *L. welshimeri* 6b SLCC5334 and *L. seeligeri* 1/2b SLCC3954 to identify phylogenetic relationships.

In order to enable phylogenetic clustering we created a well-supported (bootstrap >80%) core-genome tree based on an alignment of all concatenated core genes (2018) of 19 strains using Mugsy [71] (Figure 3A/B). This tree shows distances between strains based on small adaptations inside mutually conserved genes, which translate into an approximate timeline when assuming consistent rates of evolution. We found that strains of species *L. monocytogenes* clustered inside three clearly separated lineages in support of previous observations [7,8]. Lineage III contains serotypes 4a and 4c, lineage II includes 1/2a, 1/2c, 3a and 3c and lineage I bears strains of serotypes 1/2b, 3b, 4b, 4d, 4e and 7. Differentiation leading to separate serotypes apparently had little impact on the placement of branches apart from the general lineage. We identified the closest relationships between strains of different serotypes being 1/2b SLCC2755, 7 SLCC2482 (termed phylogenomic group 1 or PG1) and 4e SLCC2378, 4b F2365 (PG2) in lineage I, as well as 1/2a EGD-e, 1/2c SLCC2372, 3c SLCC2479 (PG3) in lineage II, with the exception of clonal strains 08–5578 and 08–5923 which both belong to serotype 1/2a. There is a clear correlation of PGs with previously determined CCs, whereby PG1 strains were classified as CC3, PG2 strains as CC1, and PG3 strains as CC9 [7]. Strains of serotypes 4e and 4d were found on a branch displaying strain 4b L312 as its oldest ancestor in support of a previous hypothesis indicating serotype 4b as ancestral state for serotypes 4e and 4d [7].

We additionally clustered all strains based on the accessory gene content (presence/absence of 2953 genes) to identify the impact of gene-scale indels, which includes most horizontal gene transfer events [72] (Figure 3C). This methodology was shown to be biased towards a tree topology that parallels convergence in lifestyle and thus displays a phenotypical relationship among the compared strains [73]. The resulting tree was found to be well supported (>80% bootstrap) with the exception of the placement of branches neighboring the central *L. monocytogenes* junction, implying early indels and recombination that lead to inconsistent topologies.

If only gene gain and loss are taken into account, lineages of *L. monocytogenes* are closely related to other listerial species, indicating that large evolutionary timeframes shown by the SNP-based core-genome tree resulted in a low number of conserved gene-scale indels.

The opposite is apparent when considering phylogenomic groups, which were found to be closely related in the core-genome tree but to a much lesser degree considering gene content, implying a number of young indels. Interestingly, phylogenomic groups are located at the end of shorter common branches in the gene content tree, which is due to a small number of exclusively conserved genes (PG1: 28, PG2: 20, PG3: 22, primarily hypothetical and truncated genes) (Additional file 1). Thus, strains of phylogenomic groups can be considered closely related but do not necessarily share the same niche or phenotype. Other branches are supported by a varying number of conserved and predominantly hypothetical genes (ex. 4b L312, 4b F2365, 4d ATCC19117, 4e SLCC2378: 18 genes; 3b SLCC2540, 1/2b SLCC2755, 7 SLCC2482: 5 genes) that are distributed along the chromosomes in small modules.

We identified three topological changes between core-genome and gene content tree hinting at shared indels that run contrary to the phylogenomic signal of core-genome SNPs. Removal of genes related to mobile genetic elements (34% of accessory genes) from the gene content matrix resulted in a topology very similar to the core-genome tree. Thus, large-scale insertions, which resulted mainly from bacteriophage integration, run contrary to the “true” phylogenetic signal by inserting many genes in one event as well as by putative parallel insertions into different strains. The only remaining difference was observed considering a common branch for strains of lineage III and apathogenic species, highlighting small-scale indels as causative force. This supports a previous hypothesis suggesting lineage III as a possible deleterious intermediate state between lineages I/II and apathogenic species [7,9,21,74,75].

Interestingly, the majority of accessory genes of species *L. monocytogenes* were either scattered along the chromosomes (46%) or found inside hyper variable regions (20%

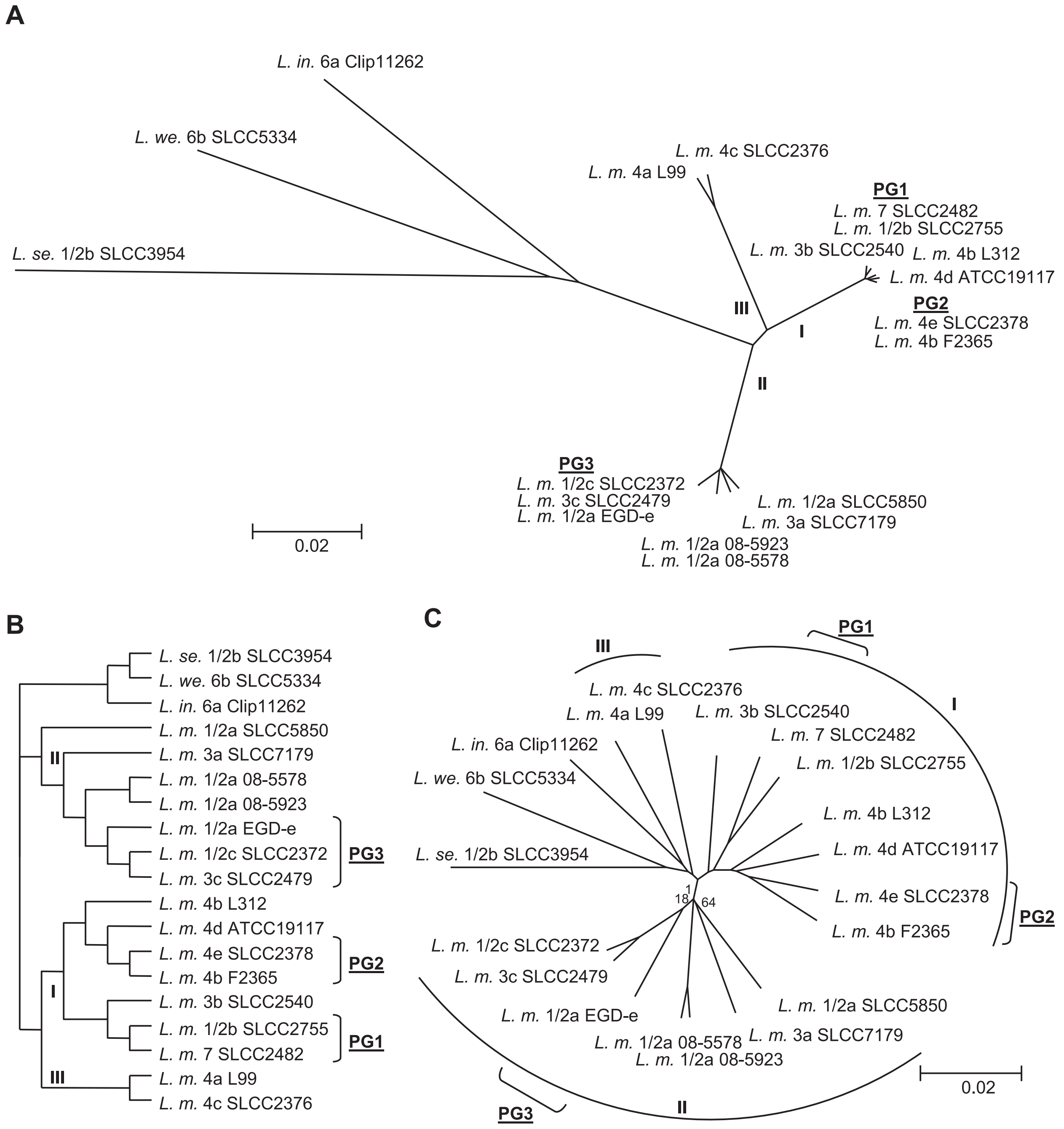


Figure 3 Phylogenomic and -genetic trees. (**A**) Neighbor joining tree based on an alignment of 2018 mutually conserved core genes (amino acid identity >60%, coverage >80%) of 19 strains of genus *Listeria*. Bootstrap support of 100 replicates was always found to be above 80% and thus omitted. Lineages of *L. monocytogenes* are marked in roman letters and phylogenomic groups (PG) describe closely related strains. (**B**) Data of panel A transformed to as cladogram to highlight branching. (**C**) Neighbor joining tree of gene content (presence/absence). Only bootstrap support values below 80% (100 replicates) are indicated.

when excluding MGE) and thus likely originated from a wide range of diversifying forces. Gradual change seems to be a superior factor for the evolution of gene content of *Listeriae* when compared to large-scale insertions of multiple genes by mobile elements.

In summary, tree topologies based on a core-genome alignment and gene content were found to be highly similar despite the obfuscating influence of mobile genetic elements. Other studies on *Rickettsia/Orienta* species and *E.coli/Shigella* found considerable differences in the respective phylogenies indicating more distinct evolutionary histories for the gene repertoires involved [44,76]. The relative correspondence of SNPs and gene-scale indels in genus *Listeria* could be a result of differential acquisition and loss of genes in accordance to various evolutionary descents as previously described considering other genera [77,78].

Frequent loss and disruption of known virulence-associated genes may explain observed phenotypic attenuations

About one third of the genes which displayed compelling evidence for involvement in the infectious process were found to be absent or to code for a truncated protein in at least one of the strains studied, putatively impacting the disease phenotype (Additional file 9) [18,19,79-84]. Rates of mortality of larvae in the *Galleria mellonella* model system indicative of pathogenicity showed that strains of serotype 4b killed most larvae, followed by 1/2c SLCC2372, 3a SLCC7179, 1/2a EGD-e, 1/2b SLCC2755, 3b SLCC2540 and 3c SLCC2479 (Figure 4, Additional file 10). The remaining strains displayed a low degree of pathogenicity in this model, which was described to emulate many aspects of *Listeria* infection seen in vertebrates [85]. Nonetheless, limits of the insect model in forecasting effective human infection become obvious regarding human listeriosis outbreak strains 1/2a 08-5923/08-5578, which only lead to low rates of host mortality following *Galleria* infection. Approximately half of the strains compared in this study were furthermore found to be virulence attenuated as assessed by low invasion rates of epithelial cells [50]. *Galleria* mortality and HeLa cell invasion rates correlated for 6 strains (1/2b SLCC2755, 4b L312, 4c SLCC2376, 7 SLCC2482, 4e SLCC2378, 4d ATCC19117), while 4 strains killed the majority of larvae without being able to invade HeLa cells (3a SLCC7179, 3c SLCC2479, 1/2c SLCC2372, 3b SLCC2540). The latter observation indicates that the respective strains are able to invade other cell types in order to infect an invertebrate host. In order to assess maximum growth rates in a rich medium, the compared strains were furthermore grown in BHI medium at 37°C (Additional file 11). The only outlier was found to be strain 1/2a SLCC5850, which grew considerably slower than the other strains.

In order to correlate phenotypes with genomic differences we performed detailed analyses of virulence-associated genes that allow us to present hypotheses on the evolutionary descent of these changes (Additional file 12). In short, deletions affecting primary virulence genes *prfA* (1/2a SLCC5850), *plcA* (3a SLCC7179), *inlA* (3c SLCC2479), and *inlB* (4b F2365) were identified in four strains [32,86]. A number of surface-associated genes were found to be absent from strains of lineage III and especially from strain 4a L99 [7-9,21,75]. Further deletions which putatively interfere with regulation of the SigB regulon during stress are related to genes *rsbS* (1/2c SLCC2372), *rsbV* (4d ATCC19117) and *rsbU* (3c SLCC2379) [87-89]. The BHI growth attenuation of 1/2a SLCC5850 may result from the specific absence of

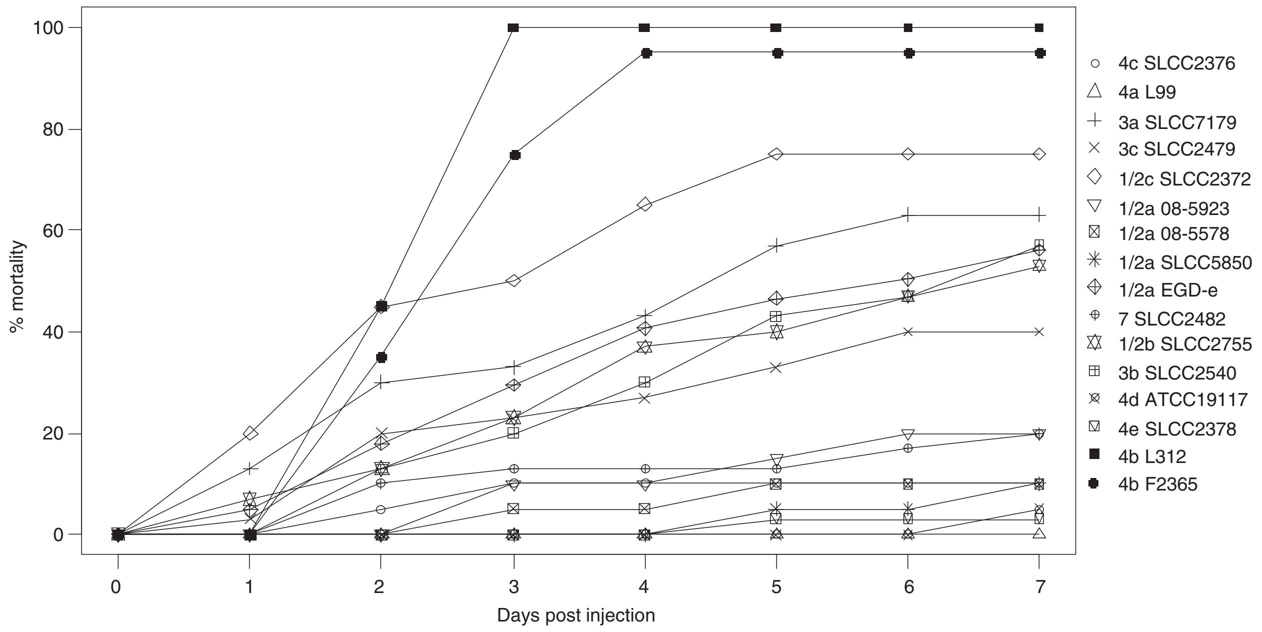


Figure 4 ***Galleria mellonella*** **mortality rates.** Mortality rates of *Galleria mellonella* larvae over the course of seven days post injection. Respective standard deviations can be found in the supplementary material (Additional file 10).

12 genes found in all other compared strains, coding for various proteins involved in energy production/conversion and metabolism (Additional file 1).

In conclusion, strains of *L. monocytogenes* frequently lose determinants of pathogenicity leading to virulence-attenuated phenotypes, which may be advantageous in some environments, especially considering lineage III [7-9,21,75]. Interestingly, highly invasive and/or pathogenic strains of serotypes 4b, 1/2a, 1/2b, and 1/2c also displayed a range of deletions here, indicating a certain amount of redundancy of these functions [18,31,32].

Distribution of surface-associated genes displays conserved lineage-backbones with strain-specific adaptations

A detailed examination was undertaken to spot relevant patterns of presence or absence of surface-associated genes mediating interaction with the environment and the infected host, and to invoke evolutionary explanations (Additional file 13, Additional file 14).

To conclude, genes bearing P60 or LysM domains showed little variation among the strains studied (Additional file 15) [22,23]. Between 6 and 16 non-core lipoprotein coding genes were identified, indicating some differentiation. These were frequently located in chromosomal hotspots of horizontal gene transfer and found inside or adjacent to prophage insertions, hinting at putative methods of transmission. Interestingly, all strains of epidemic lineage I show an exclusive gene (*LMOf2365_1974*) with both LPXTG and GW domains, which may become a future research target when considering the role of cell wall anchored modulators of virulence or pathogenicity.

Internalins are involved in cell adhesion and invasion of host cells and contain a leucine-rich repeat (LRR) domain indicated in protein-protein interaction (Additional file 16) [24-26]. InlB B-repeats represent a hallmark of previously described virulence-associated internalins [90], and were identified in 15 clusters, thus increasing the probability of the respective genes to be involved in host-pathogen interaction. The distribution of putative internalins revealed that only four of 42 homology clusters are mutually conserved, confirming previous observations of diversity, especially considering lineages II and III [91,92]. A number of known virulence-associated internalins were absent in a subset of strains, putatively resulting in a reduced number of infectable cell types (lineage III: *inlC* and *inlF*, 4a L99: *inlGHE*, *inlI* and *inlJ*, 3c SLCC2479: *inlA*, 4b F2365: *inlB*) [9,21,32,75,92]. The absence of *inlC* in strains of lineage III may have been caused by a deleterious transposition moving two adjacent lipoprotein coding genes (*lmo1264-5*) by approximately 600kb to replace the internalin (Additional file 17). Interestingly, we identified different versions of *inlF* and *inlJ* in lineage I as compared to lineages II/III, putatively resulting in different adhesion properties and implicated in host tropism [93]. Only one internalin was found to be specific and mutually conserved for lineage I (*LMOf2365_0805*), indicating this gene for further research regarding virulence.

Taken together, we found that most surface-associated genes are either mutually conserved or were likely present in an early ancestor of a lineage, implying a fixed core-functionality that is rarely complemented by strain-specific additions confirming previous observations [22-24]. Nonetheless, we identified a number of novel surface-associated genes, including their distribution among all serotypes of species *L. monocytogenes*, thereby presenting a pool of candidates for future analysis considering virulence and pathogenicity.

Ancestral genes of serotypes, serogroups and lineages reveal new marker and virulence-associated genes while strain-specific genes rarely represent an obvious extension of functionality

In order to identify conserved ancestral genes which may be important for the differentiation of lineages, we collected genes that were found in all strains of a lineage (>60% amino acid identity, >80% coverage) and absent in all strains of other lineages (Additional file 18). Thus, 33 lineage-III-specific, 22 lineage-II-specific and 14 lineage-I-specific core genes could be identified, which are largely supported by previous microarray-based studies [9,47]. Due to analyses of genetic localization and sequence composition, we want to propose the hypothesis that ancestral strains of lineage I and III diverged from lineage II by loss of genes related to carbohydrate metabolism and gain of hypothetical and surface-associated genes. This theory is based on the following observations: (1) distinct lineage core genes of lineage II predominantly include PTS systems and ABC transporters involved in carbohydrate metabolism organized in three operon-like islands, while those of lineages I and III mainly consist of scattered hypothetical and surface-associated proteins, (2) specific core genes of lineage II display no deviation from the average G/C content of the respective chromosome or codon usage disparities frequently associated with horizontal gene transfer, (3) strains of lineages I and III contain putative sequence remnants of some of these genes (*lmo0734*, *lmo1060*, ~60bp with >75% nucleotide identity), (4) neighborhood and sequence of specific core genes of lineages I and III show more ambiguous patterns including putative insertions, especially considering surface-related proteins (data not shown). According to this hypothesis, ancestral strains of lineages I and III have lost genes related to carbohydrate metabolism and instead gained genes coding for surface-associated proteins serving different needs considering nutrients and interaction with the environment.

The adaptation of strains of lineage III furthermore included loss of genes implicated in food preservation measures, pathogenicity, or virulence as previously described [9,10,21,48,94,95]. We identified 45 genes found to be conserved in 13 out of 14 strains of predominantly human listeriosis-related lineages I and II while being absent from both strains of lineage III (Additional file 19). These comprise genes coding for 16 hypothetical proteins, 14 metabolic enzymes, 6 surface-associated proteins and 4 transcriptional regulators. Affected metabolic pathways include non-mevalonate isoprenoid, fructose and arginine biosynthesis, as well as a nitroreductase and a hydrolase [9,10,21,48,94,95]. Other genes related to stress resistance exclusively conserved in these lineages include the intracellularly up-regulated A118-like prophage rest also known as monocin or *lma*-operon [21]. Furthermore, lineage III does not contain genes coding for multiple internalins and amidases associated with invasion (*inlF, inlC, lmo0129, lmo0849*) [9,21,47]. In summary, strains of less virulent and less pathogenic lineage III mainly differ from the other two lineages by loss of genes involved in metabolism, stress resistance and surface-associated functions implied in adaptation to the complex inter- and intracellular environment inside the host, as well as resistance towards food preservation measures [9,21,47].

We also tried to identify exclusive indels for serogroups or –types that are represented by at least two strains in this analysis in order to uncover ancestral sequences (Additional file 20). We found nine genes to be specific for all strains of serogroup 4, while 16 genes are specifically absent, most of which were already described to be responsible for differences in teichoic acid composition [32,95]. Neither strains of serogroups 3 or 1/2, nor of serotypes 1/2a or 4b show exclusive gene indels, indicating that the respective variable antigens either result from minor changes inside coding genes, from differences located in intergenic regions (ex. promoters, imperfect automatic prediction of ORFs, operon structures, etc.) or from heterogeneous causes.

In order to assess the impact of recent adaptations, strain-specific genes were examined (Additional file 21). Between 11 (3c SLCC2479) and 177 (4a L99) genes per strain were classified as specific, including 0 (4b L312) to 93 (4a L99) genes inserted by a set of previously determined mobile genetic elements dominated by specific prophages. Up to 37 strain-specific genes were found to be fragments of genes either split or truncated by the insertion of a premature stop-codon ("pseudogenes"). Most of these are transporters, metabolic enzymes or regulators and in many cases associated with virulence or pathogenicity as described previously [8]. Strains 1/2a SLCC5850 and 7 SLCC2482 displayed an overrepresentation of fragmentary CDS, which may mark the recent onset of a reductive adaptation. Strain 4b L312 was isolated from cheese and shows a specific insertion of an additional lactose/cellobiose PTS (*LMOL312_2315-20*), which could represent an adaptation to dairy products. A specific element found in strain 3b SLCC2540 resembles the bacteriocin transport and resistance system lantibiotic sublancin 168 (*LMOSLCC2540_2733-40*, up to 28% amino acid identity at 100% coverage) [96]. We found no homologue to the *sun*A bacteriocin peptide, suggesting either export of a different bacteriocin or an exclusive function in resistance to these molecules. Interestingly, eight non-homologous restriction-modification systems were also found to be strain-specific, confirming observations of their "selfish" and competitive nature [97].

Small non-coding RNA candidates of *L. monocytogenes* are largely conserved within the species

Previous transcriptomic analyses uncovered 210 regulatory sRNA candidates expressed in *L. monocytogenes,* some of which have been implicated in adaptation to iron limitation, oxidative stress, low temperature or intracellular survival [35-41]. We identified homologues of these in all compared strains in order to identify patterns associated with evolutionary descent and possible involvement in the infectious process using sRNAdb [98] (Additional file 22).

Only 43 of these were found to be accessory sRNAs, defined as being absent from at least one compared strain, including 20 sRNAs that are only present in a subset of strains of lineage II. Approximately half of those differentially distributed sRNAs, that were previously suggested to be involved in virulence or pathogenicity by growth attenuation of deletion mutants in mice (*rli33-1, rli38, rli50*) or by intracellular up-regulation in macrophages (*rli24, rli28, rli29, rliC, rli85, rli95, rli48, rli98, rliG*) were also exclusively present in a varying subset of strains of lineage II [35,36]. It should be noted that this subset never included strain 3a SLCC7179, implying that ancestral strains of 3c SLCC2479, 1/2c SLCC2372 and serotype 1/2a contained a specific range of sRNAs in order to adapt to the environment and to modulate the infectious process.

We found only *rli38, rli62,* and *rliG* to be specifically present in strain 1/2a EGD-e, whereby the latter two sRNAs inserted as part of specific prophage A118 (MGE-13). Transcriptional activation of prophage genes was reported previously, but an impact on phenotype due to prophage-related sRNAs has still to be elucidated in species *L. monocytogenes* [21,50].

Interestingly, strain 3a SLCC7179 shows a fragmented homologue of *ssrA* (*tmRNA*, 391/500 bp = 78% coverage) necessary for the *trans*-translation of mRNAs that lack a natural stop-codon. Some strains of *E. coli* contain an alternative sRNA termed *afrA* (*yhdL*), which can serve as a possible replacement but was found to be absent from all compared strains [99]. Thus, we speculate that either the shortened *ssrA* gene is still functional or that species *L. monocytogenes* or specifically strain 3a SLCC7179 harbor another yet unknown system to recycle stalled ribosomes and incomplete polypeptides.

In summary, evolution of small non-coding RNAs represents an ongoing process in species *L. monocytogenes.* This excludes all riboswitches found to be mutually conserved in all compared chromosomes, strengthening a hypothesis implicating *cis* acting RNA regulation as an ancient mechanism [36]. Small non-coding RNA transcriptomic analysis of strains of lineages I and III will now be required to uncover their specific regulatory networks on this level.

LisDB – a comparative genomics server for the *Listeria* research community

A large part of the analysis presented in this study is based on the GECO comparative genomics software [53]. We have created a public web-server that includes all published chromosomes and plasmids of genus *Listeria,* as well as a subset of genomes of related genera. The main function of this tool is the identification of homologous genes between replicons to uncover relationships of genomic regions or complete pan-genomic distributions. These data can be visualized graphically or exported in the form of tab-delimited lists. Among the latter are matrices sorted for conservation in selected replicons or for synteny according to a reference strain. Gene gain and loss between two replicons can be identified and nucleotide or amino acid sequences can be exported. GECO-LisDB is accessible at the following address: http://bioinfo.mikrobio.med.uni-giessen.de/geco2lisdb.

Conclusions

Listeria monocytogenes represents a well-characterized pathogen and model system for infection research. Extension of fully sequenced genomes by 11 strains to include all serotypes of the species allowed evolutionary analyses of unprecedented depth. Comparative examination in conjunction with public data revealed that (i) the species pan-genome is highly stable but not closed, (ii) accessory genes are mainly located in defined chromosomal regions (nine hyper variable hotspots, nine different prophages, three transposons, and two mobilizable islands) constituting primary loci of gene-scale species evolution, (iii) potentially functional CRISPR/Cas systems of different subtypes are infrequent but may shape genome diversity, (iv) evolutionary distances observed between lineages of *L. monocytogenes* and apathogenic species are mostly the result of SNPs rather than gene-scale indels that are rarely commonly inherited, highlighting the potential impact of small-scale mutation on long-term development, (v) frequent loss or truncation of genes described to be vital for virulence or pathogenicity was confirmed as a recurring pattern, especially for lineages II and III.

The presence or absence of genes among all serotypes of species *L. monocytogenes* uncovered by this study will be helpful for further diagnostic, phylogenetic and functional research, and is assisted by the comparative genomic GECO-LisDB analysis server (http://bioinfo.mikrobio.med.uni-giessen.de/geco2lisdb).

Methods

Sequencing

The 11 isolates to be sequenced were selected to achieve full coverage of serotypes of species *L. monocytogenes* as previously characterized by MLST, PFGE, and MALDI-TOF [7,100] (Table 1). DNA was purified per strain using Epicentre's MasterPure gram-positive DNA purification kit as recommended by the manufacturer and ten μg of genomic DNA were used for library-preparation following the manufacturer`s constructions (Roche 454 Life Science GS FLX Shotgun DNA Library manual). Sequencing was performed on a 454 GS-FLX system using GS FLX Standard Chemistry. Between 213437 and 297585 reads per strain were *de novo* assembled with the GS Assembler (Newbler 1.1.03.24). The resulting contigs were compared to published strains of *L. monocytogenes* covering major lineages (eg. 4a L99, 1/2a EGD-e, 4b F2365) using Mauve for scaffolding purposes. Differing layouts were assessed manually and joined to a preliminary consensus order. PCR-based techniques followed to close the remaining gaps partially assisted by Minimap (unpublished software) to identify specific primer pairs. This software combines BLASTN and Primer3 in order to identify primer candidates located at the edge of each contig. Primer candidates were selected to not target repetitious sequences (>70% nucleotide identity at >50% coverage). PCRs were sequenced with Sanger ABI Big Dye technology (Applied Biosystems). Sanger reads were incorporated into the assembly using the GAP4 software package v4.11 and SeqMan (Lasergene 5). A total of 487 gaps were closed this way resulting in finished sequences covered from either high-quality 454-reads or Sanger-reads. The completed chromosomes achieved mean coverages between 16-26x and 99.67–99.93% of the bases carried Q40 or higher quality scores. The final gap in the chromosome of *L. monocytogenes* 7 SLCC2482 was marked with a sequence of 100 Ns. Sequencing and finishing procedures were carried out by the Goettingen Genomics Laboratory (Goettingen, Germany), the Institute of Medical Microbiology of the Justus-Liebig University (Giessen, Germany), Roche (Germany), and Agowa (Berlin, Germany). All replicons were deposited in the EMBL database (see Table 1 for accession numbers).

Annotation

Automatic annotation was performed by GenDB, which includes steps for the identification of protein coding sequences (CDS), rRNA and tRNA genes as well as similarity searches against major gene and protein databases [101]. The annotation was enriched using a separate bidirectional best BLASTP step (>80% amino acid identity, >90% coverage) to incorporate data from *L. monocytogenes* 4a L99 (EMBL-Bank: FM211688) and the surface protein prediction software Augur using default parameters [102]. Further annotation was extracted from publications dealing with specific classes of genes such as CRISPR/Cas [70] and known internalins [103]. All information obtained was joined and mapped onto a list of clusters bearing all genes of eleven strains (homology >80% amino acid identity, >90% coverage) using GECO [53] and manually curated according to the following rules with decreasing relevancy: (1) homology to a known gene group (e.g. Cas, internalin, surface-associated) (2) homology to a coding sequence from strain 4a L99, (3) classification as a surface-associated protein-coding gene according to Augur, (4) at least partial homology (>60% amino acid identity, >80% coverage) to a gene family found in Pfam [104] (5) or at least partial homology (>60% amino acid identity, >80% coverage) to a gene found in the NCBI nr database. A manual scan of the complete chromosomes using the GECO visualization interface revealed a number of genes that were fragmented (at least 25% shorter than orthologous genes of reference strains 4a L99, 1/2a EGD-e, and 4b F2365) due to the presence of premature stop-codons and thus annotated as putative fragmentary genes. All automatic annotations were adapted in order to achieve congruent annotations for modules of genes. If no annotation was possible according to these rules, the respective putative protein-coding gene was labeled as a hypothetical protein.

Comparative analyses

Homologous coding sequences were identified by BLASTCLUST [105] as implemented in the comparative genomics software GECO [53]. The standard similarity criterion was set to a minimum of 60% amino acid identity and 80% coverage of both proteins. Chromosomal regions were checked manually using the comparative genome browser of GECO in order to find orthologous CDS which satisfied the homology criteria and were located in a syntenic region in comparison to a reference strain. In some cases a stricter analysis based on 80% amino acid identity and 90% coverage was additionally employed to reduce the number of false positives. In order to avoid excessive redundancy, we denote only one gene of a homologous cluster in brackets, which can be further assessed using either the GECO LisDB server (http://bioinfo.mikrobio.med.uni-giessen.de/geco2lisdb) or the supplementary homology matrix (Additional file 1).

Pan-genome analysis

The pan-genome size of *L. monocytogenes* was predicted based on the chromosomes of 16 sequenced strains compared in this study. We employed the standard BLASTCLUST homology cutoff of 60% amino acid identity and 80% coverage for this analysis. Chromosomes were added 10000 times in a randomized order without replacement, and the number of core (mutually conserved), and accessory (found in at least one but not all strains) genes was noted using GECO. Since mean and median values for each step showed little variation, mean numbers of gene classes were plotted. In order to predict a possible future pan-genomic distribution for this species we performed a power law fitting as described previously [55].

Identification of large insertions

The colinearity of chromosomes of *L. monocytogenes* allowed a relatively simple method to identify large insertions. First we masked the sequence inversion surrounding the oriC in strain 08–5923 (LM5923_2737-0270) and 08–5578 (LM5578_2788-0270) by reordering coding sequences to follow the usual chromosomal layout as found in strain 1/2a EGD-e. CDS were then compared in a bidirectional best BLASTP analysis using similarity criteria of more than 60% amino acid identity and 80% coverage of both CDS. Core-CDS existing in all compared strains were identified by single linkage clustering (AB + BC = ABC). All core-CDS showing a break in the synteny (translocation, inversion) relative to reference strain 1/2a EGD-e were removed from the pool. Finally, the number of CDS located between syntenic core-CDS was plotted as a bar chart per strain. Exact borders of mobile genetic elements were identified based on annotation, deviation of GC-content and comparative analysis with sequenced phages and strains of genus *Listeria*.

CRISPR/Cas analysis

Spacer/repeat-arrays were identified with PILER-CR 1.02 and CRT 1.1 using standard parameters with the exception of maximum repeat length, which was increased to 40 [68,69]. Resulting arrays were combined and controlled manually leading to the removal of eleven false positives inside LRR- and LPXTG-domain containing coding sequences. Consensus sequences of repeats of remaining loci were employed for a BLASTN search versus chromosomes of all strains resulting in the identification of multiple decaying spacer/repeat modules that had been ignored by Piler and CRT due to repeat sequence mismatches of up to 20%. Spacers were compared to 10 published bacteriophages of genus *Listeria* (A006: NC_009815, A118: NC_003216, A500: NC_003216, A511: NC_009811, B025: NC_009812, B054: NC_009813, P100: NC_009813, P35: NC_009814, P40: EU855793, PSA: NC_003291), 16 chromosomes and 4 plasmids of strains of this study and the NCBI nt-database using

BLASTN. Alignments showing up to 1 mismatch were deemed homologous. Finally, all spacers where compared to each other using BLASTCLUST considering perfect matches only and mapped to mirror the order of spacers inside the respective loci to visualize the degree of relatedness (Additional file 3, software BlastclustTo-Matrix available upon request). Softening of the homology cutoffs to 80% nucleotide identity at 80% coverage did not result in a meaningful increase of matches. *cas* genes were identified by sequence homology to published data found in the NCBI NT database and Pfam [104].

Phylogeny

A phylogenetic core-genome tree was created based on mutually conserved core CDS of all compared strains including out-group strains *L. innocua* 6a Clip11262, *L. welshimeri* 6b SLCC5334 and *L. seeligeri* 1/2b SLCC3954. These were extracted from a GECO homology matrix (amino acid identity >60%, coverage > 80%) (Additional file 1) following removal of all clusters showing paralogues. A total of 2018 protein coding genes were concatenated resulting in approximately 2 Mb of nucleotide sequence information per strain. The data was aligned using Mugsy [71] and resulting locally collinear blocks were joined per strain and imported into MEGA5 and SplitsTree4 [106,107]. Based on the alignment we created multiple phylogenomic trees (maximum parsimony, minimum evolution, neighbor joining) including 100 bootstrap replicates. Since tree topology was identical in all cases and relative branch lengths showed little variation, we only present trees based on the neighbor joining algorithm.

In order to identify the impact of indels on phylogeny we built a second tree based on the presence and absence of 2953 accessory genes using GeneContent [72]. Distance between strains was calculated with the Jaccard coefficient [108] and a tree was inferred using the neighbor joining reconstruction method including 100 bootstrap replicates.

Identification of surface-associated genes and putative internalins

Surface-associated genes were identified based on sequence similarity to known motifs (P60, LysM, GW, LRR, LPXTG, lipo) using various Hidden Markov Models (HMM) and SignalP as implemented by Augur [102]. A domain was considered present if HMM e-value < 10 and HMM score > 5. All surface-associated homology matrices were created using a higher standard cutoff (80% amino acid identity, 90% coverage) in order to achieve a higher degree of resolution and thus identify even small amounts of sequence dissimilarity. Clusters showing paralogous CDS were manually split according to a GECO synteny analysis.

All CDS containing a leucine rich repeat (LRR) domain were assumed to be putative internalins and checked for the presence of a signal peptide. False positives and negatives as revealed by synteny analysis were corrected manually and the homology cutoff was reduced to 50% identity and 40% coverage if necessary. Apprehension of internalin-types based on predicted internalins from a previous study [103] as well as domains identified by Augur completed the analysis.

Measurement of bacterial growth

Bacterial cultures were grown over night at 37°C in brain heart infusion broth (BHI) and diluted 1:200 the next day for fresh cultures. Automated measuring at 37°C was performed using the Infinite 200 plate reader (Tecan) in 96-well plates with 150 μl volume/well.

Galleria mellonella infection model

In order to assess the degree of pathogenicity of the 16 strains studied, the insect model *Galleria mellonella* was employed [85]. While this model is unable to mimic all features of vertebrate hosts, a number of listerial virulence genes are generally needed for infection in mammals as well as in invertebrates. In short, bacteria were serially diluted using 0.9% NaCl to a concentration of 10^8 cells/ml. The dilution was plated out on BHI agar plates to calculate the inoculum injected. Ten μl (10^6 bacteria) inoculum were injected dorsolaterally into the hemocoel of last instar larvae using 1 ml disposable syringes and 0.4 × 20mm needles mounted on a microapplicator as described previously. After injection, larvae were incubated at 37°C. Larvae were considered dead when they showed no movement in response to touch. No mortality of *Galleria* larvae were recorded when injected with 0.9% NaCl. Two different versions of these independent experiments were conducted. Strains 1/2a 08–5923, 1/2a 08–5578, 1/2a SLCC5850 and 4b F2365 were injected into 10 animals each and the experiment was performed 2 times per strain. The remaining strains were injected into 20 animals each including 3 repetitions. Mean percental mortality rates of 2 × 10 and 3 × 20 larvae were noted, respectively.

Analyses of sRNAs

Multiple studies have previously determined small noncoding RNA candidates of species *L. monocytogenes* that were classified as intergenic sRNAs, antisense sRNAs, or cis-regulatory RNAs (including riboswitches) [35-41]. A consensus list was created, whereby candidate sRNAs overlapping by at least 50% were merged to one putative long transcript. Homologues of these 210 sRNA candidates were identified in all compared strains using a minimum BLASTN cutoff of 60% nucleotide identity and 80% coverage as applied by the sRNAdb software [98] (Additional file 22).

Additional files

Additional file 1: Species homology matrices. General homology matrices showing the distribution of all coding sequences among 16 strains of species *L. monocytogenes* and 19 strains of genus *Listeria* at different cutoffs. This table is sorted for maximum conservation (core genes = top, specific genes = bottom).

Additional file 2: Insertional hotspot ranges. Hotspots showing at least three separate insertions denoted by locustag ranges.

Additional file 3: Comparative genomic GECO figures of hyper variable hotspots. Comparative GECO depictions of insertional hotspots highlighting extensive mosaicism.

Additional file 4: Mobile genetic elements. Distribution of mobile genetic elements ordered by relative position in the chromosome of *L. monocytogenes* 1/2a EGD-e.

Additional file 5: Comparative genomic GECO figures of transposons ICELm1 and TN554. Comparative GECO depiction using a homology measure of 60% amino acid identity and 80% coverage. Displays content and conservation of two transposons.

Additional file 6: Comparative genomic GECO figures of IS3 elements. Comparative GECO depiction using a homology measure of 60% amino acid identity and 80% coverage. Displays duplication of IS3-like transposon.

Additional file 7: Comparative genomic GECO figure of CRISPR/Cas loci. Comparative GECO depictions of three CRISPR/Cas loci using a minimum CDS homology measure of 60% amino acid identity and 80% coverage.Cas genes and spacer/repeat arrays are framed. Locus 1 displayed no associated Cas genes. Locus 3 includes a *trans*-acting sRNA called tracrRNA that was described to compensate for a missing endoribonuclease in conjunction with host factor RNase III.

Additional file 8: CRISPR/Cas loci. Homology matrices and positions of CRISPR/Cas genes and associated arrays of three loci. Spacers were additionally mapped versus the NCBI nt database to identify possible target sequences.

Additional file 9: Known virulence genes. Homology matrix of known virulence genes.

Additional file 10: *Galleria* standard deviations. Standard deviations calculated for independent experiments considering mortality rates of *Galleria mellonella* larvae over the course of seven days post infection.

Additional file 11: Growth curves BHI. Growth of *L. monocytogenes* in BHI medium at 37°C.

Additional file 12: Detailed analyses of reductive evolution of virulence-associated genes. In-depth information about previously described virulence and pathogenicity indicated genes that are absent or truncated in one of the compared strains.

Additional file 13: Plot of Surface-associated CDS. Bar plot depicting the distribution of all surface-associated protein coding genes among studied strains.

Additional file 14: Distribution of surface-associated genes displays conserved lineage-backbones with strain-specific adaptations. Detailed analysis of presence and absence of surface-associated genes.

Additional file 15: Surface-associated CDS. Homology matrices of genes containing a surface-associated domain (NLPC/p60, LysM, GW, LRR, LPxTG, Lipobox, signal peptide).

Additional file 16: Internalins. Homology matrix of genes containing a leucine rich repeat domain and an optional signal peptide.

Additional file 17: Putative transposition of lipoproteins *lmo1264-5* in lineage III. Comparative GECO depiction using a homology measure of 80% amino acid identity and 90% coverage. Displays the putative transposition of lipoproteins *lmo1264-5* in lineage III into the locus that putatively held *inlC* previously.

Additional file 18: Lineage-specific CDS. Homology matrix of coding genes specifically present in one lineage.

Additional file 19: Lineage I/II exclusive CDS. Homology matrix of genes conserved in 13/14 strains of lineages I and II, while being absent from both strains of lineage III.

Additional file 20: Serogroup and –type ancestral indels. Homology matrix of CDS found to be commonly present or absent (ancestral indel) for either one or multiple serogroups or -types.

Additional file 21: Strain-specific CDS. Homology matrix of coding genes specifically present in one strain.

Additional file 22: Small non-coding regulatory RNAs. Homology matrix of sRNA candidates.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

CK carried out bioinformatic tasks related to the sequencing process, performed annotation and comparative analyses and drafted the manuscript. AB participated in the annotation and bioinformatic analyses and assisted in drafting the manuscript. MM performed phenotypic experiments. AS performed the genomic sequencing. AG participated in the design of the study. RD participated in the design of the study. SB assisted in drafting the manuscript. TH conceived of the study, participated in its design and coordination and assisted in drafting the manuscript. TC participated in the design of the study and assisted in drafting the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This project was funded by the German Federal Ministry of Education and Research through ERA-NET program grants sncRNAomics and SPATELIS, and the LOEWE program of the state Hessia collaborative research project Insect Biotechnology to T. H. and T. C., as well as the German Centre for Infection Research, Justus-Liebig University Giessen.
We thank Sonja Voget, who performed the majority of the sequencing experiments that this project is based on and Anita Hoeland, who constructed data visualizations. We would also like to acknowledge Mark Achtman and Jana Haase for retrieving information about isolates stored in the *Listeria* Culture Collection (SLCC). Jordan Pischimarov, Julian Krauskopf, and Sebastian Oehm assisted in genome annotation and Keith Jolley created a BIGSdb database for MLST classification. Last but not least, we would like to thank Alexandra Amend, Claudia Zörb, Nelli Schklarenko and Burkhard Linke for excellent technical assistance.

Author details

[1]Institute of Medical Microbiology, German Centre for Infection Research, Justus-Liebig-University, D-35392, Giessen, Germany. [2]Department of Genomic and Applied Microbiology and Goettingen Genomics Laboratory, Institute of Microbiology and Genetics, Georg-August University Goettingen, Grisebachstrasse 8, D-37077, Goettingen, Germany. [3]Bioinformatics Resource Facility, Center for Biotechnology, Bielefeld University, D-33549, Bielefeld, Germany. [4]ICAR Research Complex for Goa, Ela, Old Goa 403402, India.

Received: 7 September 2012 Accepted: 15 December 2012
Published: 22 January 2013

References

1. Schmid MW, Ng EY, Lampidis R, Emmerth M, Walcher M, Kreft J, Goebel W, Wagner M, Schleifer KH: **Evolutionary history of the genus *Listeria* and its virulence genes.** *Syst Appl Microbiol* 2005, **28:**1–18.
2. Graves LM, Helsel LO, Steigerwalt AG, Morey RE, Daneshvar MI, Roof SE, Orsi RH, Fortes ED, Milillo SR, den Bakker HC, *et al*: ***Listeria marthii* sp. nov., isolated from the natural environment, Finger Lakes National Forest.** *Int J Syst Evol Microbiol* 2010, **60:**1280–1288.
3. Leclercq A, Clermont D, Bizet C, Grimont PA, Le Fleche-Mateos A, Roche SM, Buchrieser C, Cadet-Daniel V, Le MA, Lecuit M, *et al*: ***Listeria rocourtiae* sp. nov.** *Int J Syst Evol Microbiol* 2010, **60:**2210–2214.
4. Vazquez-Boland JA, Kuhn M, Berche P, Chakraborty T, Dominguez-Bernal G, Goebel W, Gonzalez-Zorn B, Wehland J, Kreft J: ***Listeria pathogenesis* and molecular virulence determinants.** *Clin Microbiol Rev* 2001, **14:**584–640.

5. Allerberger F, Wagner M: **Listeriosis: a resurgent foodborne infection.** *Clin Microbiol Infect* 2010, **16:**16–23.
6. Mead PS, Slutsker L, Dietz V, McCaig LF, Bresee JS, Shapiro C, Griffin PM, Tauxe RV: **Food-related illness and death in the United States.** *Emerg Infect Dis* 1999, **5:**607–625.
7. Ragon M, Wirth T, Hollandt F, Lavenir R, Lecuit M, Le MA, Brisse S: **A new perspective on *Listeria monocytogenes* evolution.** *PLoS Pathog* 2008, **4:**e1000146.
8. Orsi RH, den Bakker HC, Wiedmann M: ***Listeria monocytogenes* lineages: genomics, evolution, ecology, and phenotypic characteristics.** *Int J Med Microbiol* 2011, **301:**79–96.
9. Doumith M, Cazalet C, Simoes N, Frangeul L, Jacquet C, Kunst F, Martin P, Cossart P, Glaser P, Buchrieser C: **New aspects regarding evolution and virulence of *Listeria monocytogenes* revealed by comparative genomics and DNA arrays.** *Infect Immun* 2004, **72:**1072–1083.
10. den Bakker HC, Didelot X, Fortes ED, Nightingale KK, Wiedmann M: **Lineage specific recombination rates and microevolution in *Listeria monocytogenes*.** *BMC Evol Biol* 2008, **8:**277.
11. Orsi RH, Sun Q, Wiedmann M: **Genome-wide analyses reveal lineage specific contributions of positive selection and recombination to the evolution of *Listeria monocytogenes*.** *BMC Evol Biol* 2008, **8:**233.
12. Kuenne C, Voget S, Pischimarov J, Oehm S, Goesmann A, Daniel R, Hain T, Chakraborty T: **Comparative analysis of plasmids in the genus *Listeria*.** *PLoS One* 2010, **5:**e12511.
13. Roche SM, Gracieux P, Milohanic E, Albert I, Virlogeux-Payant I, Temoin S, Grepinet O, Kerouanton A, Jacquet C, Cossart P, *et al*: **Investigation of specific substitutions in virulence genes characterizing phenotypic groups of low-virulence field strains of *Listeria monocytogenes*.** *Appl Environ Microbiol* 2005, **71:**6039–6048.
14. Swaminathan B, Gerner-Smidt P: **The epidemiology of human listeriosis.** *Microbes Infect* 2007, **9:**1236–1243.
15. Portnoy DA, Chakraborty T, Goebel W, Cossart P: **Molecular determinants of *Listeria monocytogenes* pathogenesis.** *Infect Immun* 1992, **60:**1263–1267.
16. Hamon M, Bierne H, Cossart P: ***Listeria monocytogenes*: a multifaceted model.** *Nat Rev Microbiol* 2006, **4:**423–434.
17. Dominguez-Bernal G, Muller-Altrock S, Gonzalez-Zorn B, Scortti M, Herrmann P, Monzo HJ, Lacharme L, Kreft J, Vazquez-Boland JA: **A spontaneous genomic deletion in *Listeria ivanovii* identifies LIPI-2, a species-specific pathogenicity island encoding sphingomyelinase and numerous internalins.** *Mol Microbiol* 2006, **59:**415–432.
18. Cotter PD, Draper LA, Lawton EM, Daly KM, Groeger DS, Casey PG, Ross RP, Hill C: **Listeriolysin S, a novel peptide haemolysin associated with a subset of lineage I *Listeria monocytogenes*.** *PLoS Pathog* 2008, **4:**e1000144.
19. Camejo A, Buchrieser C, Couve E, Carvalho F, Reis O, Ferreira P, Sousa S, Cossart P, Cabanes D: **In vivo transcriptional profiling of *Listeria monocytogenes* and mutagenesis identify new virulence factors involved in infection.** *PLoS Pathog* 2009, **5:**e1000449.
20. Mohamed W, Sethi S, Darji A, Mraheil MA, Hain T, Chakraborty T: **Antibody targeting the ferritin-like protein controls *Listeria* infection.** *Infect Immun* 2010, **78:**3306–3314.
21. Hain T, Ghai R, Billion A, Kuenne CT, Steinweg C, Izar B, Mohamed W, Mraheil MA, Domann E, Schaffrath S, *et al*: **Comparative genomics and transcriptomics of lineages I, II, and III strains of *Listeria monocytogenes*.** *BMC Genomics* 2012, **13:**144.
22. Cabanes D, Dehoux P, Dussurget O, Frangeul L, Cossart P: **Surface proteins and the pathogenic potential of *Listeria monocytogenes*.** *Trends Microbiol* 2002, **10:**238–245.
23. Bierne H, Cossart P: ***Listeria monocytogenes* surface proteins: from genome predictions to function.** *Microbiol Mol Biol Rev* 2007, **71:**377–397.
24. Milillo SR, Wiedmann M: **Contributions of six lineage-specific internalin-like genes to invasion efficiency of *Listeria monocytogenes*.** *Foodborne Pathog Dis* 2009, **6:**57–70.
25. Kasamatsu J, Suzuki T, Ishijima J, Matsuda Y, Kasahara M: **Two variable lymphocyte receptor genes of the inshore hagfish are located far apart on the same chromosome.** *Immunogenetics* 2007, **59:**329–331.
26. Boehm T: **Design principles of adaptive immune systems.** *Nat Rev Immunol* 2011, **11:**307–317.
27. Makarova KS, Haft DH, Barrangou R, Brouns SJ, Charpentier E, Horvath P, Moineau S, Mojica FJ, Wolf YI, Yakunin AF, *et al*: **Evolution and classification of the CRISPR-Cas systems.** *Nat Rev Microbiol* 2011, **9:**467–477.
28. Aklujkar M, Lovley DR: **Interference with histidyl-tRNA synthetase by a CRISPR spacer sequence as a factor in the evolution of *Pelobacter carbinolicus*.** *BMC Evol Biol* 2010, **10:**230.
29. Stern A, Keren L, Wurtzel O, Amitai G, Sorek R: **Self-targeting by CRISPR: gene regulation or autoimmunity?** *Trends Genet* 2010, **26:**335–340.
30. Deltcheva E, Chylinski K, Sharma CM, Gonzales K, Chao Y, Pirzada ZA, Eckert MR, Vogel J, Charpentier E: **CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III.** *Nature* 2011, **471:**602–607.
31. den Bakker HC, Cummings CA, Ferreira V, Vatta P, Orsi RH, Degoricija L, Barker M, Petrauskene O, Furtado MR, Wiedmann M: **Comparative genomics of the bacterial genus *Listeria*: Genome evolution is characterized by limited gene acquisition and limited gene loss.** *BMC Genomics* 2010, **11:**688.
32. Nelson KE, Fouts DE, Mongodin EF, Ravel J, Deboy RT, Kolonay JF, Rasko DA, Angiuoli SV, Gill SR, Paulsen IT, *et al*: **Whole genome comparisons of serotype 4b and 1/2a strains of the food-borne pathogen *Listeria monocytogenes* reveal new insights into the core genome components of this species.** *Nucleic Acids Res* 2004, **32:**2386–2395.
33. Mraheil MA, Billion A, Kuenne C, Pischimarov J, Kreikemeyer B, Engelmann S, Hartke A, Giard JC, Rupnik M, Vorwerk S, *et al*: **Comparative genome-wide analysis of small RNAs of major Gram-positive pathogens: from identification to application.** *Microb Biotechnol* 2010, **3:**658–676.
34. Storz G, Opdyke JA, Zhang A: **Controlling mRNA stability and translation with small, noncoding RNAs.** *Curr Opin Microbiol* 2004, **7:**140–144.
35. Toledo-Arana A, Dussurget O, Nikitas G, Sesto N, Guet-Revillet H, Balestrino D, Loh E, Gripenland J, Tiensuu T, Vaitkevicius K, *et al*: **The *Listeria* transcriptional landscape from saprophytism to virulence.** *Nature* 2009, **459:**950–956.
36. Mraheil MA, Billion A, Mohamed W, Mukherjee K, Kuenne C, Pischimarov J, Krawitz C, Retey J, Hartsch T, Chakraborty T, *et al*: **The intracellular sRNA transcriptome of *Listeria monocytogenes* during growth in macrophages.** *Nucleic Acids Res* 2011, **39:**4235–4248.
37. Mandin P, Repoila F, Vergassola M, Geissmann T, Cossart P: **Identification of new noncoding RNAs in *Listeria monocytogenes* and prediction of mRNA targets.** *Nucleic Acids Res* 2007, **35:**962–974.
38. Christiansen JK, Nielsen JS, Ebersbach T, Valentin-Hansen P, Sogaard-Andersen L, Kallipolitis BH: **Identification of small Hfq-binding RNAs in *Listeria monocytogenes*.** *RNA* 2006, **12:**1383–1396.
39. Wurtzel O, Sesto N, Mellin JR, Karunker I, Edelheit S, Becavin C, Archambaud C, Cossart P, Sorek R: **Comparative transcriptomics of pathogenic and non-pathogenic *Listeria* species.** *Mol Syst Biol* 2012, **8:**583. doi:10.1038/msb.2012.11.:583.
40. Oliver HF, Orsi RH, Ponnala L, Keich U, Wang W, Sun Q, Cartinhour SW, Filiatrault MJ, Wiedmann M, Boor KJ: **Deep RNA sequencing of *Listeria* monocytogenes reveals overlapping and extensive stationary phase and sigma B-dependent transcriptomes, including multiple highly transcribed noncoding RNAs.** *BMC Genomics* 2009, **10:**641. doi:10.1186/1471-2164-10-641.
41. Nielsen JS, Olsen AS, Bonde M, Valentin-Hansen P, Kallipolitis BH: **Identification of a sigma B-dependent small noncoding RNA in *Listeria monocytogenes*.** *J Bacteriol* 2008, **190:**6264–6270.
42. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, *et al*: **Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome".** *Proc Natl Acad Sci U S A* 2005, **102:**13950–13955.
43. Hiller NL, Janto B, Hogg JS, Boissy R, Yu S, Powell E, Keefe R, Ehrlich NE, Shen K, Hayes J, *et al*: **Comparative genomic analyses of seventeen *Streptococcus pneumoniae* strains: insights into the pneumococcal supragenome.** *J Bacteriol* 2007, **189:**8186–8195.
44. Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, Bingen E, Bonacorsi S, Bouchier C, Bouvet O, *et al*: **Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths.** *PLoS Genet* 2009, **5:**e1000344.
45. Lefebure T, Stanhope MJ: **Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition.** *Genome Biol* 2007, **8:**R71.
46. Hogg JS, Hu FZ, Janto B, Boissy R, Hayes J, Keefe R, Post JC, Ehrlich GD: **Characterization and modeling of the *Haemophilus influenzae* core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains.** *Genome Biol* 2007, **8:**R103.

47. Deng X, Phillippy AM, Li Z, Salzberg SL, Zhang W: **Probing the pan-genome of *Listeria monocytogenes*: new insights into intraspecific niche expansion and genomic diversification.** *BMC Genomics* 2010, **11**:500.
48. Glaser P, Frangeul L, Buchrieser C, Rusniok C, Amend A, Baquero F, Berche P, Bloecker H, Brandt P, Chakraborty T, *et al*: **Comparative genomics of *Listeria* species.** *Science* 2001, **294**:849–852.
49. Gilmour MW, Graham M, Van DG, Tyler S, Kent H, Trout-Yakel KM, Larios O, Allen V, Lee B, Nadon C: **High-throughput genome sequencing of two *Listeria monocytogenes* clinical isolates during a large foodborne outbreak.** *BMC Genomics* 2010, **11**:120.
50. Chatterjee SS, Otten S, Hain T, Lingnau A, Carl UD, Wehland J, Domann E, Chakraborty T: **Invasiveness is a variable and heterogeneous phenotype in *Listeria monocytogenes* serotype strains.** *Int J Med Microbiol* 2006, **296**:277–286.
51. Haase JK, Murphy RA, Choudhury KR, Achtman M: **Revival of Seeliger's historical 'Special *listeria* culture Collection'.** *Environ Microbiol* 2011, **13**:3163–3171.
52. Mainou-Fowler T, MacGowan AP, Postlethwaite R: **Virulence of *Listeria* spp.: course of infection in resistant and susceptible mice.** *J Med Microbiol* 1988, **27**:131–140.
53. Kuenne CT, Ghai R, Chakraborty T, Hain T: **GECO–linear visualization for comparative genomics.** *Bioinformatics* 2007, **23**:125–126.
54. Jolley KA, Maiden MC: **BIGSdb: Scalable analysis of bacterial genome variation at the population level.** *BMC Bioinforma* 2010, **11**:595.
55. Tettelin H, Riley D, Cattuto C, Medini D: **Comparative genomics: the bacterial pan-genome.** *Curr Opin Microbiol* 2008, **11**:472–477.
56. Phillippy AM, Deng X, Zhang W, Salzberg SL: **Efficient oligonucleotide probe selection for pan-genomic tiling arrays.** *BMC Bioinforma* 2009, **10**:293.
57. Cossart P, Archambaud C: **The bacterial pathogen *Listeria monocytogenes*: an emerging model in prokaryotic transcriptomics.** *J Biol* 2009, **8**:107.
58. Hain T, Steinweg C, Kuenne CT, Billion A, Ghai R, Chatterjee SS, Domann E, Karst U, Goesmann A, Bekel T, *et al*: **Whole-genome sequence of *Listeria welshimeri* reveals common steps in genome reduction with *Listeria innocua* as compared to *Listeria monocytogenes*.** *J Bacteriol* 2006, **188**:7405–7415.
59. Loessner MJ, Inman RB, Lauer P, Calendar R: **Complete nucleotide sequence, molecular analysis and genome structure of bacteriophage A118 of *Listeria monocytogenes*: implications for phage evolution.** *Mol Microbiol* 2000, **35**:324–340.
60. Promadej N, Fiedler F, Cossart P, Dramsi S, Kathariou S: **Cell wall teichoic acid glycosylation in *Listeria monocytogenes* serotype 4b requires *gtcA*, a novel, serogroup-specific gene.** *J Bacteriol* 1999, **181**:418–425.
61. Burrus V, Pavlovic G, Decaris B, Guedon G: **The ICESt1 element of *Streptococcus thermophilus* belongs to a large family of integrative and conjugative elements that exchange modules and change their specificity of integration.** *Plasmid* 2002, **48**:77–97.
62. Pei L, Palma M, Nilsson M, Guss B, Flock JI: **Functional studies of a fibrinogen binding protein from *Staphylococcus epidermidis*.** *Infect Immun* 1999, **67**:4525–4530.
63. Blatny JM, Godager L, Lunde M, Nes IF: **Complete genome sequence of the *Lactococcus lactis* temperate phage phiLC3: comparative analysis of phiLC3 and its relatives in lactococci and streptococci.** *Virology* 2004, **318**:231–244.
64. Orsi RH, Borowsky ML, Lauer P, Young SK, Nusbaum C, Galagan JE, Birren BW, Ivy RA, Sun Q, Graves LM, *et al*: **Short-term genome evolution of *Listeria monocytogenes* in a non-controlled environment.** *BMC Genomics* 2008, **9**:539.
65. Rabinovich L, Sigal N, Borovok I, Nir-Paz R, Herskovits AA: **Prophage excision activates *Listeria* competence genes that promote phagosomal escape and virulence.** *Cell* 2012, **150**:792–802.
66. Chan YC, Raengpradub S, Boor KJ, Wiedmann M: **Microarray-based characterization of the *Listeria monocytogenes* cold regulon in log- and stationary-phase cells.** *Appl Environ Microbiol* 2007, **73**:6484–6498.
67. Schaferkordt S, Chakraborty T: **Identification, cloning, and characterization of the *lma* operon, whose gene products are unique to *Listeria monocytogenes*.** *J Bacteriol* 1997, **179**:2707–2716.
68. Edgar RC: **PILER-CR: fast and accurate identification of CRISPR repeats.** *BMC Bioinforma* 2007, **8**:18.
69. Bland C, Ramsey TL, Sabree F, Lowe M, Brown K, Kyrpides NC, Hugenholtz P: **CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats.** *BMC Bioinforma* 2007, **8**:209.
70. Haft DH, Selengut J, Mongodin EF, Nelson KE: **A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes.** *PLoS Comput Biol* 2005, **1**:e60.
71. Angiuoli SV, Salzberg SL: **Mugsy: fast multiple alignment of closely related whole genomes.** *Bioinformatics* 2011, **27**:334–342.
72. Gu X, Zhang H: **Genome phylogenetic analysis based on extended gene contents.** *Mol Biol Evol* 2004, **21**:1401–1408.
73. Dutilh BE, den Noort V, van der Heijden RT, Boekhout T, Snel B, Huynen MA: **Assessment of phylogenomic and orthology approaches for phylogenetic inference.** *Bioinformatics* 2007, **23**:815–824.
74. Chen J, Jiang L, Chen X, Luo X, Chen Y, Yu Y, Tian G, Liu D, Fang W: ***Listeria monocytogenes* serovar 4a is a possible evolutionary intermediate between *L. monocytogenes* serovars 1/2a and 4b and *L. innocua*.** *J Microbiol Biotechnol* 2009, **19**:238–249.
75. Chen J, Xia Y, Cheng C, Fang C, Shan Y, Jin G, Fang W: **Genome sequence of the nonpathogenic *Listeria monocytogenes* serovar 4a strain M7.** *J Bacteriol* 2011, **193**:5019–5020.
76. Georgiades K, Merhej V, El KK, Raoult D, Pontarotti P: **Gene gain and loss events in *Rickettsia* and *Orientia* species.** *Biol Direct* 2011, **6**:6.
77. Snel B, Bork P, Huynen MA: **Genome phylogeny based on gene content.** *Nat Genet* 1999, **21**:108–110.
78. van Schaik W, Top J, Riley DR, Boekhorst J, Vrijenhoek JE, Schapendonk CM, Hendrickx AP, Nijman IJ, Bonten MJ, Tettelin H, *et al*: **Pyrosequencing-based comparative genome analysis of the nosocomial pathogen *Enterococcus faecium* and identification of a large transferable pathogenicity island.** *BMC Genomics* 2010, **11**:239. doi:10.1186/1471-2164-11-239.
79. Begley M, Sleator RD, Gahan CG, Hill C: **Contribution of three bile-associated loci, *bsh, pva,* and *btlB*, to gastrointestinal persistence and bile tolerance of *Listeria monocytogenes*.** *Infect Immun* 2005, **73**:894–904.
80. Mraheil MA, Billion A, Mohamed W, Rawool D, Hain T, Chakraborty T: **Adaptation of *Listeria monocytogenes* to oxidative and nitrosative stress in IFN-γ-activated macrophages.** *Int J Med Microbiol* 2011, **301**:547–555.
81. Lebreton A, Lakisic G, Job V, Fritsch L, Tham TN, Camejo A, Mattei PJ, Regnault B, Nahori MA, Cabanes D, *et al*: **A bacterial protein targets the BAHD1 chromatin complex to stimulate type III interferon response.** *Science* 2011, **331**:1319–1321.
82. Ryan S, Begley M, Hill C, Gahan CG: **A five-gene stress survival islet (SSI-1) that contributes to the growth of *Listeria monocytogenes* in suboptimal conditions.** *J Appl Microbiol* 2010, **109**:984–995.
83. Joseph B, Mertins S, Stoll R, Schar J, Umesha KR, Luo Q, Muller-Altrock S, Goebel W: **Glycerol metabolism and PrfA activity in *Listeria monocytogenes*.** *J Bacteriol* 2008, **190**:5412–5430.
84. Kirchner M, Higgins DE: **Inhibition of ROCK activity allows InlF-mediated invasion and increased virulence of *Listeria monocytogenes*.** *Mol Microbiol* 2008, **68**:749–767.
85. Mukherjee K, Altincicek B, Hain T, Domann E, Vilcinskas A, Chakraborty T: ***Galleria mellonella* as a model system for studying *Listeria* pathogenesis.** *Appl Environ Microbiol* 2010, **76**:310–317.
86. Chen Y, Ross WH, Whiting RC, Van SA, Nightingale KK, Wiedmann M, Scott VN: **Variation in *Listeria monocytogenes* dose responses in relation to subtypes encoding a full-length or truncated internalin A.** *Appl Environ Microbiol* 2011, **77**:1171–1180.
87. Kang CM, Brody MS, Akbar S, Yang X, Price CW: **Homologous pairs of regulatory proteins control activity of *Bacillus subtilis* transcription factor sigma(b) in response to environmental stress.** *J Bacteriol* 1996, **178**:3846–3853.
88. Palma M, Cheung AL: **sigma(B) activity in *Staphylococcus aureus* is controlled by RsbU and an additional factor(s) during bacterial growth.** *Infect Immun* 2001, **69**:7858–7865.
89. Chaturongakul S, Boor KJ: **RsbT and RsbV contribute to sigmaB-dependent survival under environmental, energy, and intracellular stress conditions in *Listeria monocytogenes*.** *Appl Environ Microbiol* 2004, **70**:5349–5356.
90. Ebbes M, Bleymuller WM, Cernescu M, Nolker R, Brutschy B, Niemann HH: **Fold and function of the InlB B-repeat.** *J Biol Chem* 2011, **286**:15496–15506.
91. Tsai YH, Orsi RH, Nightingale KK, Wiedmann M: ***Listeria monocytogenes* internalins are highly diverse and evolved by recombination and positive selection.** *Infect Genet Evol* 2006, **6**:378–389.

92. Jia Y, Nightingale KK, Boor KJ, Ho A, Wiedmann M, McGann P: **Distribution of internalin gene profiles of *Listeria monocytogenes* isolates from different sources associated with phylogenetic lineages.** *Foodborne Pathog Dis* 2007, **4:**222–232.
93. Balandyte L, Brodard I, Frey J, Oevermann A, Abril C: **Ruminant rhombencephalitis-associated *Listeria monocytogenes* alleles linked to a multilocus variable-number tandem-repeat analysis complex.** *Appl Environ Microbiol* 2011, **77:**8325–8335.
94. Ryan S, Begley M, Gahan CG, Hill C: **Molecular characterization of the arginine deiminase system in *Listeria monocytogenes*: regulation and role in acid tolerance.** *Environ Microbiol* 2009, **11:**432–445.
95. Zhang C, Zhang M, Ju J, Nietfeldt J, Wise J, Terry PM, Olson M, Kachman SD, Wiedmann M, Samadpour M, *et al*: **Genome diversification in phylogenetic lineages I and II of *Listeria monocytogenes*: identification of segments unique to lineage II populations.** *J Bacteriol* 2003, **185:**5573–5584.
96. Dubois JY, Kouwen TR, Schurich AK, Reis CR, Ensing HT, Trip EN, Zweers JC, den Dijl JM: **Immunity to the bacteriocin sublancin 168 Is determined by the SunI (YolF) protein of *Bacillus subtilis*.** *Antimicrob Agents Chemother* 2009, **53:**651–661.
97. Kobayashi I: **Behavior of restriction-modification systems as selfish mobile elements and their impact on genome evolution.** *Nucleic Acids Res* 2001, **29:**3742–3756.
98. Pischimarov J, Kuenne C, Billion A, Hemberger J, Cemic F, Chakraborty T, Hain T: **sRNAdb: a small non-coding RNA database for gram-positive bacteria.** *BMC Genomics* 2012, **13:**384. doi:10.1186/1471-2164-13-384.:384-13.
99. Himeno H: **Novel factor rescues ribosomes trapped on non-stop mRNAs.** *Mol Microbiol* 2010, **78:**789–791.
100. Barbuddhe SB, Maier T, Schwarz G, Kostrzewa M, Hof H, Domann E, Chakraborty T, Hain T: **Rapid identification and typing of listeria species by matrix-assisted laser desorption ionization-time of flight mass spectrometry.** *Appl Environ Microbiol* 2008, **74:**5402–5407.
101. Meyer F, Goesmann A, McHardy AC, Bartels D, Bekel T, Clausen J, Kalinowski J, Linke B, Rupp O, Giegerich R, *et al*: **GenDB–an open source genome annotation system for prokaryote genomes.** *Nucleic Acids Res* 2003, **31:**2187–2195.
102. Billion A, Ghai R, Chakraborty T, Hain T: **Augur–a computational pipeline for whole genome microbial surface protein prediction and classification.** *Bioinformatics* 2006, **22:**2819–2820.
103. Bierne H, Sabet C, Personnic N, Cossart P: **Internalins: a complex family of leucine-rich repeat-containing proteins in *Listeria monocytogenes*.** *Microbes Infect* 2007, **9:**1156–1166.
104. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, *et al*: **The Pfam protein families database.** *Nucleic Acids Res* 2012, **40:**D290–D301.
105. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403–410.
106. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: **MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods.** *Mol Biol Evol* 2011, **28:**2731–2739.
107. Huson DH, Bryant D: **Application of phylogenetic networks in evolutionary studies.** *Mol Biol Evol* 2006, **23:**254–267.
108. Wolf YI, Rogozin IB, Grishin NV, Koonin EV: **Genome trees and the tree of life.** *Trends Genet* 2002, **18:**472–479.

doi:10.1186/1471-2164-14-47
Cite this article as: Kuenne *et al.*: Reassessment of the *Listeria monocytogenes* pan-genome reveals dynamic integration hotspots and mobile genetic elements as major components of the accessory genome. *BMC Genomics* 2013 **14**:47.