Graz University of Technology
Faculty of Technical Chemistry, Chemical and Process Engineering, Biotechnology
Institute of Molecular Biotechnology

Revitalization, whole-genome sequencing and genotypic analysis of up to 95-year old *Listeria monocytogenes* isolates of the historic "Special Listeria Culture Collection"

Master Thesis

written by

Patrick HYDEN

Submitted in fulfillment of the requirements
for the academic degree Master of Science
in the field of Bioinformatics

Graz, February 2015

Supervisor:

Univ.-Prof. Dipl.-Biol. Dr.rer.nat. Christoph Wilhelm SENSEN

……………………………………………………..

EIDESSTATTLICHE ERKLÄRUNG

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Graz, am …………………………… ………………………………………………..

(Unterschrift)

Englische Fassung:

STATUTORY DECLARATION

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

…………………………… ………………………………………………..

date (signature)

Danksagung

An erster Stelle möchte ich mich bei meiner Familie für ihre bedingungslose Unterstützung bedanken. Vor allem gilt der Dank meinen Eltern Wolfgang und Eva Hyden, und meinen Großeltern Dr. Hans und Ingrid Leikauf, für die finanzielle und persönliche Unterstützung während meines gesamten Studiums.

Ich bedanke mich bei meinem Betreuer Univ.-Prof. Dipl.-Biol. Dr.rer.nat Christoph Wilhelm Sensen für die Betreuung und Begleitung über den Rahmen meiner Masterarbeit hinaus, und für wertvolle Anregungen und Diskussionen.

Weiters danke ich Dr. Werner Ruppitsch, Dr. Ariane Pietzka, Dr. Alexander Indra, Dr. Burkhard Springer und Prof. Dr. Franz Allerberger für die Betreuung und Begleitung bei der Agentur für Gesundheit und Ernährungssicherheit (AGES) und die Möglichkeit ein derart aktuelles Thema bearbeiten zu dürfen.

Meinen besonderen Dank möchte ich auch an alle KollegInnen bei der AGES an den Standorten Graz Beethovenstraße und Wien Währingerstraße richten. Ausdrücklich möchte ich mich bei Bernhard Prewein, Sarah Lepuschitz, Marion Blaschitz und Anna Lenkh für die schöne Zeit und die wertvollen Gespräche bedanken.

Ganz speziell möchte ich mich auch bei Prof. Dr. Thomas Rattei von der Universität Wien für die Unterstützung bei der Bearbeitung von bioinformatischen Fragestellungen, wertvolle Diskussionen und den Zugang zum Computercluster der Division of Computational Systems Biology bedanken.

Zu guter Letzt möchte ich mich auch bei all meinen Freunden und StudienkollegInnen für die schöne Zeit in Graz bedanken.

Abstract

Listeria monocytogenes is a foodborne pathogen causing severe infections like sepsis and meningitis, with mortality rates of about 30% in mainly immune-compromised individuals. Although in many European countries extensive screening is performed to prevent large outbreaks, the causative food source in most sporadic cases could not be identified. To find correlating isolates for clinical cases and food samples, the gold standard method pulsed-field gel electrophoresis (PFGE) was shown to be not discriminative enough. The emergence of next generation sequencing allows new phylogenetic methods, like core genome multi locus sequence typing (cgMLST) to evolve. Those techniques allow investigating genomic distances between isolates and strains. PFGE was used for the last decade and large databases exist for *L. monocytogenes* in reference laboratories, while the availability of whole-genome sequencing data for comparison is thus far rare. In this thesis, parts of the historic "Special Listeria Culture Collection" (SLCC) were re-cultivated in order to obtain genome sequences of 40- to 95-years old isolates, with the eldest being preserved since 1921. 484 SLCC isolates were successfully revitalized and 192 of them were chosen for whole-genome sequencing. Additionally all Austrian clinical isolates between 2012 and 2015 and 84 isolates with a similar PFGE pattern as an ongoing outbreak in Germany were sequenced. In summary, a total of 501 genomes were sequenced and subsequently analyzed by cgMLST. To create maximum information for cgMLST, different data processing strategies *i.e.* assembling methods were compared. For the data produced in this thesis, the assembly engine SPAdes returned the most reliable assembly results. Phylogenetic trees were generated using single nucleotide polymorphism based analysis and cgMLST and verified the information drawn from cgMLST. In the near future, the large genome dataset generated during this thesis of isolates covering almost a timespan of a century, may allow interesting insights into the recent evolution of *L. monocytogenes*.

Kurzfassung

Listeria monocytogenes ist ein Erreger, der durch kontaminierte Lebensmittel verbreitet wird. Er kann schwere Infektionen wie Sepsis und Meningitis mit Mortalitätsraten von etwa 30% verursachen und führt in erster Linie bei immungeschwächten Personen zu Listeriose. Obwohl in vielen europäischen Ländern ein umfangreiches Screening durchgeführt wird, um Ausbrüche zu verhindern, kann die Ursache in den meisten sporadischen Fällen bisher nicht ermittelt werden. Isolate aus klinischen Fällen und Lebensmittelproben werden dazu bisher mittels der Standardmethode „Pulsfeld- Gel-Elektrophorese" (PFGE) untersucht, jedoch wurde gezeigt, dass diese keine ausreichende Tiefe zulässt, um Ausbrüche vollständig aufzuklären (Salipante *et al.*, 2015). Der Fortschritt in der DNA-Sequenziertechnologie ermöglichte es, die neue Methode *core-genome multi locus sequence typing* (cgMLST) zu entwickeln, mit welcher der genomische Abstand zwischen Isolaten detailliert untersucht werden kann. Im Rahmen dieser Arbeit wurde die historische „Special Listeria Culture Collection" (SLCC) teilweise wiederbelebt, um Genomsequenzen der 40- bis 95-Jahre alten Isolate zu erhalten. Das älteste Isolat aus dem Jahr 1921 wurde dabei in der Deutschen Sammlung von Mikroorganismen und Zellkulturen hinterlegt (DSM 101804, Hyden *et al.*, 2016). Insgesamt wurden 484 SLCC Isolate erfolgreich rekultiviert und 192 von ihnen Ganz-Genom sequenziert. Zusätzlich wurden alle österreichischen Humanisolate zwischen 2012 und 2015 und 84 weitere Isolate mit einem ähnlichen PFGE Muster eines Ausbruchs in Deutschland sequenziert. Insgesamt wurden 501 Genome generiert, die anschließend durch cgMLST analysiert wurden. Es wurden verschiedene *assembly* Strategien verglichen und die Ergebnisse aus cgMLST miteinander verglichen. Bei dieser Art der Datenverarbeitung erzeugte der *de-novo assembler* „SPAdes" die besten Resultate. Stammbäume wurden mit *single nucleotide polymorphism* basierter Analyse und cgMLST erzeugt, um die Ergebnisse von cgMLST zu validieren. Die im Rahmen dieser Arbeit sequenzierten Isolate, die beinahe die Zeitspanne von einem Jahrhundert abdecken, könnten künftig noch interessante Einblicke in die jüngsten Entwicklungen von *L. monocytogenes* ermöglichen.

Contents

1. Introduction

1.1. *Listeria monocytogenes* relevance as foodborne pathogen

Listeria monocytogenes Murray, 1926 (Pirie, 1940) is a Gram-positive foodborne pathogen, the cause of listeriosis (Farber and Peterkin, 1991). Although it can be found in soil, plants and water, *L. monocytogenes* is commonly acquired by contaminated food (Allerberger and Wagner, 2010), mostly via so-called ready-to-eat – RTE. From the gastrointestinal environment it can invade mammalian cells with intracellular growth, replication and spread to adjacent cells. This way *L. monocytogenes* causes invasive listeriosis, which manifests in severe infections, such as sepsis and meningitis, in immuno-compromised adults (Allerberger and Wagner, 2010). It can also cross the placental barrier and infect fetuses from infected mothers (Disson *et al.* 2008). The disease is often severe in specific high risk groups (*i.e*. immuno-compromised individuals, persons > 60 years old or newborn) with reported mortalities of up to 30 % (Allerberger and Wagner, 2010). In the European Union, 1,763 cases of listeriosis were confirmed in 2013, reporting constantly increasing numbers of cases since 2008 (press release EFSA and ECDC, 2015). Apart from this low incidence the disease is a great burden: in the Netherlands in 2011 *L. monocytogenes* caused € 4.6 to € 9.4 million in damages as the ninth severe foodborne pathogen (Mangen *et al.* 2015). According to Hoffmann *et al.* it causes $ 2.6 billion per year in the US, which is the third-highest burden on public health after *Toxoplasma gondii* ($3.0 billion) and *Salmonella enterica* ($3.3 billion) of food-related pathogens (Hoffmann *et al.* 2012). Both authors evaluated the impact on life quality, Hoffmann *et al.* in quality adjusted life years (QALY) and Mangen *et al.* in disability adjusted life years (DALY), which majorly contributed to the high costs estimated.

In 2014, 47 cases of invasive listeriosis were recorded by the Austrian Agency for Health and Food Security (Huhulescu, 2015). This is an annual incidence of 0.55 per 100,000 inhabitants. The mortality was 26 % (12 of 47) (Huhulescu, 2015). From 1997 to 2014 the number of infections was increasing (figure 1).

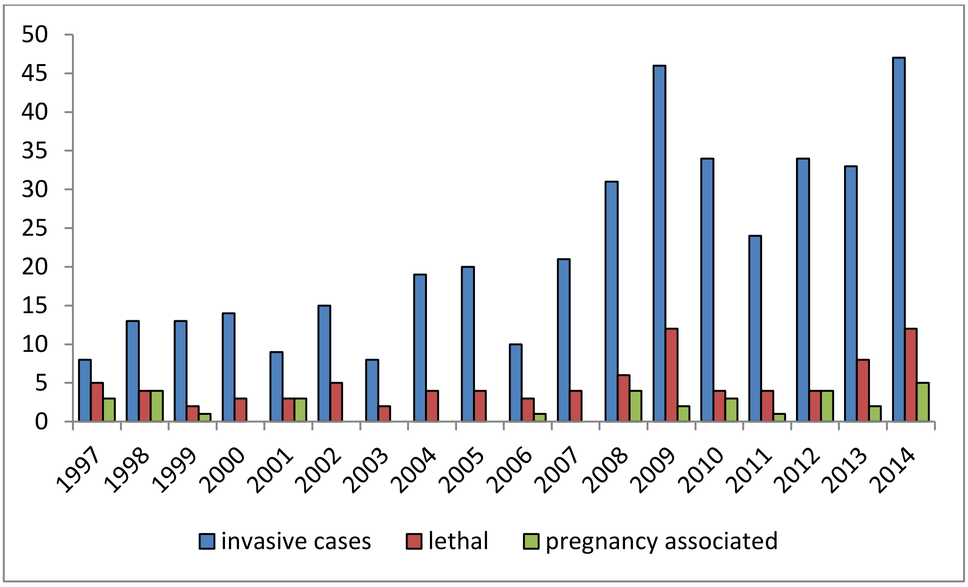

Figure 1: Cases of invasive listeriosis in Austria from 1997 to 2014. Data was taken from the annual report for listeriosis in Austria of 2014 (Huhulescu, 2015).

Unlike most foodborne pathogens *L. monocytogenes* is able to grow at low temperatures, low moisture and high salt concentrations (Gandhi and Chikindas, 2007). It is adapted to survive in conditions of the gastrointestinal tract, such as low pH (Smith *et al.* 2013) and presence of bile salts (Dussurget *et al.* 2002). Some strains are able to form strong biofilms on surfaces important in food industries *i.e.* stainless steel, ceramics and synthetic materials within 24 hours (Doijad *et al.* 2015). Foods which are not heated to 65°C or require storage at refrigerated temperatures, such as cheese or meat, have been reported to be infectious. Biofilm formation and resistance to the commonly used disinfectant benzalkonium chloride (Mullapudi *et al*. 2008) led to persistence in meat processing facilities (reviewed in Carpentier and Cerf, 2011).

Those characteristics, paired with the high severity of infections, make *L. monocytogenes* a great threat for the food industry. From November 2011 to September 2015 277 food recalls were listed by the FDA (see www.fda.gov/Safety/Recalls/) caused by *Listeria monocytogenes*.

There are frequent outbreaks, which are often associated with defective food treatment in ready-to-eat food production (Todd and Notermans, 2011). A selection of reported outbreaks in the last 20 years is listed in table 1.

Table 1: A selection of reported listeriosis outbreaks between 1995 and 2014.

Year(s) of occurrence	**Product(s) involved**	**Country**	**Confirmed cases**	**Mortalities**	**Reference**
1995/96	Cold smoked rainbow trout	Sweden	9	?	Tham *et al.*, 2000
1998/99	Frankfurters, deli meat	USA	108	14 + 4 unborn	Mead *et al.*, 2006
2000	Homemade Mexican style cheese	USA	12 of 16 pregnant, 3 newborn	3 unborn	MacDonald *et al.*, 2005
2005	Soft cheese	Switzerland	10	3 + 2 abortions	Bille *et al.*, 2006
2008	Cheese	Canada, Quebec	38	?	Gaulin *et al.,* 2012
2008	Deli meat	Canada, Ontario	56	21	Ontario Ministry of Health, 2009
2009/10	Acid curd cheese ("Quargel")	Austria, Germany and Czech Republic	25+8+1	5+3+0	Fretz *et al.*, 2010
2010	Hog head cheese, Deli meat	USA	10	?	Centers for Disease Control and Prevention (CDC), 2011
2011	Cantaloupe	USA	147	33 + 1 unborn	McCollum *et al.*, 2013
2013	"Rullepølser", Deli meat	Denmark	38	15	"The Local" online news journal, 2014
2013	Cold meat, deli meat	Sweden	27	4	Online-News "TheForeigner.no," 2014

2014	Caramel apples	USA	35	7	Centers for Disease Control and Prevention (CDC),2015

Large outbreaks are rare. The majority of listeriosis infections is sporadic and their sources remain unknown (Allerberger and Wagner, 2010).

1.2. Special Listeria Culture Collection (SLCC)

Since the introduction of solid media to microbiology, individual microbiologists started to assemble collections of single species strains. While most of them have vanished today, some survived as stab cultures or lyophilisates and contain invaluable treasures. One of those collections was accumulated by Prof. H. P. Seeliger, consisting of isolates between 1921 and 1987. (Haase *et al.*, 2011)

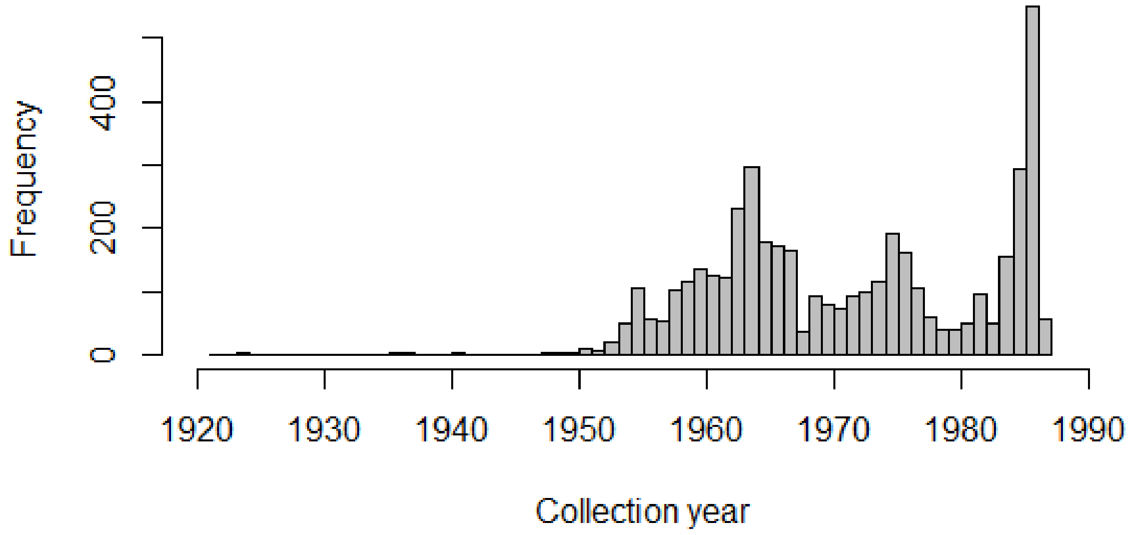

Figure 2: Distribution of the isolates collected by H. P. Seeliger for the Special Listeria Culture Collection (SLCC). Histogram was generated with data from 4,404 isolates revived by Haase *et al.* 2011.

After retiring in 1989, his colleague Prof. H. Hof in Mannheim, University of Heidelberg inherited the so-called “Special Listeria Culture Collection” (SLCC). In 2008 major parts of the collection were moved from Heidelberg, Germany to the Environmental Research Institute, University College Cork, Ireland to Mark Achtman. There, the SLCC was moved from ambient temperature to 10°C for the first time. Detailed information about each strain, which existed in handwritten and typed documents in German before, was translated to English and made available online (Haase *et al.,* 2011). The collection consisted of 11,066 tubes and 8,643 lyophils for a total of 6,451 strains.

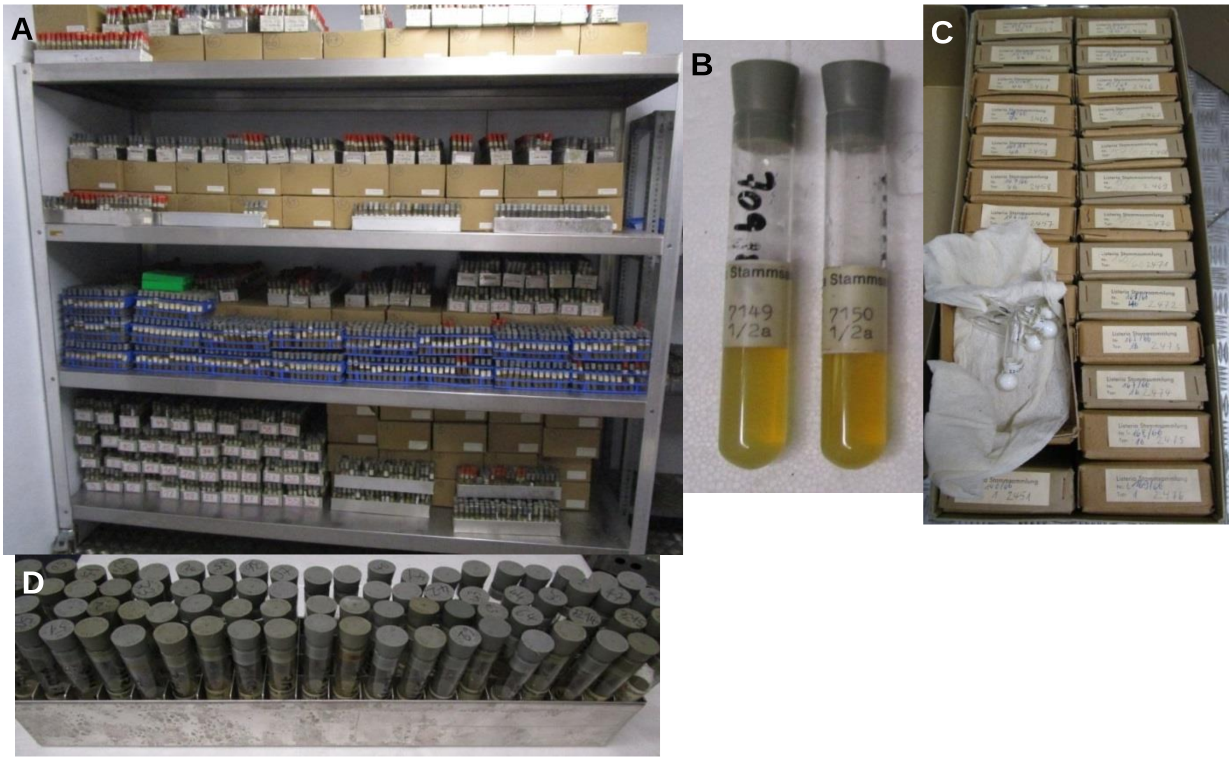

Figure 3: Photographs taken from the current state of original tubes and vials of the Special Listeria Culture Collection (SLCC). A: complete overview of the storage space the SLCC requires. **B:** Glass tubes with stab agar with rubber stops of two different isolates labeled with SLCC number and serotype. **C**: Box packaging of vials with lyophilizate. **D**: Metal rack with 80 stab cultures.

After the study in Cork was finished, the whole collection was taken into the care of the Austrian Agency for Health and Food Safety and transferred to Graz, Austria. From then on, the SLCC was stored in a 4°C refrigerated room in the basement.

In March 2015, I started to re-cultivate several strains to see whether the isolates are still viable after thousands of kilometers of travel, years of storage and nondescripted tries of re-cultivation.

With the re-cultivation and analysis of strains originating in the first half of the 20th century, a great historic treasure of microbiology was studied in more detail for the first time. Our hypothesis was, that it might be possible to answer questions concerning the genetic shift and acclimatization of *Listeria monocytogenes* to the altered environment and human lifestyle.

1.3. Identification of outbreak sources

Before the rise of DNA-based methods for subtyping of bacterial species in the 1990's, typing by phenotypic characteristics was state of the art. Based on antigenic agglutination, serotyping was applied to *Listeria* genus first in 1940 (Paterson, 1940) and further developed by (Donker-Voet, 1959; Seeliger and Hoehne, 1979) to the today valid scheme with 13 different serotypes: 1/2a, 1/2b, 1/2c, 3a, 3b, 3c, 4a, 4ab, 4b, 4c, 4d, 4e and 7.

Although serotyping is still used to describe *L. monocytogenes* strains, it has been complemented by the serogroup PCR, which assigns the serotypes to 4 groups to differentiate quickly between the serotypes 1/2a, 1/2b, 1/2c and 4b, which account for 98 % of all clinical isolates (Doumith *et al*., 2004).

Those methods are useful for first level typing but not for differentiation on the strain level. Different genotyping methods were introduced such as random amplification of polymorphic DNA (RAPD) (Giovannacci *et al.*, 1999), PCR–restriction enzyme analysis (PCR–REA) and genomic macrorestriction using rare cutting enzymes with pulsed-field gel electrophoresis (PFGE). All three methods are based on differently sized DNA products which form distinct bands on electrophoresis and patterns which can be compared for each strain. According to Giovannacci *et al.* RAPD and PFGE have higher discriminative power compared to phenotypic methods (Giovannacci *et al.*, 1999). Hence PFGE was standardized using *Apa*I and *Asc*I enzymes (Graves and Swaminathan, 2001) which made DNA "fingerprints" comparable and became international standard. The European Food Safety Authority (EFSA) developed standard operation procedures which produce comparable results as the US Network "PulseNet" (Michelon *et al.*, 2015), which recently tightened PFGE as the gold standard.

With the advance of DNA sequencing, multi locus sequence typing (MLST) was introduced for *Neisseria meningitidis* (Maiden *et al.*, 1998). The method is based on assigning numbers to each different allele for each locus, while the combination of alleles defines the sequence type (ST). A few years later, this method was applied to *L. monocytogenes* but was found to be less discriminative than PFGE, but suggesting that the results would be easier to compare in different laboratories (Salcedo *et al.*, 2003). Nowadays in the standard procedure for MLST on *Listeria monocytogenes*, seven housekeeping genes are (partially) sequenced: *acbZ* (ABC-Transporter), *bgl*A (beta-glucosidase), *cat* (catalase), *dap*E (succinyl diaminopimelate desuccinylase), *dat* (D-amino acid aminotransferase), *ldh* (lactose dehydrogenase) and *lhk*A (histidine kinase) (Ragon *et al.*, 2008). Each sequence is restricted to the exact length of the reference, insertions or deletions (indels) are not allowed. The MLST database for *Listeria monocytogenes* is available at www.pasteur.fr/mlst.

Multi virulence locus sequence typing (MVLST) was based on six different genes, which were virulence related, but the method was reported to have the same discriminative power as MLST (Zhang *et al.*, 2004). For outbreak investigation, MLST/MVLST were reported to have less discriminative power than PFGE (Haase *et al.*, 2014).

The ongoing evolution of DNA-sequencing technology enables analysis of sequences based on whole-genome sequencing data. Like the MLST scheme the allele-based sequence typing was adopted to whole genomes. Hence a core genome MLST (cgMLST) scheme was created, using most annotated genes of the *L. monocytogenes* strain EGD-e (GenBank accession number NC_003210.1). Accordingly the genes were divided in two groups resulting in a “core genome”, consisting of 1701 genes and an “accessory” gene pool, consisting of 1166 genes (Ruppitsch *et al.*, 2015).

Genotyping based on WGS provides high resolutions and allow quantifying the genetic difference between two isolates. Further DNA fingerprinting methods, like PFGE face serious problems (Achtman, 2008). The methods have to be standardized between laboratories. Although today a standard protocol exists in the European Union (Michelon *et al.*, 2015), PFGE patterns from different experimental protocols are impossible to compare. WGS data is by far easier to distribute and is available for future comparison or different methods, independent from the next-generation sequencing technology used. Further, clusters identified with PFGE were shown to be insufficiently described for outbreak investigation (Lienau *et al.*, 2011).

Apart from strain typing and analysis of human infections, genomic analysis of *L. monocytogenes* isolates found in food products is important to prevent outbreaks and sporadic infections in the future. For most clinical isolates there are no appropriate genetic profiles known from food isolates thus far. A complete elucidation is almost impossible for single cases of listeriosis today (Allerberger and Wagner, 2010).

1.4. Next generation sequencing

Phylogenetic studies based on sequence alignment or single nucleotide polymorphisms (SNPs) were performed increasingly with the availability of gene sequences. The first bacterial genome was completely sequenced in 1995 (Fleischmann *et al.*, 1995), while the first *L. monocytogenes* genome was published in 2001 (Glaser *et al.*, 2001). In the early 2000's, only sequences of highly conserved regions were used for SNP-based genotyping, *e.g.* the *sigB* gene was used to distinguish between the lineages of *Listeria monocytogenes* (Moorhead *et al.*, 2003). Whole-genome comparison was used quite early to describe general differences between lineages or close related species (Glaser *et al.*, 2001) or to explore characteristics of clinical versus laboratory strains of *Mycobacterium tuberculosis* (Fleischmann *et al.*, 2002). The bottleneck was the need of complete genomes which acquisition was both laborious and cost intensive.

Until 2008, Sanger-based capillary sequencing was the cheapest DNA sequencing method on the market, costing at least $ 500 per Million of bases (Mega basepairs, Mb) raw data sequenced (Wetterstrand, 2015). These 500-600 bp long sequences needed a minimum coverage of six, for one single *L. monocytogenes* genome the cost would be about $ 9,000. As shown in figure 4, the advance and application of next generation sequencing (NGS) technologies have reduced the costs of raw sequencing data dramatically (Metzker, 2010). Though, the read length of sequences obtained was reduced, resulting in higher coverage required for assembling the genome (Wetterstrand, 2015).

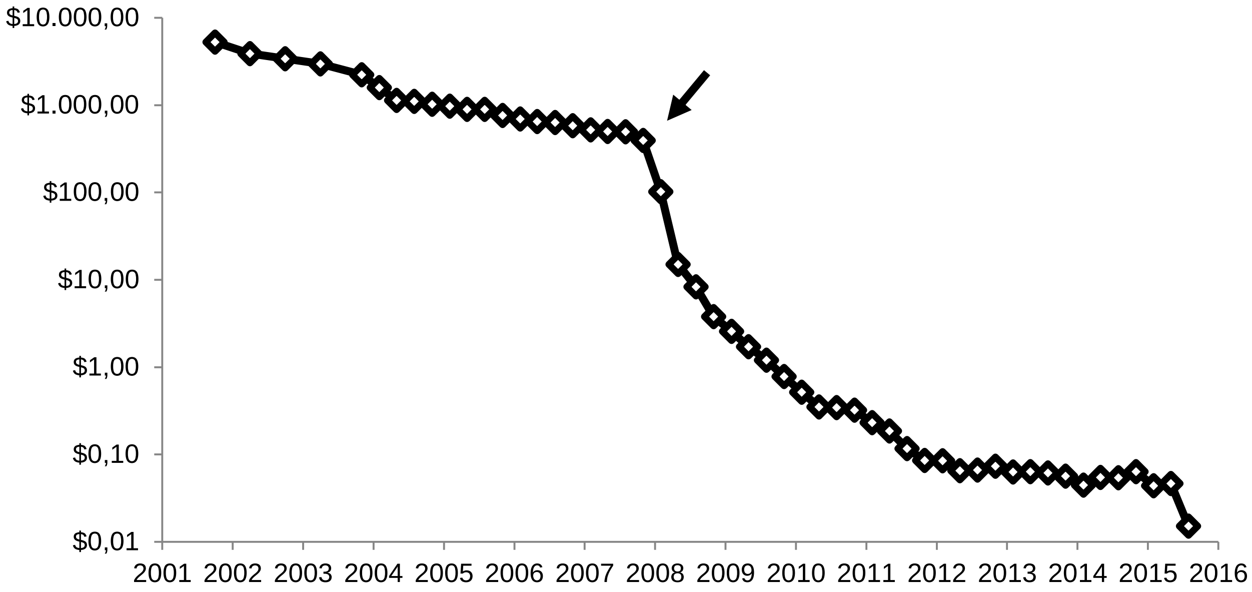

Figure 4: Decreasing costs per raw Mb sequencing in USD over the past fifteen years. The figure is derived from data collected by the National Human Genome Research Institute (NHGRI) accessible at the website www.genome.gov/sequencingcosts/. (Wetterstrand, 2015)

One of the first method used for DNA sequencing was based on radioactively labeled dideoxynucleotides which discontinued in vitro DNA-replication via polymerase (Atkinson *et al.*, 1969). The fragments were separated by gel electrophoresis. Nowadays known as Sanger sequencing, the method was published as “chain termination method” and was capable of generating sequences up to a few hundred bases (Sanger *et al.*, 1977). The technique has been optimized by using fluorescent dyes for termination, automated laser detection and capillary electrophoresis. At maximum a sequencing device with 384 parallel capillaries was produced, which could be “utilized for massively parallel genetic analysis” (Emrich *et al.*, 2002). In only nine months continuous work of 65 people, 27.2 million reads in mate-pairs were sequenced on a Sanger capillary sequencer for the first human genome, which had 14.8 billion base pairs in total (Venter *et al.*, 2001).

For the next-generation of DNA sequencing, a large variety of sequencing technologies were developed. The major technologies are 454 pyrosequencing, distributed by Roche, reversible termination sequencing by synthesis (Illumina), sequencing by ligation (Applied Biosystems) and semiconductor sequencing, also known as IonTorrent (Thermo Fisher) (Hodkinson and Grice, 2015).

454 pyrosequencing was the first of those methods that was commercially available in 2004 (Stein, 2008). Pyrosequencing is based on a real-time reaction of pyrophosphate,

which is released during elongation of DNA replication and used by a luciferase to produce a light signal (Ronaghi *et al.*, 1998). The technology has been optimized over the last decade. By the end of 2015 there were three devices available on the market (Roche Website, 2015). Two benchtop machines GS Junior/Junior+ capable of producing sequences around 400/700 bp read length and total yields of 35/70 Mb in 10/18 hours runtime. And the newest version of GS FLX+ with up to 1,000 bp read length and 700 Mb in 23 hours runtime (Roche Website, 2015). The technology produces over 99 % of the reads in with less than 0.1 % error rate, although pyrosequencing was reported to have problems with homopolymer regions (Huse *et al.*, 2007). In 2013 it was revealed that Roche was closing 454 Life Sciences and the production of sequencers using this technology would end in 2016 (Bio-IT World Website, 2013).

Sequencing by ligation “SOLiD” was also an early NGS technology beginning in 2007 (Stein, 2008), which allowed up to 75 bp read length. The technology was commercialized by Applied Biosystems, which merged with Invitrogen in 2008 into the company “Life Technologies” (San Francisco Business Times, 2008), which itself was taken over by Thermo Fisher in 2014 (Berkrot, 2013). The latest device using SOLiD technology was presented in 2010: the 5500xl W allows up to 320 Gb yield in paired end 2x 50 bp mode (Thermo Fisher Website, 2015).

Life Technologies also acquired the semiconductor sequencing technology called IonTorrent in 2010 (GenomeWeb Website, 2010). IonTorrent uses semiconductors detecting H^+-ions produced by incorporation of desoxynucleotides (dNTPs) during DNA-replication. Instead of using light for detection, ion-sensitive field-effect transistors (ISFET) in integrated circuits directly measure a change in electric potential (Rothberg *et al.*, 2011). The IonTorrent Personal Genome Machine (PGM) was released in 2010 (Rusk, 2011) as the first benchtop sequencer for mid-scale sequencing projects with low entry costs (Pollack, 2011). The Ion PGM allows 200 or 400 bp reads with a maximum expected output of up to 2 Gb in only 4.4 hours produced by the latest chipset the “Ion 318 Chip v2”. Larger versions available are the Ion Proton (200 bp read length, 10 Gb maximum output) and the Ion S5, which promises 10-15 Gb within 24 hours from DNA to sequence data (thermofisher.com, 2015). Like the pyrosequencing technology, multiple nucleotides of the same type can be incorporated in the same cycle, which leads to serious homopolymer detection problems (Bragg *et al.*, 2013).

The currently leading technology with a market share of about two thirds is distributed by Illumina (Steinbock and Radenovic, 2015; Karow, 2015). Initially invented by Solexa, the method is based on sequencing by synthesis: on a solid surface called “flow cell” DNA molecules are locally amplified to clusters. All DNA strands are further elongated by a fluorescently labeled nucleotide, while for each nucleotide a different marker is in use. The fluorophores are excited by two different light sources to detect the base of each cluster. The labeling also acts as a reversible chain terminator and is cleaved after imaging. The cycle of synthesis, imaging and cleavage is repeated in the flow cell depending on the desired read length up to 300 bp (Morey *et al.*, 2013).

A special feature of this technology is the ability to sequence each cluster in a paired-end mode. After sequencing the first strand, the second DNA strand is completely synthesized and can further be sequenced beginning from the 5’ end. This was important for Illumina reads to compete with other technologies in the beginning, when only 36 to 75 bp could be sequenced (Fullwood *et al.*, 2009). Today paired-end sequencing is a huge advantage, allowing up to 2x300 bp of both ends of DNA fragments up to 1,500 bp long (Illumina MiSeq User Guide, 2015). Illumina’s sequencers in the high-throughput segment are unchallenged with the HiSeq X Ten allowing up to 1,800 Gb output with one single run (Illumina Website, 2015). The benchtop versions MiSeq/MiSeq Dx hold the highest market share in the benchtop segment (Karow, 2015) and generate the most accurate reads compared to competing IonTorrent PGM and 454 Junior (Loman *et al.*, 2012). Systematic errors arise from PCR amplification where genomic regions with extreme GC content are under-represented (Aird *et al.*, 2011).

All of these NGS methods use PCR amplification before sequencing. Using a large group of identical sequences allows high accuracy, but has a drawback: PCR limits the maximum sequencing length and introduces bias (Aird *et al.*, 2011). The third generation of sequencing will abandon amplification prior to sequencing and is based on single molecule sequencing. The first method developed, which used an immobilized polymerase as a real time sequencing engine in combination with Zero Mode Waveguide technique to detect incorporation of nucleotides showed accuracy of 75 to 85 % (Korlach *et al.*, 2008, Eid *et al.*, 2009).

The biggest advantage using this technology distributed by Pacific Bioscience's launched in 2010, might be the long reads (Rusk, 2011). The initial device of PacBio allowed around 1 kb long reads, while the now available PacBio RS II and the PacBio Sequel promise averages of 15 kb and up to 60 kb long sequences. In comparison to Illumina the sequencing costs are more expensive and create lower yields (7 Gb compared to 1,800 Gb). Still long read sequencing technologies are promised to have a bright future (opiniomics.org, 2015).

A different way of single molecule detection is the nanopore technology. Single stranded DNA is ratchet though a pore in a thin layer in nanometer dimensions, giving electrically measurable signals. Protein pores in lipid biolayers, fabricated solid-state or plastic materials were used as nanopores (Branton *et al.*, 2008). Oxford Nanopore Technologies (ONT) is using biological pores in solid-state polymer layers (Bayley, 2015) and is about to launch sequencers in different scales: the MinION has about the size of a cellphone, the PromethION as the benchtop sequencer and the GridION as the largest version (nanoporetech.com, 2015). Currently the ONT MinION is the only version available (Mikheyev and Tin, 2014) in a limited access program with a reported accuracy between 75-85 % (Jain *et al.*, 2015). Apart from the revolutionary technique, the MinION is most innovative by being the first portable sequencer (Steinbock and Radenovic, 2015).

Although the third generation sequencing methods for now are behind in terms of accuracy, the long read length compared to NGS high throughput data might solve problems in whole-genome analysis produced by incomplete draft assemblies today (Rhoads and Au, 2015).

1.5. Extensive bioinformatics is needed for sequencing data handling

After DNA sequencing was established, the demand for sequences larger than read length grew rapidly. The 48,502 bp large genome of bacterophage λ was the first virus to be sequenced completely (Sanger *et al.*, 1982). The genome was shattered mechanically and enzymatically and cloned into vectors to amplify the sequences. For this random sequencing, methods had to be developed to re-assemble the genome (Staden, 1979). The genome of *Haemophilus influenzae* was the first bacterial genome published (Fleischmann *et al.*, 1995), using a random shotgun sequencing approach and the software tool TIGR for assembly. Obstacles as repeat regions, chimeras and sequencing errors were pointed out and dealt with (Sutton *et al.*, 1995).

With automatic sequencers and increasing throughput alternative base calling software and quality scores were introduced. The software tool *phred* used log-transformed error rates today applied for every sequencing result:

$$q = -10 \times log_{10}(p)$$

In this formula q is the phred- or quality-score and p the estimated error probability for the base (Ewing *et al.*, 1998).

Also the need for efficient assembling algorithms for short reads, in comparison to Sanger-sequenced reads, increased. *De-novo* assembly is possible when reads overlap, therefore a sufficient amount of oversampling is needed. In general this is referred to as "coverage" of a genome assembly. Also, in a random sequencing approach the average coverage should be considered to ensure most of the genome is represented in the reads (Fleischmann *et al.*, 1995). Shorter reads require higher coverage while a larger amount of reads increases computational resources needed (Miller *et al.*, 2010). Also short NGS sequences result in lower assembly qualities, which can be observed as gaps being longer and more frequent (Chaisson *et al.*, 2004). Paired-end reads facilitate assembly and help to resolve repeat regions. Reads are considered to be paired, when the relative orientation and the separation on the target is known. This additional information can be exploited by assembler software. The assembler software groups reads into contiguous sequences ("contigs"), which are grouped by the use of paired-end reads to scaffolds. The consensus sequence of contigs is output as FASTA file (Miller *et al.*, 2010). The methods behind assemblers are mainly graph based: the overlap-layout-consensus (OLC) assembly, de-Bruijn graphs (DBG) or greedy algorithms. A graph is an abstraction of nodes and edges, a collection of edges creates a path.

Basically OLC assemblers are used to compare all reads pairwise and determine overlaps by alignment. The graph is formed by reads representing nodes and overlaps being edges. Then paths are calculated, which return contig consensus sequences. Obviously for pairwise alignment of reads the, first step for the OLC algorithm can be computationally expensive for next generation sequencing data consisting of billions of short reads (Li *et al.*, 2010). Examples for OLC assemblers are: TIGR (Sutton *et al.*, 1995), Celera (Miller *et al.*, 2008), ARACHNE (Batzoglou *et al.*, 2002) and Newbler (Margulies *et al.*, 2005).

Assemblers based on DBG initially generate short sequences called “k-mers”, with equal length from each read, *i.e.* a sequence AGGTCT has 4 k-mers with the length k=3: AGG, GGT, GTC and TCT, or 2 k-mers with k=5: AGGTC and GGTCT. K-mers overlapping by “k minus one” bases are connected by edges. The graph is solved by a Hamiltonian cycle (Compeau *et al.*, 2011), which visits each k-mer exactly once and creates the minimal length genome. In most assemblers the graph is simplified by different tasks and removes certain nodes and edges. The application of DBG in genome assembly was pioneered by Pevzner *et al.* (2001) developing the first DBG assembler named Euler. A large variety of assemblers are available for short or long NGS reads. Examples for DBG assemblers are: Velvet, SPAdes, Ray, ALLPATHS-LG and IDBA.

The Velvet assembler was designed to use very short Illumina or SOLiD reads with a maximum k-mer size of 31 (Zerbino and Birney, 2008). By adjusting the “MAXKMERLENGTH” compilation parameter of the open source code and using the Perlscript VelvetOptimiser (available on bioinformatics.net.au/software.velvetoptimiser.shtml) the assembler can be run iterative automatically using different k-mer values and returning the best assembly.

The SPAdes assembler (Bankevich *et al.*, 2012) was developed for bacterial single-cell and multi-cell assembly. SPAdes uses iteratively several k-mer values. It is constantly updated to support the newest NGS technologies and is taking advantage of different read correction tools (available under bioinf.spbau.ru/spades).

Ray (Boisvert *et al.*, 2012) is an assembler which makes use of multiple read libraries of different NGS brands simultaneously. The process is scalable and one single assembling can be run on up to 100,000 CPU cores and allows Meta genome *de-novo* assembly in the Ray$_{\text{Méta}}$ version.

The ALLPATHS-LG assembler was designed to produce high quality assemblies of large mammalian genomes with default parameters (Gnerre *et al.*, 2011). Hence, it requires at least a “fragment library” *i.e.* Illumina, 454 or IonTorrent short read library and a jumping library e.g. mate-pair Illumina reads. Adding a PacBio read library to the two Illumina read libraries, the assembler can produce almost complete bacterial genomes with a low number of mismatches/SNPs (Ribeiro *et al.*, 2012; Liao *et al.* 2015). Alternatively ONT MinION reads where shown to result in equal good assemblies as PacBio reads (Sovic *et al.* 2015).

The Maryland Super-Read Celera Assembler (MaSuRCA) creates extended reads by using read pair information, called “super reads”. Those super reads are assembled using a modified version of CABOG, which itself is a modification of the Celera assembler. It is said to perform as good as ALLPATHS-LG with solely Illumina data (Zimin *et al.*, 2013).

Many assemblers have been compared in the past seeking for the best assembler for certain applications and critical assessment of assemblies. The “Assemblathon” (assemblathon.org) is a competition where sequencing data is provided to teams which use their developed assembler. For the “Assemblathon 1” a bacterial genome (Earl *et al.*, 2011) and the “Assemblathon 2” three vertebrate genomes were assembled and compared (Bradnam *et al.*, 2013). The third Assemblathon contest is currently running. The first two competitions and the “Genome Assembly Gold –standard Evaluation” (GAGE) paper (Salzberg *et al.*, 2012) showed that the ALLPATHS-LG and SOAPdenovo returned the best assemblies comparing the NG50 value, which measures both gene content and contig size. GAGE-B, being the sequel of GAGE used 250x2 bp and 100x2 bp paired-end Illumina reads of 5 different bacterial species showed that the MaSuRCA assembler resulted in the largest N50 values for most experiments. In a few cases, SPAdes and SOAPdenovo were equal or even better than MaSuRCA, while ALLPATHS-LG needed a jumping library and was therefore not applicable (Magoc *et al.*, 2013). The most recent comparison paper assessed the performance of *de-novo* assemblers on IonTorrent PGM data and Illumina MiSeq data in different lengths. Therefore, only assemblers allowing single-end read data were used which excluded ALLPATHS and MaSuRCA. The results indicated, that MiSeq data is more robust on the choice of assemblers, while IonTorrent data clearly benefited from error correction tools utilized in certain assembler pipelines e.g. BayesHammer in SPAdes (Jünemann *et al.*, 2014).

Each of the assembler comparison studies discussed that there is no universal assembler who performs best for every dataset. Jünemann *et al.* (2014) reported that the choice of the assembler depends on the NGS platform, the demands on assembly quality and the available computational resources.

Apart from *de-novo* assembly, assemblies can be produced using a reference genome – if possible. Reference mapping, also referred to as re-sequencing was especially useful with very short reads in the early days of Illumina and SoLID NGS data of up to 50 bp in length and was used for variant calling as reviewed by Bentley (2006). Apart from the right software a closely related complete reference genome is needed (Trapnell and Salzberg, 2009).

The major problem for *de-novo* assemblers are large repetitive regions, which are larger than the read-length. In prokaryotic genomes the largest global repeat is the rDNA operon, which is about 7 kb long (Treangen *et al.*, 2009). To overcome this problem and produce nearly complete genomes it would be necessary to have long read libraries like Pacific Biosciences sequence data (Bashir *et al.*, 2012) of reads > 7,000 bp. While short read shotgun sequencing produces data with very high accuracy, long read sequencing technologies have only 80 % at maximum. Using assemblers, which are able to cope with a combination of long reads and short reads, may produce almost complete genomes (Koren and Phillippy, 2015). It was reported, that such a procedure would be affordable for less than $ 1000 (Koren *et al.*, 2013).

2. Objectives

Listeria monocytogenes is a serious threat to modern food industry (Carpentier and Cerf, 2011). The Gram-positive bacterium is able to grow at low temperatures, form biofilms and may cause severe infections in humans. Nowadays the infection source for most individual clinic cases is unknown, although different typing methods are applied to food isolates routinely. With the advance of next-generation sequencing, the costs for in-house sequencing became affordable and can be used to challenge old-fashioned highly sophisticated genotyping methods. The Austrian Agency for Health and Food Safety (AGES) starts to use whole-genome sequencing (WGS) of *L. monocytogenes* for genotyping.

The aim of this thesis was to partially revive historic *L. monocytogenes* isolates from the Special Listeria Culture Collection. 100 of the oldest strains and 200 human isolates collected in Austria or Germany of different time periods should be analyzed by WGS. In order to compare revitalized isolates to recent clinical isolates, about 200 additional isolates collected between 2008 and 2015 in Austria should be sequenced using a benchtop Illumina MiSeq sequencer. The massive output of generated WGS data should be analyzed by a novel genotyping method named core genome multi locus sequence typing (cgMLST). Therefore I aimed to validate the method by comparing it to single nucleotide polymorphism (SNP) based methods and to optimize the data processing procedure for *L. monocytogenes* sequence data. Also, WGS of a selection of historic strains might allow identifying potential changes in *L. monocytogenes* genomes.

3. Material and methods

3.1. Re-Cultivation of SLCC stab cultures

The isolates were selected to find the oldest possible isolates. Additionally isolates were selected for human isolates from Germany between 1954 and 1970. Information about the strains' collection years and sources were taken from an Excel-sheet published along with the study by Haase *et al.* (2011). Serotype information was taken from the test tubes' labels.

The stab cultures were rinsed with 2-3 ml tryptose phosphate broth (TPB) and small pieces of agar were transferred with the media to sterile 15 ml tubes (Falcon). The liquid culture was incubated at 37°C for up to 7 days. When the media was turbid, the suspension was streaked to Columbia agar with 5 % sheep blood (COS, bioMiereux). Single beta-hemolytic colonies were streaked to the selective RAPID'L.mono Agar (RLM, Bio-Rad). Colonies which grew in blue color, typical for *L. monocytogenes*, were streaked to COS agar and glycerol stocks (CryoBank Yellow, VWR) thereof were prepared and stored at -80°C

3.2. Serogroup determination by multiplex PCR

Multiplex PCR was performed as previously described by Doumith *et al.* (2004). Sample DNA was extracted with Chelex®-100 (Bio-Rad Laboratories) previously described by Walsh *et al.* (1991). The primers were obtained from TIB MOLBIOL GmbH in a concentration of 100 µM. The enzyme-nucleotide mix "Muliplex PCR Kit" (QIAGEN) was used and an Eppendorf MasterCycler® EP S as thermocycler. Gel electrophoresis was performed using 2 % agarose E-Gel® with E-Base™ electrophoresis device (Invitrogen). The gels were documented using a Gel Doc XR+ (Bio-Rad Laboratories).

3.3. Genome isolation and sample preparation for sequencing

Approximately 9*10^{10} cells were re-suspended in 160 µl P1 buffer (QIAGEN) and 20 µl of lysozyme (100 mg/ml) was added. The suspension was incubated on a thermocycler at 37°C and 900 rpm for 1 hour. For isolation of high-molecular weight genomic DNA the HMW MagAttract Kit (QIAGEN) was applied following the kits protocol.

Quantification of the DNA was performed utilizing the Qubit® system (Life Technologies). The Qubit® reagent specifically for double-stranded DNA (dsDNA) was used, which binds to dsDNA and is measureable by fluorescence in the Qubit® fluorometer. 20 µl of each sample were added to 180 µl working solution and incubated for 2 minutes at room temperature. The fluorometer calculated the dsDNA concentration automatically.

DNA was diluted to 0.2 ng/µl using a total of 1 ng genomic DNA in 5 µl. For library preparation for multi-sample sequencing, the Illumina Nextera XT DNA Library Preparation Kit was used. The high molecular weight DNA strands were simultaneously tagged and fragmented by specific enzymes. In this step each DNA strand receives the sequence complementary to the sequencing primers. Further the DNA fragments were amplified with different primers to add 8 bases long indices at the 5' and 3' ends. This allows up to 96 samples per sequencing run with 12 x 8 indices. For *L. monocytogenes* 32 (4 x 8) or 48 (6 x 8) samples were sequenced in a single run. After amplification, the samples were normalized and then pooled to allow an optimal balance in sample composition in the final mix. The Illumina Nextera XT Indices Kit (24 indices, 96 samples) was used as listed in table 2.

Table 2: Indices used in the Illumina Nextera XT DNA Library Preparation Kit for multiplexing

Index i7 adapter	Sequence	Index i5 adapter	Sequence
N701	TAAGGCGA	S517	GCGTAAGA
N702	CGTACTAG	S502	CTCTCTAT
N703	AGGCAGAA	S503	TATCCTCT
N704	TCCTGAGC	S504	AGAGTAGA
N705	GGACTCCT	S505	GTAAGGAG
N706	TAGGCATG	S506	ACTGCATA
N707	CTCTCTAC	S507	AAGGAGTA
N708	CAGAGAGG	S508	CTAAGCCT
N709	GCTACGCT		
N710	CGAGGCTG		
N711	AAGAGGCA		
N712	GTAGAGGA		

3.4. Next generation sequencing on Illumina MiSeq

For every sequencing-run on the Illumina MiSeq a sample sheet was generated. The comma separated value file (CSV) was created using the Illumina Experiment Manager v1.9 by choosing the library preparation kit and inserting every sample ID with the correspondent indices used. Further the application “Fastq only” was selected. With the prefilled Illumina MiSeq Reagent Kit v3 around 600 sequencing cycles are possible, therefore 301 and 301 cycles were used in a paired-end sequencing mode. Additionally the 8 bp long indices were sequenced. In total 618 sequencing cycles were run. The reads were automatically de-multiplexed and the sequencing adapter and indices were removed with default parameters. For every sample two FASTQ-files were generated with the raw reads by the MiSeq on board software bcl2fastq allowing a single mismatch in the index sequences.

3.5. Software tools used

Apart from standard text processing and calculation software, specialized software tools were used for different tasks. Their versions and references are shown in table 3.

Table 3: Software tools used in the entire work are sorted in alphabetic order. License refers to either C (Commercial) or OS/F (Open Source/Free)

Name	Version	License	Citation
Illumina Experiment Manager	1.9	C	illumina.com
FastQC	0.11.3	OS	bioinformatics.babraham.ac.uk/projects/fastqc/
BWA	0.7.12	OS	Li and Durbin, 2009
Velvet	1.1.04	OS	Zerbino and Birney, 2008
SAMtools	1.2	OS	Li, 2011
SPAdes	3.5.0	OS	Bankevich *et al.*, 2012
Trimmomatic	0.32	OS	Bolger *et al.*, 2014
SeqSphere$^+$	2.4	C	ridom.com/seqsphere/
RAxML	8.2.2	OS	Stamatakis, 2015
Rstudio	0.99.446	OS/C	rstudio.com
VarScan	2.3.6	OS	Koboldt *et al.*, 2012

MaSuRCA	2.3.2	OS	Zimin *et al.* 2013
ITOL	3.0	F	Letunic and Bork, 2011
DeCypher CLI, Tera-BLAST^{TM}	8.7.0.4	C	timelogic.com

3.6. Quality assessment of the raw reads

For quick quality estimation of the sequencing run, FastQC was run.

The unprocessed reads were mapped using BWA-mem and the *de-novo* assemblies generated by Velvet as the required reference. The insert length information was collected using the standard error output for paired-end libraries. The output was collected and average insert sizes were calculated. Visualization was performed using Rstudio.

3.7. Contamination quantification

Whole-genome sequencing data were mapped to *de-novo* assemblies using BWA-mem. Non-mapped reads were collected using SAMtools. Both *de-novo* assembled contigs and non-mapped reads were compared to the NCBI nucleotide database using Tera-BLAST^{TM} with parameter settings as following: word-size 11, extension threshold 20, nucleic match 2, nucleic mismatch -3, open penalty -5, extend penalty -2, as database a local copy of the NCBI nucleotide database downloaded on 23th December 2014 was used. For each read, the top hit with over 95 % identity was taken and the reads were sorted by organism. The *de-novo* assembly BLAST results were filtered by organism. Reads mapped to regions of contigs with high identity to sequences belonging to other genus than *Listeria* were collected using SAMtools.

3.8. Sequences from the European Nucleotide Archive (ENA)

Apart from my sequencing effort, raw read files of 73 sequenced isolates of a previous study (Ruppitsch *et al.*, 2015) were downloaded from the European Nucleotide Archive (ENA). The isolates were sequenced paired-end on an Illumina Miseq with read-length of 250 bp at the University of Münster. The isolates and ENA accession numbers are listed in appendix table 2.

3.9. SeqSphere+ assembly and mapping

The all-in-one solution package Ridom SeqSphere$^+$ was used as the primary sequence analyzing software regarding MLST/cgMLST. The software supports iterative multiple *de-novo* assembling using Velvet (Zerbino and Birney, 2008) which uses automated k-mer estimation and heuristics. Also including the option for previous quality trimming, default settings were used as following: minimum average of Q30 in a window of 20 bp.

3.10. Different assembly methods

For pre-processing, the reads were trimmed using TRIMMOMATIC (Bolger *et al.* 2014). The function *ILLUMINACLIP* removed remaining adapter or index sequence fragments, so-called technical sequences, using the NexteraPE sequences included. A maximum seed mismatch of 2, the palindrome clip threshold of 30 and a simple clip threshold of 15 were set as parameters. The *SLIDINGWINDOW* function was used to trim all reads to the minimum average quality of Q30 in a range of 20 bp. The unpaired reads were concatenated to one FASTQ file and used further as additional single end library for SPAdes and BWA mapping.

SPAdes (Bankevich *et al.* 2012) was run using the options *careful* and *assembly-only*. As recommended for Illumina reads with lengths greater than 150 bp the k-values 21, 33, 55, 77, 99 and 127 were used.

MaSuRCA (Zimin *et al.* 2013) was run using a *jellyfish hash* of 100,000,000. All other parameters in the configuration file were left as default setting, using the trimmed paired reads as paired-end library.

The read files were mapped to the assembly FASTA files using BWA-mem. Each file was converted to BAM format and sorted using SAMtools functions *view* and *sort*.

The choice of the right complete genome is most crucial for SNP calling (Trapnell and Salzberg, 2009). The best reference was calculated by using a multi sequence FASTA-file which consists of 57 *Listeria monocytogenes* genomes as reference for read mapping (table 4). Plasmid sequences were removed from each fasta file and only the chromosome of each strain was used for mapping. The read files were mapped using

BWA-mem algorithm untrimmed and down-sampled to 18,000 read pairs (approximately 10 million base pairs). The resulting BAM file was sorted and converted to a pileup file using SAMtools *mpileup*. A Python script was written to count the bases and coverage for every reference genome. The reference with the highest number of bases covered was used to map the pre-processed reads.

Table 4: Overview of all reference sequences used to find the most appropriate.

Reference ID	Strain name	Accession No.
1	EGDe	NC_003210.1
2	08-5578	NC_013766.2
3	08-5923	NC_013768.1
4	SLCC5850	NC_018592.1
5	SLCC2755	NC_018587.1
6	SLCC2372	NC_018588.1
7	SLCC7179	NC_018593.1
8	SLCC2540	NC_018586.1
9	SLCC2479	NC_018589.1
10	HCC23	NC_011660.1
11	L99	NC_017529.1
12	07PF0776	NC_017728.1
13	Clip80459	NC_012488.1
14	J1-220	NC_021830.1
15	J1816	NC_021829.1
16	L312	NC_018642.1
17	LL195	NC_019556.1
18	SLCC2376	NC_018590.1
19	ATCC19117	NC_018584.1
20	SLCC2378	NC_018585.1
21	SLCC2482	NC_018591.1
22	M7	NC_017537.1
23	FSL R2-561	NC_017546.1

24	10403S	NC_017544.1
25	Finland 1998	NC_017547.1
26	J0161	NC_017545.1
27	R479a	NZ_HG813247.1
28	J2-031	NC_021837.1
29	CFSAN007956	NZ_CP011397.1
30	F2365	NZ_CP011397.1
31	R2-502	CP006594.1
32	C1-387	NC_021823.1
33	J2-064	NC_021824.1
34	J2-1091	NC_021825.1
35	N1-011A	NC_021826.1
36	J1776	NC_021839.1
37	J1817	NC_021827.1
38	J1926	NC_021840.1
39	WSLC1001	NZ_CP007160.1
40	WSLC1042	NZ_CP007210.1
41	6179	NZ_HG813249.1
42	EGD	NC_022568.1
43	NE dc2014	NZ_CP007492.1
44	CFSAN006122	NZ_CP007600.1
45	Lm60	NZ_CP009258.1
46	NTSN	NZ_CP009897.1
47	IZSAM Lm hs2008	NZ_CP010346.1
48	N2306	NZ_CP011004.1
49	CFSAN008100	NZ_CP011398.1
50	L2074	NZ_CP007689.1
51	L1846	NZ_CP007688.1
52	L2625	NZ_CP007687.1
53	L2624	NZ_CP007686.1

54	L2676	NZ_CP007685.1
55	L2626	NZ_CP007684.1
56	CFSAN023463	NZ_CP012021.1
57	LM850658	CP009242.1

3.11. Core genome multi locus sequence typing (cgMLST)

Based on the constraints of ordinary multi locus sequence typing (MLST) the software SeqSphere$^+$ allows typing by 1701 genes of *L. monocytogenes* as described by Ruppitsch *et al.* (2015). The *de-novo* assemblies were cgMLST analyzed completely in SeqSphere$^+$ and allele-numbers were automatically assigned using Ridom's allele database. New alleles were submitted to the nomenclature server. For cgMLST and MLST only sorted BAM files were used to control quality of new variants. Assemblies generated by Velvet were stored in FASTA and ACE files. The ACE file stores the reads mapped and were processed by SeqSphere$^+$.

3.12. Single nucleotide polymorphism

For better handling and identification of SNP positions, reference mappings were used. The best reference of all genomes listed in table 4 was calculated for each isolate and the reads were pre-processed as described earlier. The results of the best five references found for each strain were summed up, and the reference ranked highest over the whole number of strains to be compared was taken for read mapping with BWA-mem. The resulting SAM file was converted to BAM format, sorted and piled up using Samtools functions *view* and *mpileup,* to get MPILEUP formatted files and uncompressed VCF files. The MPILEUP file was used for SNP calling using VarScan, the SNPs were filtered to a minimum coverage of 8, occurrence on forward and reverse reads and a minimal variant frequency of 50 %. A python script was written in co-operation with Prof. Rattei, from the University of Vienna, to parse the SNPs and non-SNPs of each chosen isolate to aligned pseudo FASTA sequences. Non-SNP positions were looked up in the MPILEUP file and for non-covered positions gaps were parsed. The sequence alignment was used to calculate a phylogenetic tree using RAxML with the substitution matrix GTRCAT. Visualization was performed using the Interactive Tree of Life on itol.embl.de (Letunic and Bork, 2011).

4. Results

4.1. Re-Cultivation of SLCC *Listeria monocytogenes* isolates

Approximately 10% of 6,451 *Listeria* isolates in the "Special Listeria Culture Collection" (SLCC) were selected for re-cultivation. 591 stabs and 56 lyophilisates were rinsed with fresh media and incubated. 484 out of 647 isolates (74.8 %) in total were successfully re-cultivated, while 90 showed no growth in liquid media. 44 strains had no hemolytic activity or grew in untypical morphology for *L. monocytogenes*. 30 isolates showed no growth, insufficient growth or no specifically colored colonies on RLM selective media agar-plates.

4.2. Serogroup determination by multiplex PCR

For 303 isolates a multiplex PCR was performed to determine the serogroup. Except for the lyophilisates (53 isolates), which were determined in the scheme of Paterson (Paterson, 1940), the serotype was documented in the scheme valid since the 1980's (Seeliger and Hoehne, 1979). Those serotypes were translated to expected serogroups of the PCR results (table 5).

Table 5: Serogroup-serotype correlation table (Doumith *et al.* 2004)

Serogroup	Serotype
IIa	1/2a, 3a
IIb	1/2b, 3b, 7
IIc	1/2c, 3c
IVb	4b, 4d, 4e
Spp	4a, 4ab, 4c and *Listeria* spp.

303 isolates were analyzed in total, 125 were from serogroup IVb, 108 from serogroup IIa, 39 from serogroup IIb, 19 from serogroup IIc and 12 were confirmed as *Listeria* spp., but not assigned to one of those 4 groups (spp.).

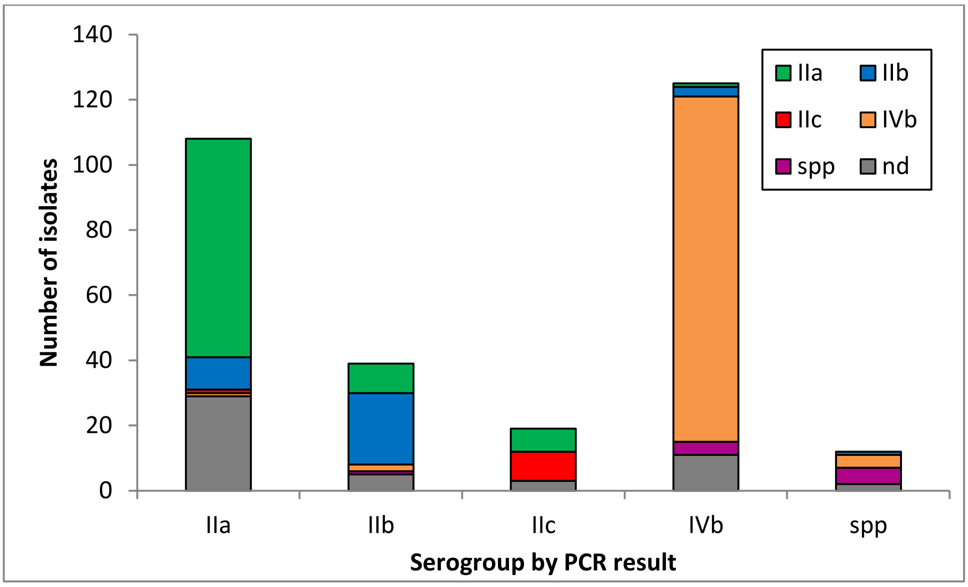

Figure 5: Summary of the 303 PCR results. The stacked bars represent the frequency of each PCR result. The sub-bars indicate the previously expected serogroup, which was derived from the documented serotype.

As shown in figure 5, the majority of PCR results supported the serotype information obtained from the SLCC documentation.

4.3. Quality assessment of raw read data

For sequencing read libraries the base quality is a highly important property. The “base calling” software interpreting the raw measured data, “calls” the base and assigns quality scores for each base, dependent on the assumed probability of an error. On the Illumina Miseq, the image files are automatically and continuously processed while sequencing. A nice overview of the read files including different base quality graphs was generated using FastQC (figure 6).

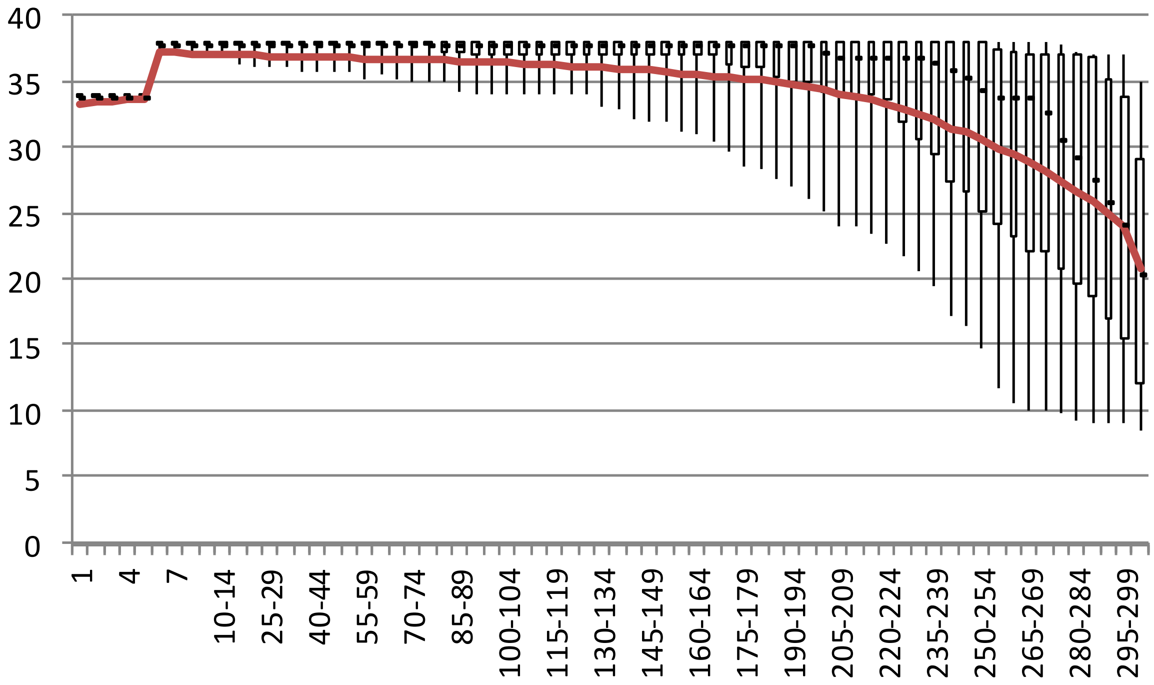

Figure 6: Quality scores across all bases produced by FastQC of sample 930005/11. The unprocessed forward read file of this sample is taken as representative for all sequences. Red: mean PHRED score, black: boxplot with Q1 and Q3 quartiles, median and P10 and P90 percentiles

As seen in figure 6, the sequencing quality is over a PHRED (Ewing *et al.*, 1998) level of 30 for the first 150 bp in almost every sequence, but the average is rapidly decreasing in the second 150 bp. Still the average quality of the 300th sequenced base is around 20, which is an error probability of 1 %. Quality trimming revealed that the average quality of the reverse read was systematically lower than the quality of the forward read. From 9 % up to 40 % of all reverse reads did not pass the quality trimming and filtering, while the forward reads passed the filtering in 99 % of all cases. The values varied in a short range for each sequencing run, dependent on the cluster density. The high amount of low quality reverse reads was independent from the size of the DNA molecules sequenced.

The insert size is another important parameter of shotgun sequencing. It describes the length of the target fragment flanked by sequencing adapters. The average insert size over all samples sequenced was 294 bp, which is less than the read length setting of 300 bp. Meaning, that in slightly more than 50 % of all fragments the same bases were sequenced on the forward read and the reverse read (figure 7).

The sequences downloaded from the ENA had an average insert size of 341 bp, while only 250 bp were sequenced.

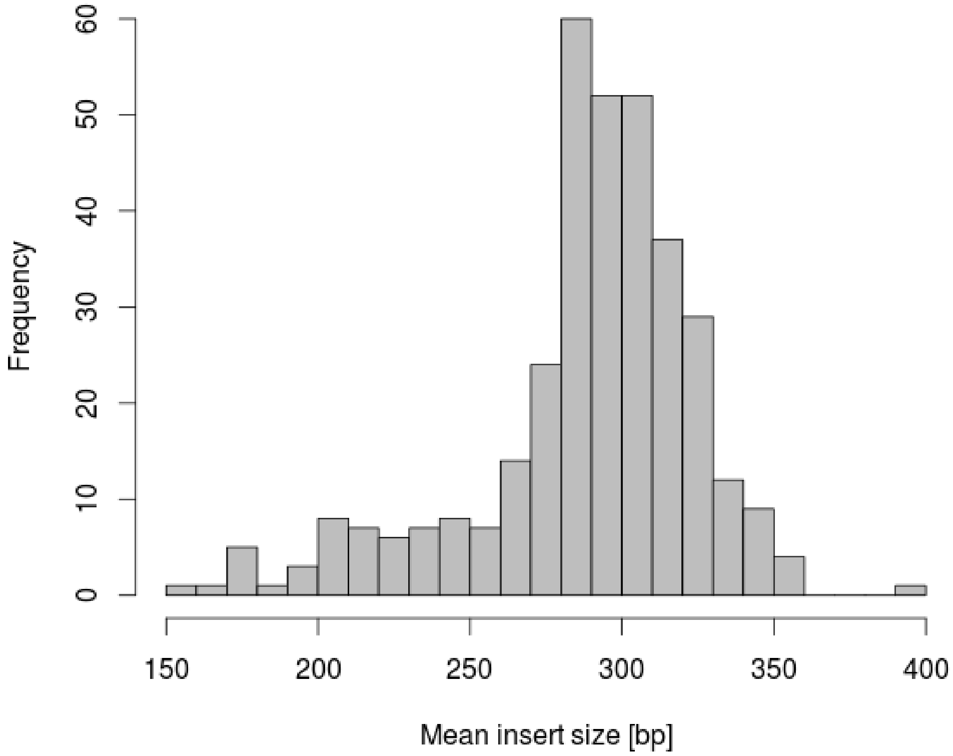

Figure 7: Calculated mean insert size histogram of all isolates sequenced in this thesis. The insert size is lower than the read length. This results in adapter read-through, less trimmed output and a partial loss of the “paired-end” advantage.

A statistical analysis of the sequence lengths for all reads of isolate 930005/11 shows the distribution of lengths (figure 8). The amount of sequences shorter than 300 bp is slightly higher than the amount of sequences with 300 or 301 bp.

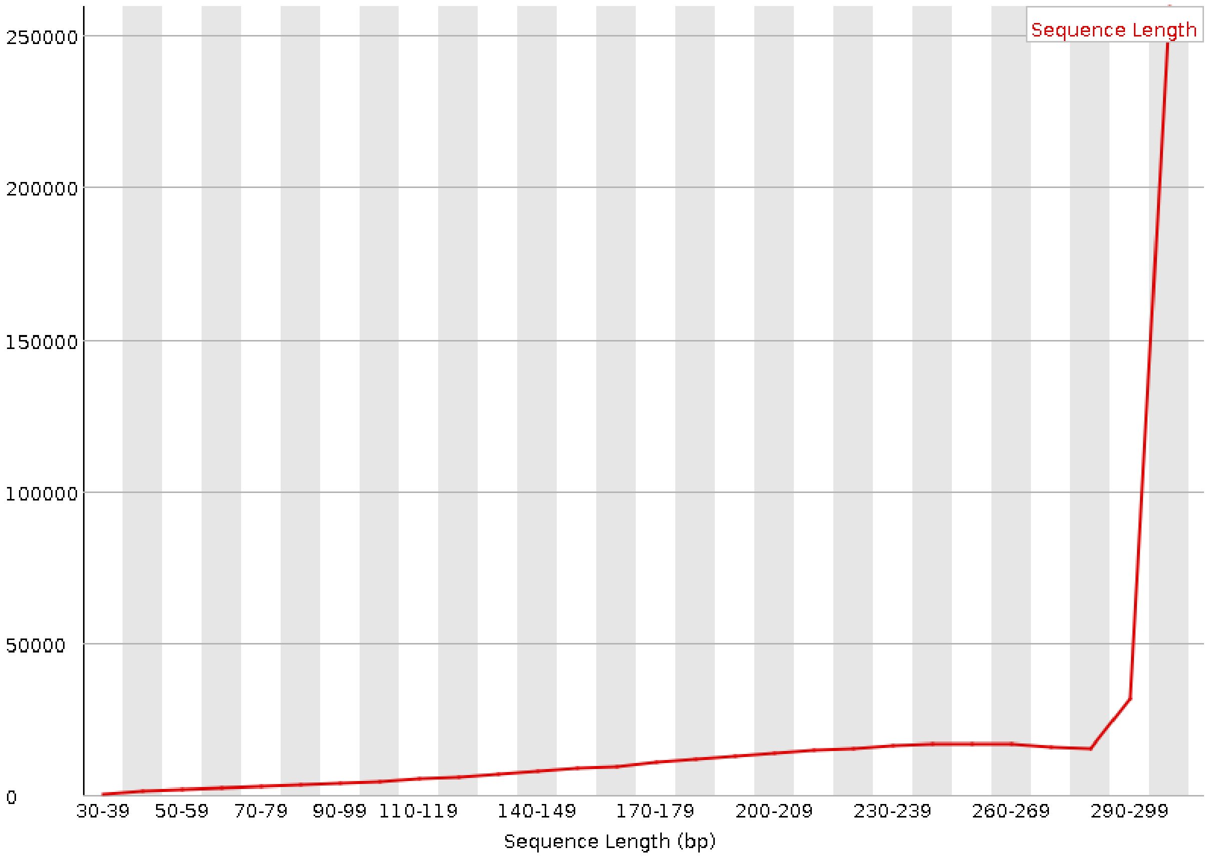

Figure 8: Distribution of the read lengths for isolate 930005/11. The graph was produced using FastQC for analysis of untrimmed forward sequences as generated by Illumina MiSeq.

Although 284 thousand read sequences are shorter than the read length used on the MiSeq, almost no sequences were found by FastQC originating from adapter or index sequences used in the library preparation or on the sequencer (figure 9).

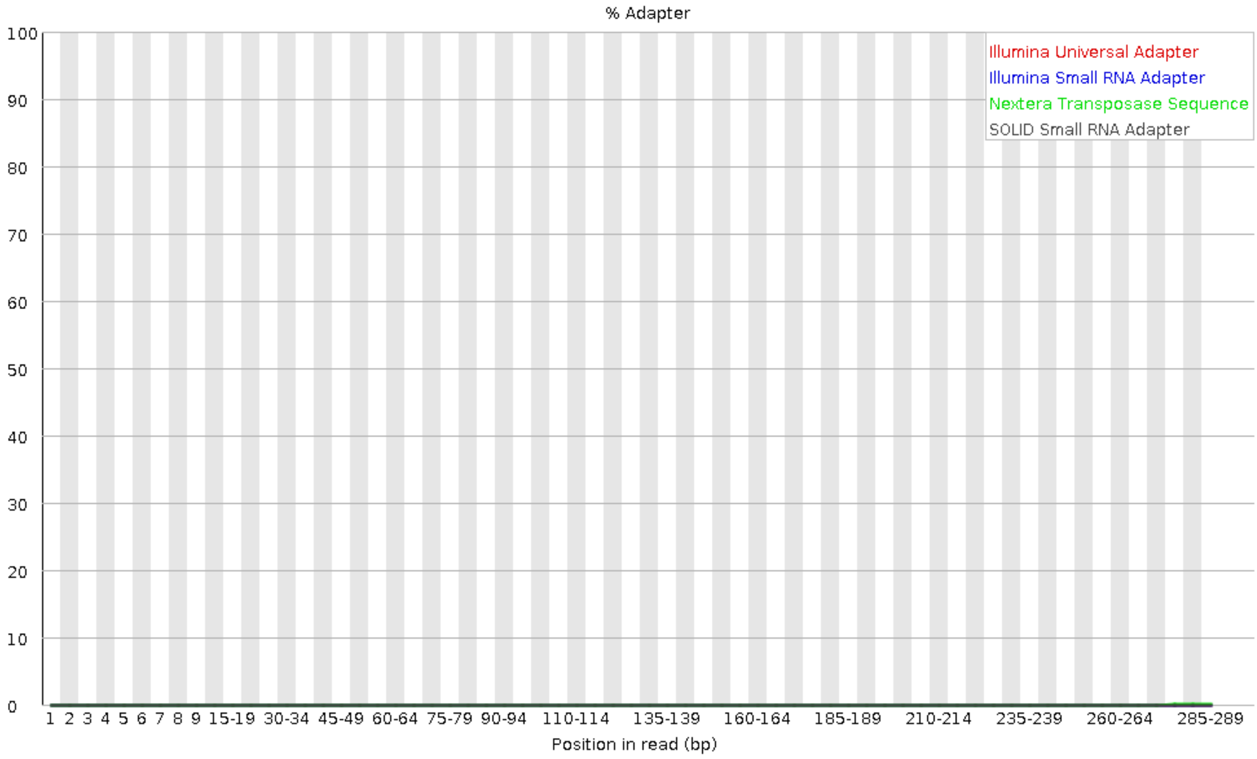

Figure 9: Content of sequences found for isolate 930005/11, which originated from adapter- or index-sequences. The graph was produced using FastQC for analysis of untrimmed forward sequences as generated by Illumina MiSeq.

4.4. Average contamination level estimation

The contamination levels were mainly quantified by collecting reads, which could not be mapped to *de-novo* assemblies. Also the contigs of all assemblies were searched for foreign DNA content using BLAST. Except for four isolates, the average contamination amount was below 1 % with an average of 2,382 or 0.19 % possibly contaminated reads. The distribution of uncontaminated reads excluding isolates SLCC180, L16-12, L22-12 and 930046/11 as outliers is shown in figure 10.

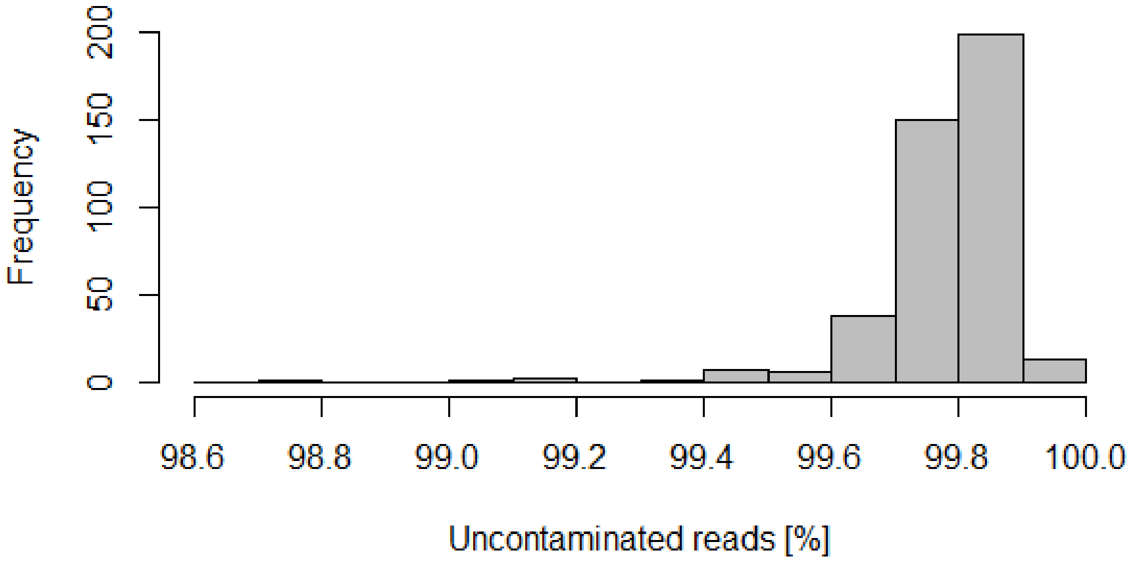

Figure 10: A very high amount of sequences was uncontaminated in the majority of datasets of whole-genomes sequences. Four out of 501 isolates are outliers which are not shown here.

In 26 isolates one contig was possibly contaminated, in five isolates two contigs and in four isolates three or more contigs were possibly contaminated. The contaminations mostly consisted of plasmids of uncultured species, *Paenibacillus larvae* or belonging to *Staphylococcus* genus. Except for two isolates, the contaminations were restricted to particular small contigs unrelated to *Listeria* genetic information. In isolates MRL-14/00747 and R-61951 about 15 kb long regions, embedded in larger contigs, returned a 98 % identity to *Enterococcus sp.* draft genome with the accession number FP929058.1. In fact the affected contigs were completely aligned to *Listeria monocytogenes* genomes SLCC2755 and N1-011A with a 98 % identity.

The sequencing data of isolate SLCC180 were contaminated by a different *Listeria monocytogenes* strain to a high extend. The actual contamination level is impossible to determine, the contamination itself was detected during the assembly process. Since the Velvet assembly was defective, SPAdes was utilized, returning larger contigs. After read mapping and analysis using core genome MLST, several hundred genes contained low frequency nucleotide variants of about 35-40 % of reads representing a different allele.

The highest amount and ration of possibly contaminated reads was found in the L22-12 dataset. 405,188 reads were collected as unmapped of which 99,043 reads had high identity to plasmids occurring in different *Staphylococcus* species. The majority of reads returned no BLAST hit (271,613 reads). Further 51,553 reads were mapped to a single 401 bp large contig which sequence had 100 % identity to a plasmid of the *Staphylococcus* genus.

Large quantities of potential contaminations were found in the isolates 930046/14 (181,645 reads) and L16-12 (173,431 reads). Isolate L16-12 is the isolate with the highest number of reads mapped to possibly contaminated contigs (65,323 reads on four different contigs).

Table 6: Overview of potentially contaminated contigs of isolate L16-12 with counted reads and top hit. 4 different contigs have high coverage regions without similarity to *Listeria spp.* genomes.

Contig size	**Region**	**Number of Reads**	**E-value**	**Organism (% identity)**
277	23-55	8890	0.000409	*Bacillus sp.* (96 %)
416	1-122 298-416	9363 9421	2.2e-041 2.7e-040	*Lysinibacillus sp.* 13S34 (91 %) *Bhargavaea sp.* DMV9 plasmid pBSDMV9 (93 %)
277	1-277	21243	1.2e-105	P*aenibacillus larvae* plasmid pMA67 (91 %)
310	1-159 152-310	6582 9824	3.8e-062 3.8e-062	*Bhargavaea sp.* DMV9 plasmid pBSDMV9 (94 %) *Lysinibacillus sp.* 13S34 (94 %)

Additionally to the contaminations listed in table 6, in L16-12 108,108 unmapped reads were collected. Only 17,758 reads returned clear BLAST hits, while 16,273 of them were accounted to *Paenibacillus larvae* plasmid pPL374.

The contigs of isolate 930046/14 showed only one possibly contaminated region with 846 reads mapped to 4448 bp long sequence. The sequence returned similarities to an *Enterococcus faecium* plasmid to 97 % identity but also to *Listeria* plasmid pLMR479a to 98 % identity. Hence 180,799 unmapped reads were collected, of which

46,595 returned BLAST hits. A majority of 30,794 reads were accounted to *Paenibacillus larvae* plasmid pPL374.

As shown for the three highly contaminated isolates, the proportion of unmapped reads without BLAST hits is higher than reads which return hits. For the rest of the isolates, which had a low contamination level, only 11.6 % of all unmapped reads were successfully identified by BLAST.

4.5. Comparison of different *de-novo* assembling tools using cgMLST

90 isolates were analyzed by different methods for assembling or mapping with subsequent core genome multi locus sequence typing (cgMLST): BWA-mapping to the best reference and the *de-novo* assemblers MaSuRCA, SPAdes and Velvet. For each isolate one Illumina MiSeq sequenced paired-end read library was created. The reads were pre-processed the same way as described, and for the BWA mapping the best reference was calculated as described. The same 1,701 genes were analyzed in all four draft assemblies for 90 isolates (table 7).

Table 7: Overview of all 90 samples assembled using four different methods. The reference used and the percentage of mapped reads refer to the reference based assembly with BWA. Values given represent failed and missing genes (^{1}failed/^{2}missing)

Sample ID	Reference used	% mapped to reference	BWA	SPAdes	MaSuRCA	Velvet
12025647	L2074	98,93	$3^{1}/4^{2}$	2/2	2/6	8/2
3230TP3	L2074	99,02	4/4	2/2	2/7	5/2
4548TP4	L2074	98,90	3/4	2/2	2/7	8/2
CIP104794	EGD	99,79	11/0	11/2	11/1	18/1
CIP105448	SLCC2372	99,58	4/0	4/0	4/6	8/0
CIP105449	SLCC2755	99,70	9/0	9/0	9/6	19/0
CIP105457	HCC23	99,60	15/0	15/0	15/0	19/0
CIP105458	ATCC19117	99,64	8/0	8/0	8/2	9/0
CIP105459	SLCC2378	99,61	13/1	13/1	13/3	21/1
CIP5953	J1-220	99,69	8/0	8/0	8/6	8/1
CIP7834	EGD	99,44	10/0	10/0	10/5	10/0
CIP7835	SLCC2540	99,59	7/1	7/1	7/2	7/2
CIP7836	SLCC2479	98,58	1/0	1/0	1/2	1/2
CIP7843	SLCC2482	99,43	8/0	8/0	8/1	8/0
K70-10	L2074	98,86	4/4	2/2	2/4	2/2
L10/10	L2074	98,60	4/4	2/2	2/4	2/2
L01/13	J1-220	94,02	15/2	9/2	9/4	9/2
L12/13	NE-dc2014	99,63	9/0	9/0	9/3	9/0
L13/13	SLCC2479	96,69	2/1	1/13	1/2	1/0
L14/10	L2074	98,73	4/4	2/2	2/5	2/2

L14/13	L2625	96,94	5/2	2/0	2/6	2/2
L15/12	SLCC2482	97,01	9/0	8/0	9/1	8/0
L15/13	SLCC2540	97,37	12/1	6/1	6/3	6/1
L16/10	L2074	98,77	3/4	2/2	2/2	2/2
L16/12	CFSAN006122	79,55	6/0	6/0	6/6	6/3
L16/13	CFSAN023463	98,14	10/0	6/0	6/0	6/0
L17/10	L2074	98,86	3/4	2/11	2/7	2/6
L17/12	NE-dc2014	99,69	9/0	9/0	9/0	9/0
L18/10	L2074	99,04	3/4	2/2	2/9	2/3
L18/13	L2625	97,96	4/2	3/2	3/2	3/2
L19/10	L2074	98,86	4/4	2/2	2/2	2/3
L20/09	L2074	96,34	1/4	0/2	0/8	0/2
L20/10	L2074	98,99	4/4	3/2	3/9	3/3
L20/13	J0161	91,88	6/2	6/2	6/3	6/1
L21/09	L2074	96,06	3/4	1/2	1/7	1/9
L02/13	SLCC7179	97,33	1/1	0/1	0/4	0/1
L22/12	L2625	65,59	1/0	0/0	0/3	0/1
L22/13	6179	98,15	2/0	2/0	2/5	2/0
L23/09	L2074	94,38	1/4	0/2	0/19	0/2
L23/12	Clip80459	96,29	10/0	9/0	10/3	9/0
L23/13	C1-387	97,46	6/1	5/1	5/4	5/2
L24/12	L2676	98,47	3/0	3/0	3/0	3/0
L24/13	N2306	99,62	10/1	10/4	10/3	11/1
L25/12	R479a	97,24	1/0	0/0	0/1	0/0
L25/13	N2306	99,68	8/0	8/0	8/0	8/0
L26/12	CFSAN006122	99,66	6/0	6/0	6/2	6/0
L26/13	J2-031	95,02	7/1	3/1	5/49	7/1
L27/12	NE-dc2014	98,03	10/0	9/6	10/2	9/0
L27/13	L2625	95,60	6/1	3/1	3/2	3/0
L28/12	L2625	96,76	3/2	1/3	1/1	1/0
L28/13	L2676	98,61	3/0	3/0	3/3	3/0
L29/12	J1-220	97,62	15/2	10/3	10/5	10/2
L29/13	L2625	92,19	4/3	3/1	3/6	3/1
L30/12	J1-220	99,47	7/0	7/4	7/2	7/0
L30/13	SLCC2540	97,57	9/3	7/3	7/4	7/3
L31/12	J1-220	98,26	10/0	9/0	9/1	9/0
L31/13	J2-1091	97,51	11/0	11/0	11/2	11/0
L32/12	J1-220	99,07	8/0	8/1	8/21	8/0
L33/12	Clip80459	98,17	9/0	8/0	8/1	8/0
L34/12	EGD	98,50	3/0	3/0	3/0	3/0
L35/12	SLCC2479	98,01	2/0	2/1	3/0	2/0
L35/13	L2625	97,37	5/1	3/0	3/1	3/1
L36/12	NE-dc2014	95,97	10/0	10/2	10/0	10/0
L37/12	NTSN	99,61	9/1	10/3	11/39	11/3
L38/12	WSLC1042	99,53	10/1	10/0	10/1	10/0
L40/12	J1-220	98,12	7/0	7/0	7/2	7/1
L41/12	L2074	94,84	3/4	1/2	1/3	0/2
L06/13	C1-387	97,32	6/1	5/1	5/4	5/1
L07/13	L2625	96,72	4/2	2/0	2/3	2/0
L08/13	J0161	92,36	6/3	5/3	5/2	5/2
L09/10	L2074	96,27	2/4	0/2	0/3	0/2
L09/13	C1-387	96,48	5/0	5/0	5/2	5/0
LD12-10	L2074	98,88	4/4	2/2	2/7	2/2
LD27-12	L2074	96,87	2/5	0/2	0/6	0/8
MRL-13/00230	L2074	97,36	4/4	0/8	0/4	0/4

MRL-13/00815	NTSN	97,13	10/2	9/2	9/1	9/1
MRL-13/00816	LL195	99,73	10/0	10/1	10/3	10/0
MRL-15/00014	N2306	99,21	9/0	9/0	9/0	9/0
MRL-15/00015	J2-1091	98,74	10/0	10/0	10/0	10/0
MRL-15/00016	SLCC2540	96,89	12/1	6/1	6/2	6/1
MRL-15/00032	FSL R2-561	95,56	2/1	0/0	0/3	0/0
MRL-15/00033	J0161	89,26	8/3	5/2	5/5	5/2
MRL-15/00034	SLCC2540	95,36	12/0	7/0	7/2	7/0
MRL-15/00035	J2-031	95,87	3/1	1/1	1/2	1/1
MRL-15/00063	J0161	90,26	9/3	6/2	5/2	5/2
MRL-15/00085	SLCC7179	88,91	14/36	5/31	5/39	5/31
MRL-15/00093	LL195	97,58	11/0	10/0	10/0	10/0
MRL-15/00127	J2-031	94,80	2/1	0/8	0/2	0/1
MRL-15/00128	J0161	90,20	7/3	6/2	7/3	5/2
MRL-15/00145	L2074	95,79	2/4	0/2	0/4	0/2

In 80 cases, all four methods led to the same cgMLST cluster type, while in one case two genes and in nine cases one gene differed in their allele numbers between the four methods. BWA mapping showed a different allele in eight genes than the *de-novo* assemblers, MaSuRCA led to a different result in two cases. There was one case were BWA mapping and MaSuRCA shared the same allele number, while SPAdes and Velvet shared one number. For each gene that was called wrongly in one or more methods, the SeqSphere+ software stated a warning because of a low coverage region (coverage below 5).

Apart from wrong alleles, BWA also had the highest average number of genes with frameshifts (6.4) compared to MaSuRCA (5.2), SPAdes (5.1) and Velvet (5.7).

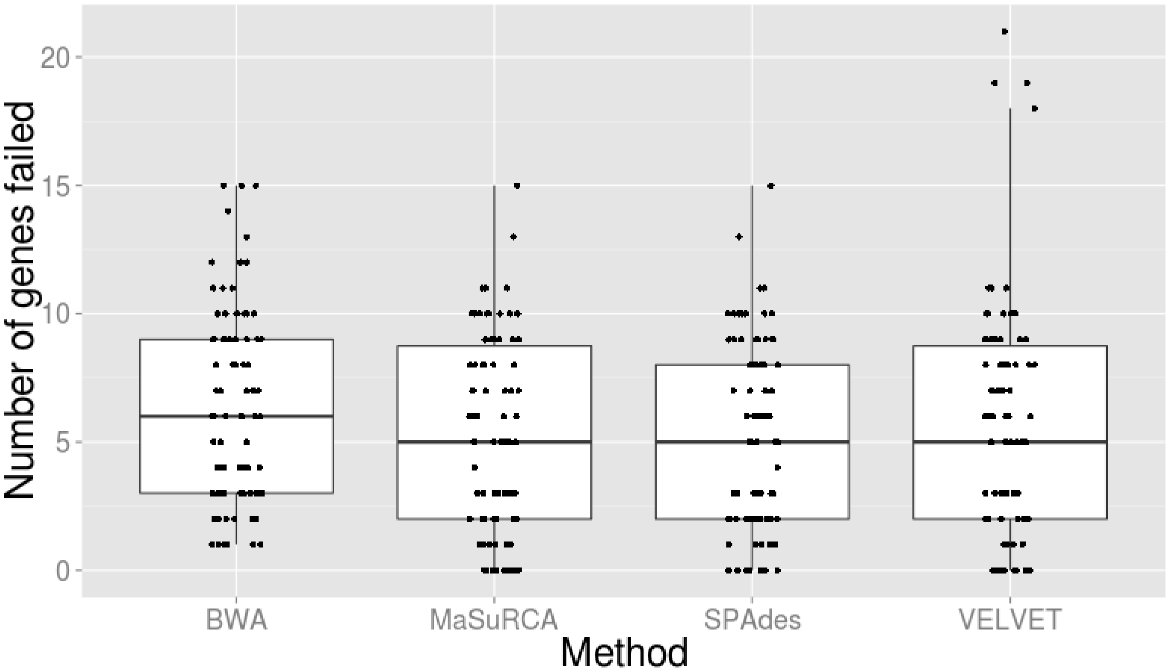

Figure 11: Boxplot of the actual numbers of genes with frameshifts or a different consensus length than the reference target gene. 90 samples were assembled using four different methods. The total number of genes with a different consensus length than its reference is plotted for each sample, represented by a dot. Genes were identified using SeqSphere$^+$. For each method a boxplot (10 %, 25 %, 50 %, 75 % and 90 % percentiles) was overlapped for better visualization.

A surprisingly high number of missing genes was encountered with the *de-novo* assembler MaSuRCA, where the isolates had 4.8 missing genes on average, compared to 1.8 (BWA), 1.8 (SPAdes) and 1.6 (Velvet) for the other methods.

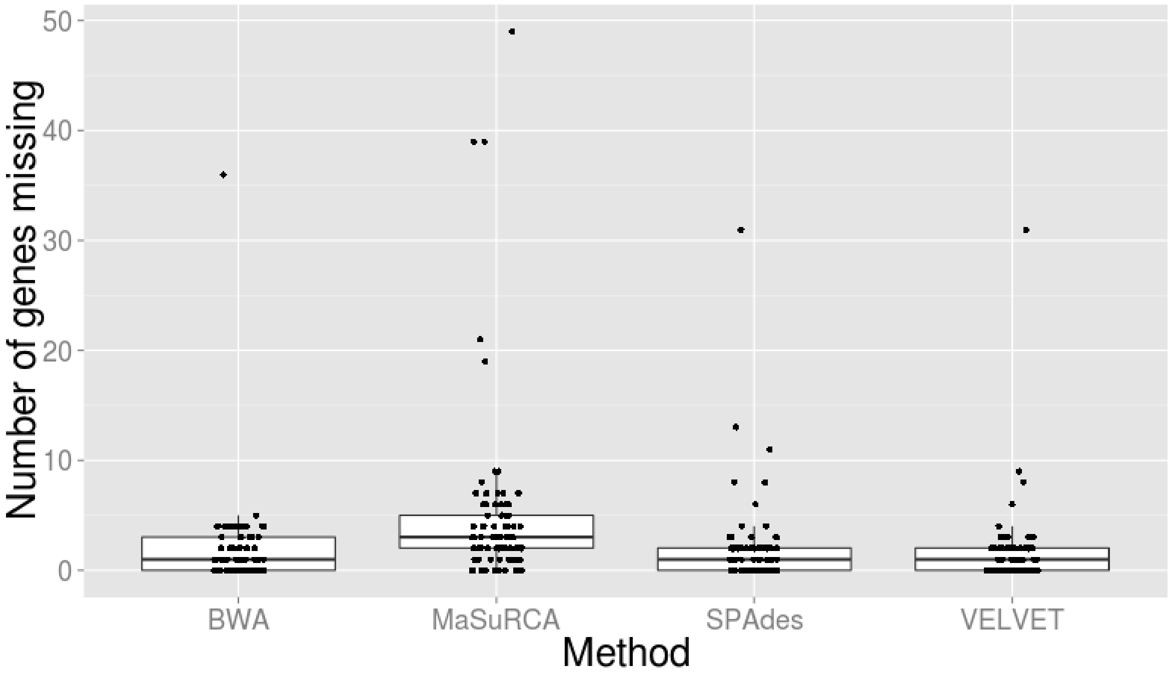

Figure 12 Boxplot of the actual numbers of genes not found in the draft assemblies by SeqSphere⁺. 90 samples were assembled using four different methods. The total number of genes not found is plotted for each sample, represented by a dot. Genes were identified using SeqSphere⁺. For each method a boxplot (10 %, 25 %, 50 %, 75 % and 90 % percentiles) was overlapped for better visualization.

All four tested methods were good enough to pass the quality standards for cgMLST in SeqSphere⁺ with more than 95 % of core genome genes passing the filter, most of the isolates even had more than 99 %.

Overall, SPAdes was the assembly engine with the lowest average of missing or failed genes (6.9) and Velvet comes very close with 7.3 genes. Assemblies created using reference mapping with BWA in sum resulted in 8.4 genes missing or failed. MaSuRCA had the highest average value (10 genes), which was caused by the highest number of missing genes over all methods.

The total amount of failed or missing genes analyzed with cgMLST for each method is shown in figure 13.

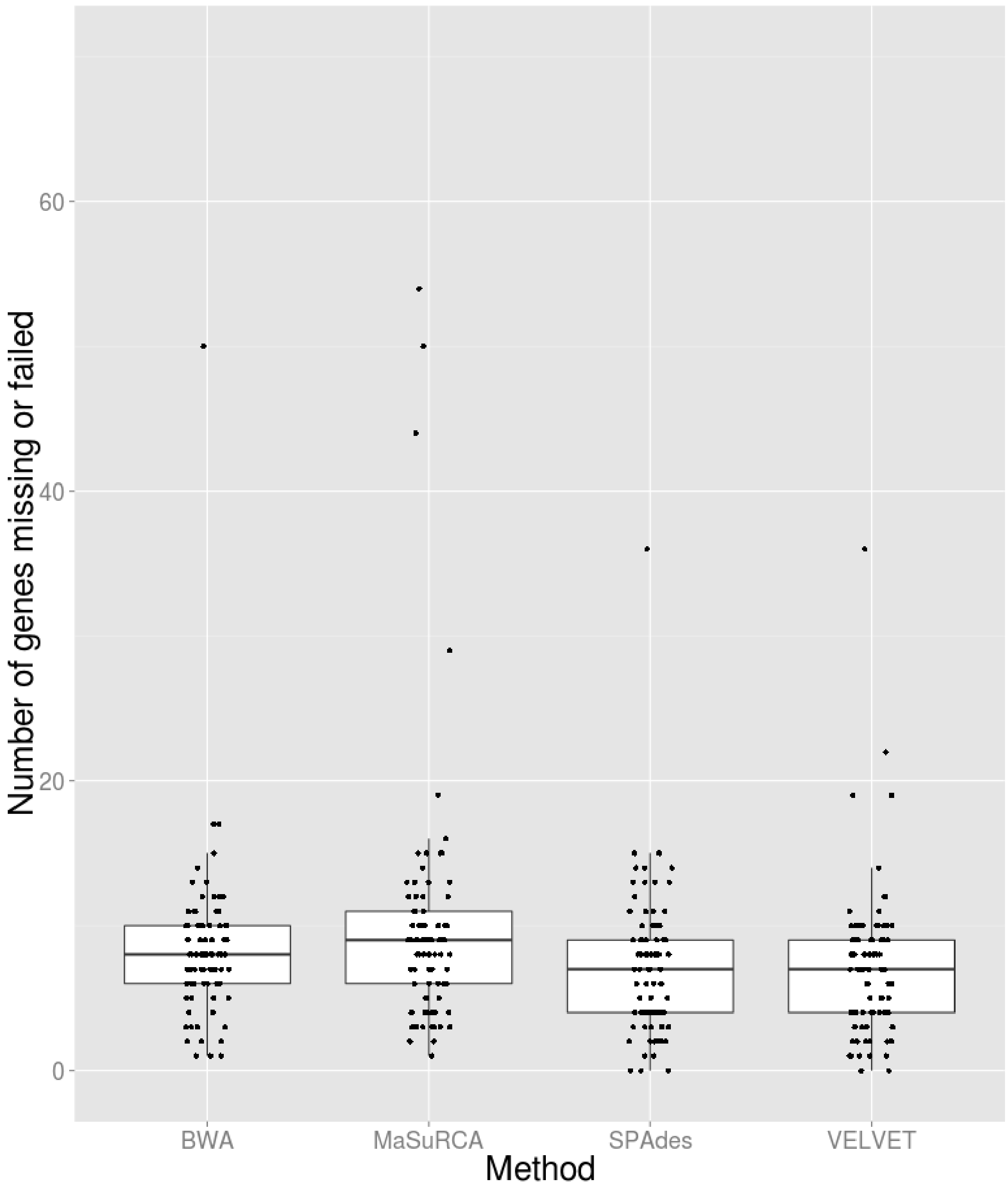

Figure 13: Boxplot of the total numbers of genes not found in the draft assemblies or genes with differing consensus length than the reference. 90 samples were assembled using four different methods (BWA-reference mapping, MaSuRCA, SPAdes and Velvet). The total number of genes not found or with wrong consensus length is plotted for each sample and each method, represented by a dot. Genes were identified using SeqSphere^{+}. For each method a boxplot (10 %, 25 %, 50 %, 75 % and 90 % percentiles) was overlapped for better visualization.

4.6. Core genome multi locus sequence typing

The simplification of the complete genome to 1,701 single genes made it possible to compare all of the sequenced *L. monocytogenes* draft genomes. Based on allele numbers, distances were calculated and the distances were drawn to a phylogenetic tree. As seen in figure 14, there are up to four different lineages. The serogroup IIb forms a distinct sub-lineage, opposed to what was reported to be a single lineage with IVb. Although they appear closer related than different isolates of IIa, IIb and most IVb isolates seem to have evolved from a common ancestor and form two clades.

Classified in the same lineage, the IIc group represents only a small subgroup of IIa. Further, some strains of rare serotypes like 4c, 4a and 4e are not related to one of the apparent lineages and highly divergent to each other.

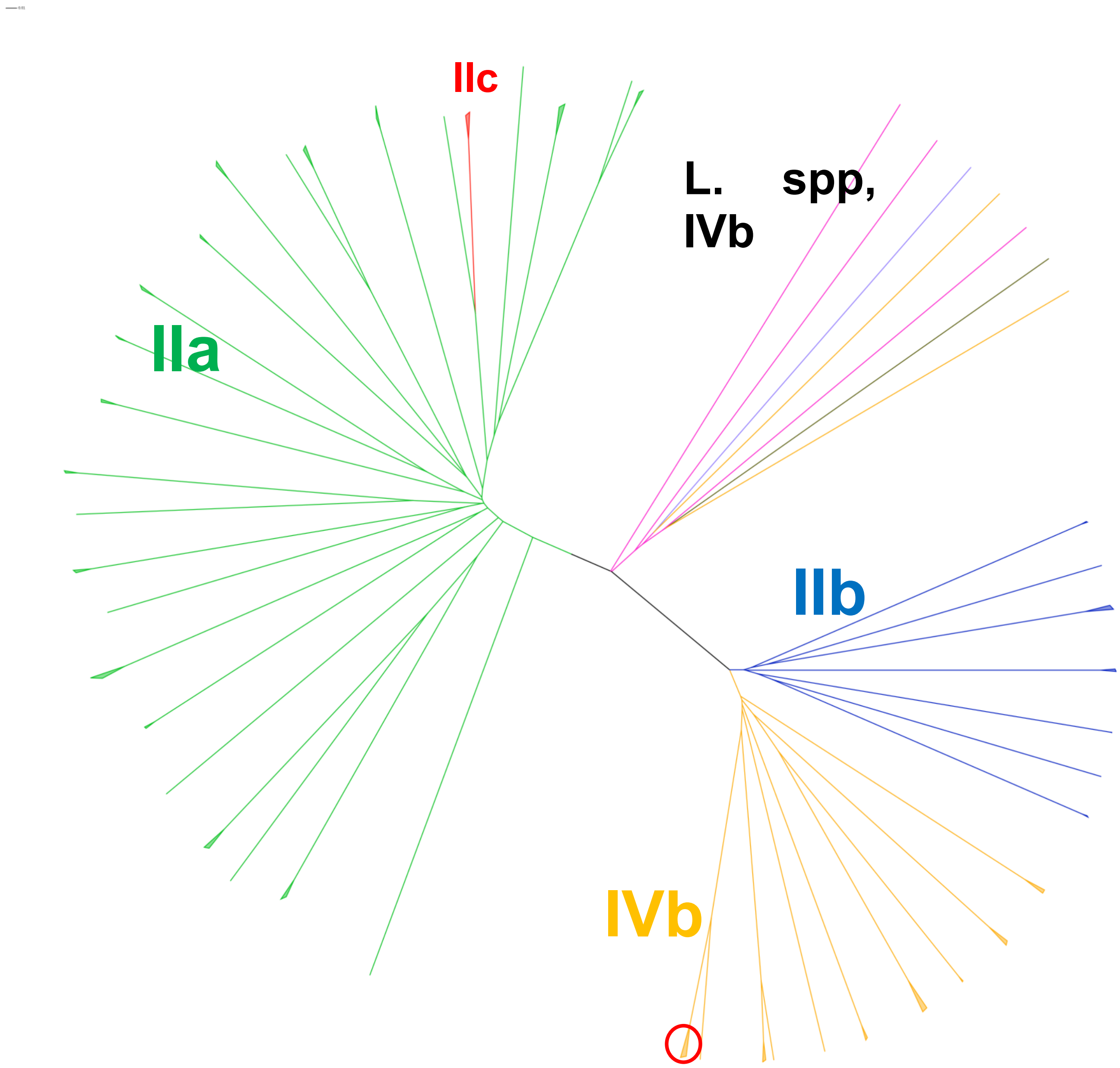

Figure 14: Phylogenetic tree of *L. monocytogenes* genomes based on cgMLST distances. Green branches represent serogroup IIa, blue branches serogroup IIb, red branches serogroup IIc and yellow branches serogroup IVb. Pink, grey and black represent 4a, 4c and 4ab. The red cycle marks the clade further discussed.

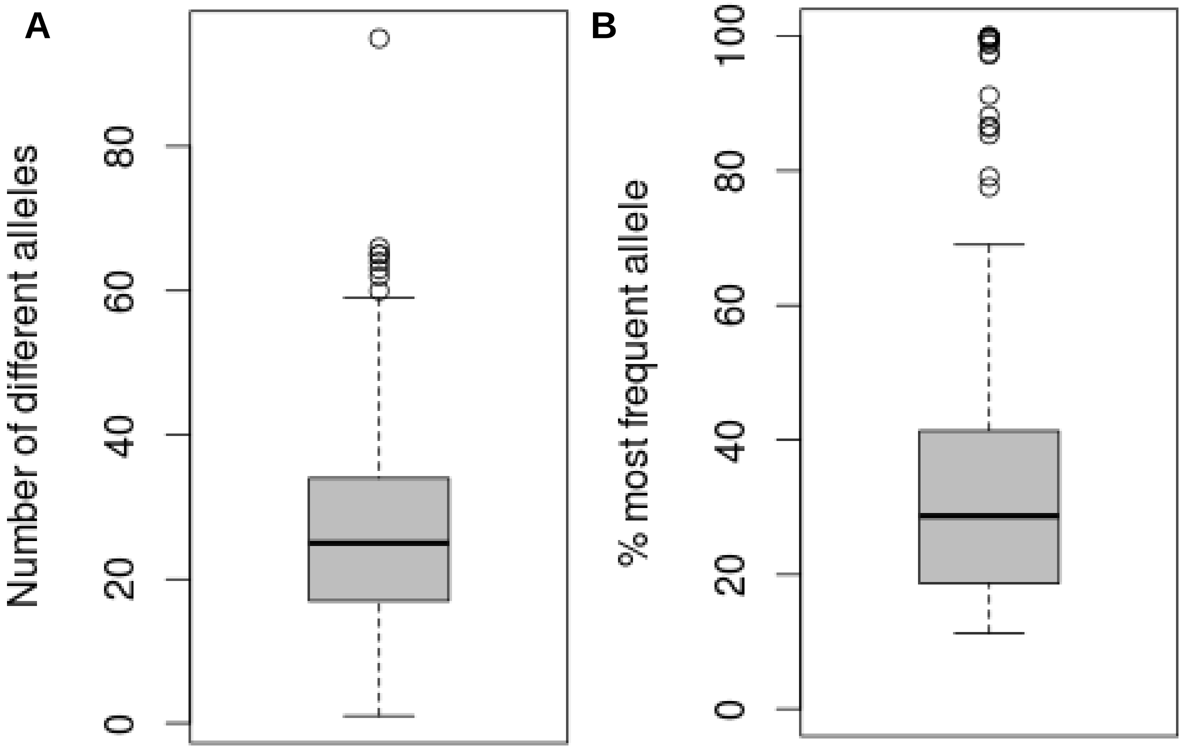

Figure 15: The boxplots show the variance in the core genome for L. monocytogenes. A: 1701 genes of the core genome were analyzed by their allele-number. The average gene had between 17 and 34 different alleles in the dataset (n=432). **B:** The percentage of the most frequent allele shows that only few genes have a predominant allele.

The core genome MLST also allows taking a look on potentially highly variant genes and highly conserved genes. As shown in figure 16A, the median of all 1701 genes had a relatively variance of 25 different alleles found in 432 isolates analyzed, while the minimum is 1, the lower quartile 17, the upper quartile 34 and the maximum is 95. The relative frequency of the most frequent allele is shown in figure 16B. The median is 28.7 %, the quartiles are 18.6 % and 41.3 % with the maximum at 100 % and the minimum at 11.3 %.

As seen in table 8 the number of possible variants does not always correlate with the ratio of the most frequent allele.

Table 8: Overview of selected gene loci sorted by amount of different alleles found. The top of the table (less than 4 different alleles) and the bottom (10 genes with the highest variance) are shown.

Gene locus	Gene product	Product description retrieved from NCBI GenBank	Total number of different alleles	Most frequent allele [%]
lmo2609	rpmJ	50S ribosomal protein L36	1	100,0
lmo2614	rpmD	50S ribosomal protein L30	1	100,0
lmo1335	rpmG	50S ribosomal protein L33	2	99,7
lmo1364	cspL	cold-shock protein	2	99,7
lmo1816	rpmB	50S ribosomal protein L28	2	99,7
lmo2616	rplR	50S ribosomal protein L18	2	99,7
lmo1797	rpsP	30S ribosomal protein S16	3	51,4
lmo2534	atpE	ATP synthase F0F1 subunit C	3	50,9
lmo2548	rpmE2	50S ribosomal protein L31	3	85,4
lmo2619	rpsN	30S ribosomal protein S14	3	52,6
lmo2623	rpsQ	30S ribosomal protein S17	3	86,4
lmo2689a	lmo2689a	Hypothetical protein	3	77,6
lmo2856	rpmH	50S ribosomal protein L34	3	99,5
…			…	…
lmo0288	lmo0288	Two-component sensor histidine kinase	59	14,9
lmo2488	uvrA	Excinuclease ABC subunit A	59	16,6
lmo0096	lmo0096	PTS mannose transporter subunit IIAB	60	21,9
lmo0788	lmo0788	Unknown function	62	16,4
lmo1320	polC	DNA polymerase III PolC	63	16,1
lmo1226	lmo1226	Uncharacterized membrane protein	64	15,4
lmo1576	lmo1576	Hypothetical protein	65	11,3
lmo0259	rpoC	DNA-directed RNA polymerase subunit beta'	65	15,6
lmo0892	rsbU	Serine phosphatase	66	22,9
lmo0210	Ldh	L-lactate dehydrogenase	95	18,1

In the upper half of table 8, there are three gene loci listed with three different alleles found but only around 50 % of isolates had the most frequent allele indicating an equal distribution over the alleles. The distribution of alleles of gene locus *lmo2619* is shown in figure 16.

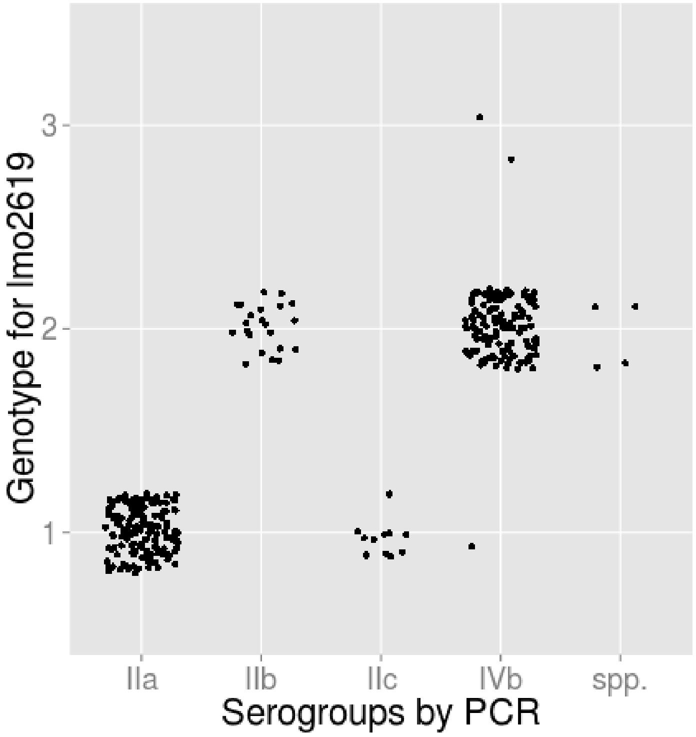

Figure 16: Distribution of genotypes for gene locus *lmo2619* over serogroups of 501 genomes. Gene locus lmo2619 has only three different alleles in 432 isolates while type 1 is mainly found in IIa and IIc isolates and type 2 in IIb, IVb and undeterminable isolates.

By gene locus *lmo2619* the serogroup can be anticipated. Serogroups IIa and IIc have genotype 1 while IIb, IVb and others have mainly genotype 2. Analysis of gene loci *lmo1797* and *lmo2534* returns similar results as locus *lmo2619*.

4.7. Evolutionary distance between SLCC isolates and recent clinical isolates

The clade indicated by the red cycle at the bottom of figure 14 seems to be very closely related in this context, but its isolates are too widespread to be appendant to an outbreak. This cluster is also shaded light in figure 17, where the isolates are displayed in a minimum spanning tree (MST). Each strain is connected by a line to the closest sequence, labeled with the total number of genes with different alleles. The red circles are clinical isolates, while the blue ones are historic SLCC isolates from different years. As seen in figure 17, isolates from the SLCC from 1921 and from years 1954 to 1980 are closely related to isolates from 2012 to 2015. Obviously there is no clonal context possible between samples from the SLCC and clinical isolates. There might be a common ancestor, which persisted over 94 years or evolved slowly. There are more examples proving that e.g. SLCC632 is just 37 alleles different from strains of the acid curd cheese outbreak occurring in 2008 and 2009.

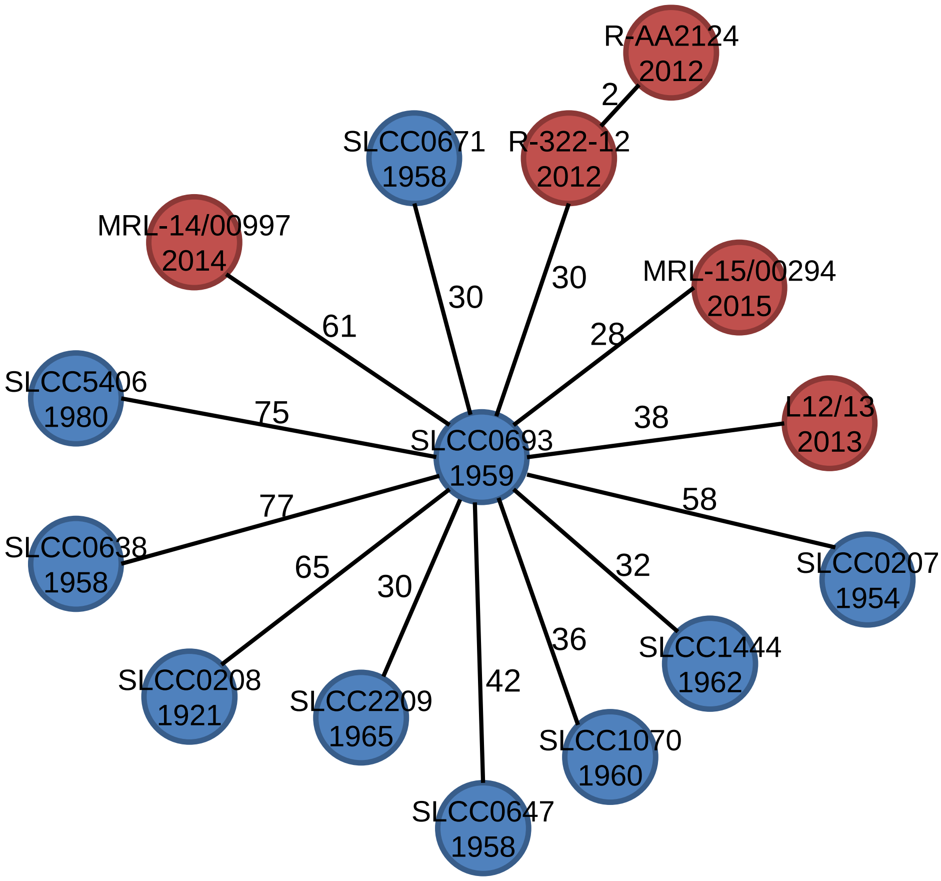

Figure 17: The minimum spanning tree shows how close historical SLCC strains (blue) and recent clinical isolates (red) can be related. The year of isolation and the strain ID are labeled. The distances are calculated by the total number of genes different to the closest isolate.

In contrast to other methods like pulse field gel electrophoresis (PFGE), the distance between remote isolates can be estimated and depicted in a number for core genome gene deviation. For *L. monocytogenes* single-nucleotide polymorphism (SNP) based methods are impracticable for comparison of remote isolates. The groups of IVb strains and IIa strains have completely rearranged genomes.

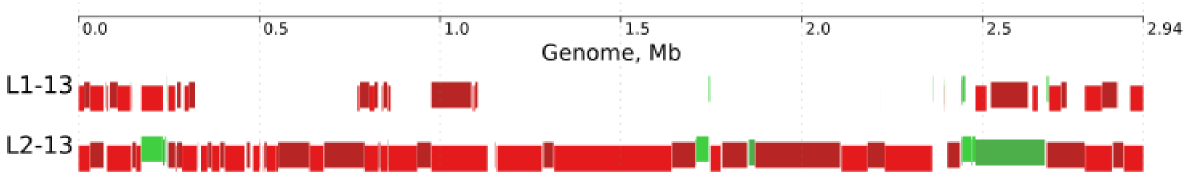

Figure 18: Genome alignment of raw assemblies to the reference genome of the *L. monocytogenes* strain EGD-e (serogroup IIa). Strain L1-13 is of serogroup IVb, strain L2-13 of serogroup IIa.

As shown in figure 18 it is not possible to align both *de-novo* assembled genomes to the EGD-e reference genome. For the genome of strain L1-13 (serogroup IVb), only 31.8 % were aligned with 4,053 SNPs per 100 kbp, while 94 % of strain L2-13 (serogroup IIa) was successfully aligned and 995 SNPs per 100 kbp were computed. Further read mapping to the EGD-e reference led to 88 % mapped reads of the IVb strain with 131,982 SNPs and 863 short indels, while 98 % of the IIa genome reads could be mapped with 26,784 SNPs and 247 short indels.

4.8. Cluster investigation by cgMLST and single nucleotide polymorphism (SNP) based analysis in comparison

As shown, cgMLST analysis can be used to type strains species specific, independent from serotypes or lineages. The simplification of genetic variation assessment to a gene-by-gene comparison saves time and computation resources. In consequence one SNP in a gene has the same phylogenetic impact as 10 SNPs in the same gene. Furthermore, intergenic regions, where the evolutionary clock is faster, are completely ignored. It could be assumed that SNP based methods would lead to more reliable results in comparison of closely related strains, while cgMLST is capable of giving a quick overview. Figure 19A shows a phylogenetic tree based on cgMLST of a small cluster with an allelic distance between 28 and 77 genes. Of the same cluster 60 strains were analyzed for SNPs in the whole genome by using a reference based method. On average, 92 % of reads were successfully mapped to the reference genome of *L. monocytogenes* strain J1-220. A total of 2,088 SNP positions were identical. On 49 positions all 60 isolates tested had the same variant, different from the reference. Excluding positions where at least one isolate had no sufficient coverage, 2,037 positions in total and 42 positions with one consistent variant remained. For better visibility, a subset of the cluster of 21 isolates is displayed in figure 19AB. Those selected 21 isolates have 1,046 consistent SNPs compared to the reference genome.

Hence 1042 SNPs, about half of the SNPs for the complete cluster, remain for phylogenetic calculations. For core genome MLST 437 of 1701 genes had more than one variant in 21 strains.

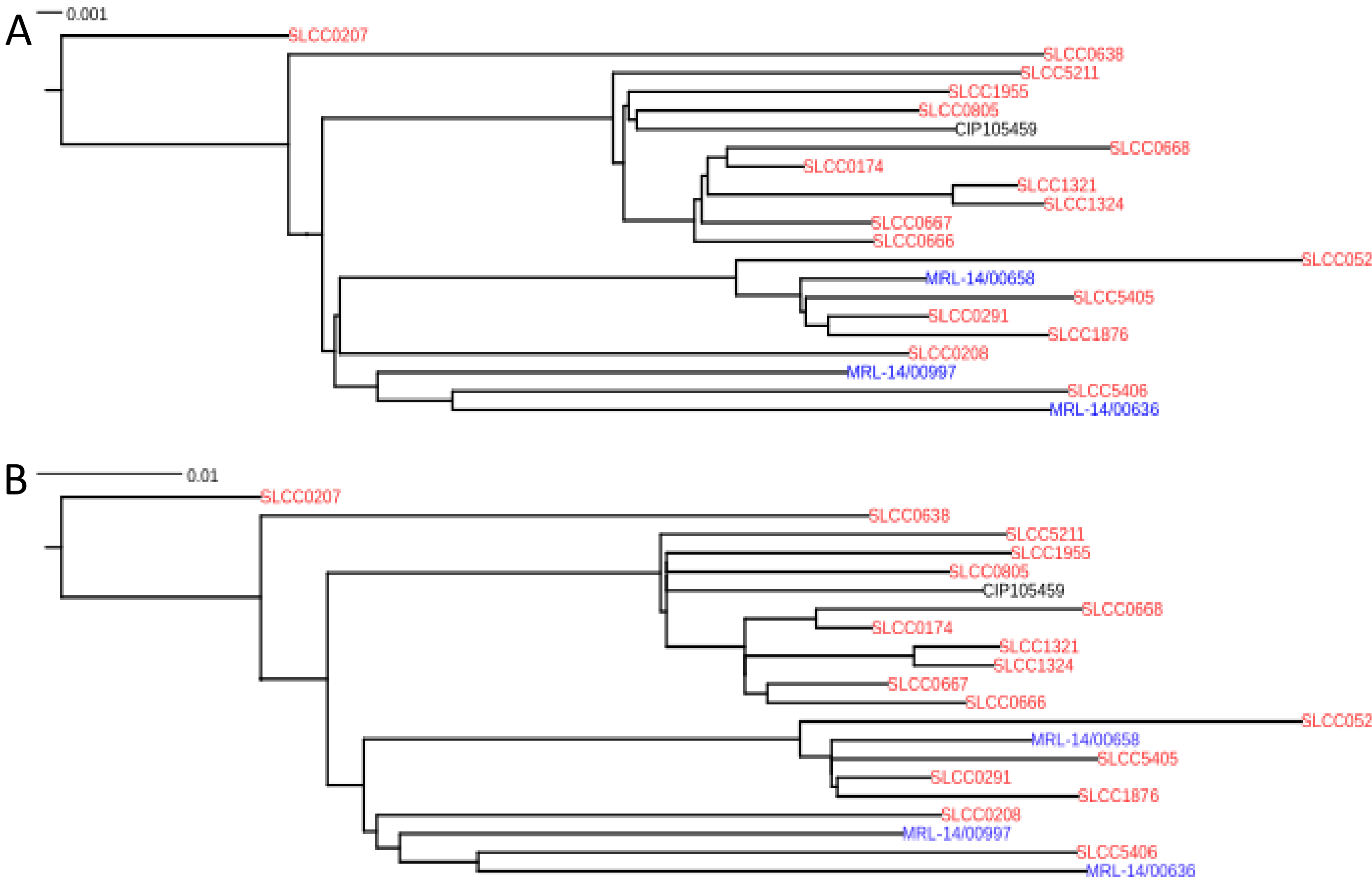

Figure 19: Difference of cgMLST analysis (A) and SNP based whole-genome analysis (B) pictured in a phylogenetic tree of a close related *L. monocytogenes* serogroup IVb cluster. Red strain names are SLCC isolates and blue names indicate human isolates from 2014. CIP105459 is a reference strain with the serotype 4e. Some branches and leaves have different lengths, the overall result is similar. Visualization was done on itol.embl.de.

Figure 19AB shows, that SNP based whole-genome analysis and cgMLST produce similar results. Although the connection of the branches and leaves is different between both phylogenetic trees in figure 19A and figure 19B, the relative distances linking two isolates are similar in both methods. Hence, the simplification to alleles is negligible for phylogenetic trees. The SNPs in coding regions are seen to be evenly distributed over the core genome, which itself represents the mutation events nicely. The distribution of SNPs over the reference genome is shown in figure 20, which supports the notion that mutations are evenly distributed.

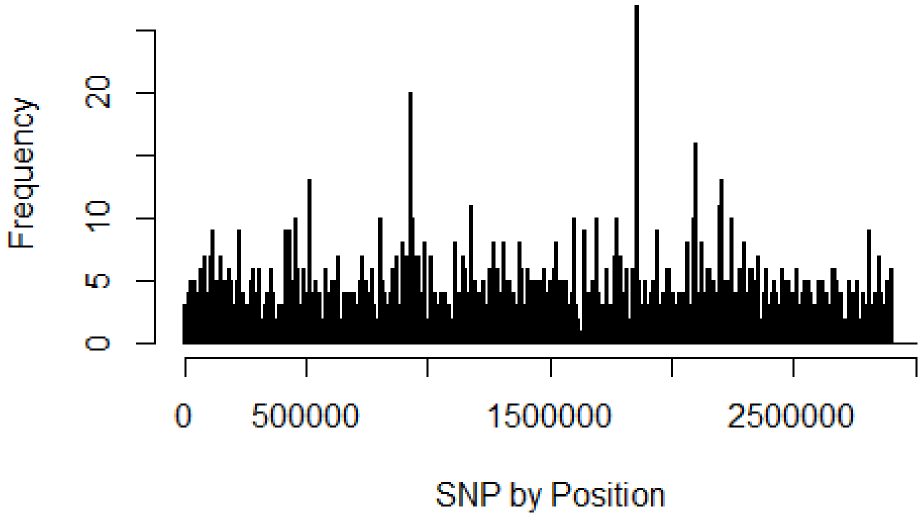

Figure 20: Distribution of SNPs over the reference genome of *L. monocytogenes* strain J1-220. 1042 SNP positions of the 21 strains calculated to a phylogenetic tree in figure 20 were accounted.

4.9. Outbreak investigation using cgMLST and SNP based phylogenetics

A few weeks after the SLCC isolates were sequenced, an outbreak was detected in Germany (Ruppitsch *et al.*, 2015b). The PFGE pattern matched a few isolates, which were collected by the Austrian-German binational reference laboratory (KL) for *Listeria* at the Austrian Agency for Health and Food Safety (AGES). Apart from 14 clinical isolates of which three were from Austria and 12 from Germany, 69 isolates from food sources from Austrian producers were sequenced. The food isolates were collected between 2011 and 2015, while the clinical isolates were collected between 2010 and 2015. The majority of clinical isolates, especially the isolates from Germany 2015 did not match any food isolates in their cgMLST cluster type (figure 21).

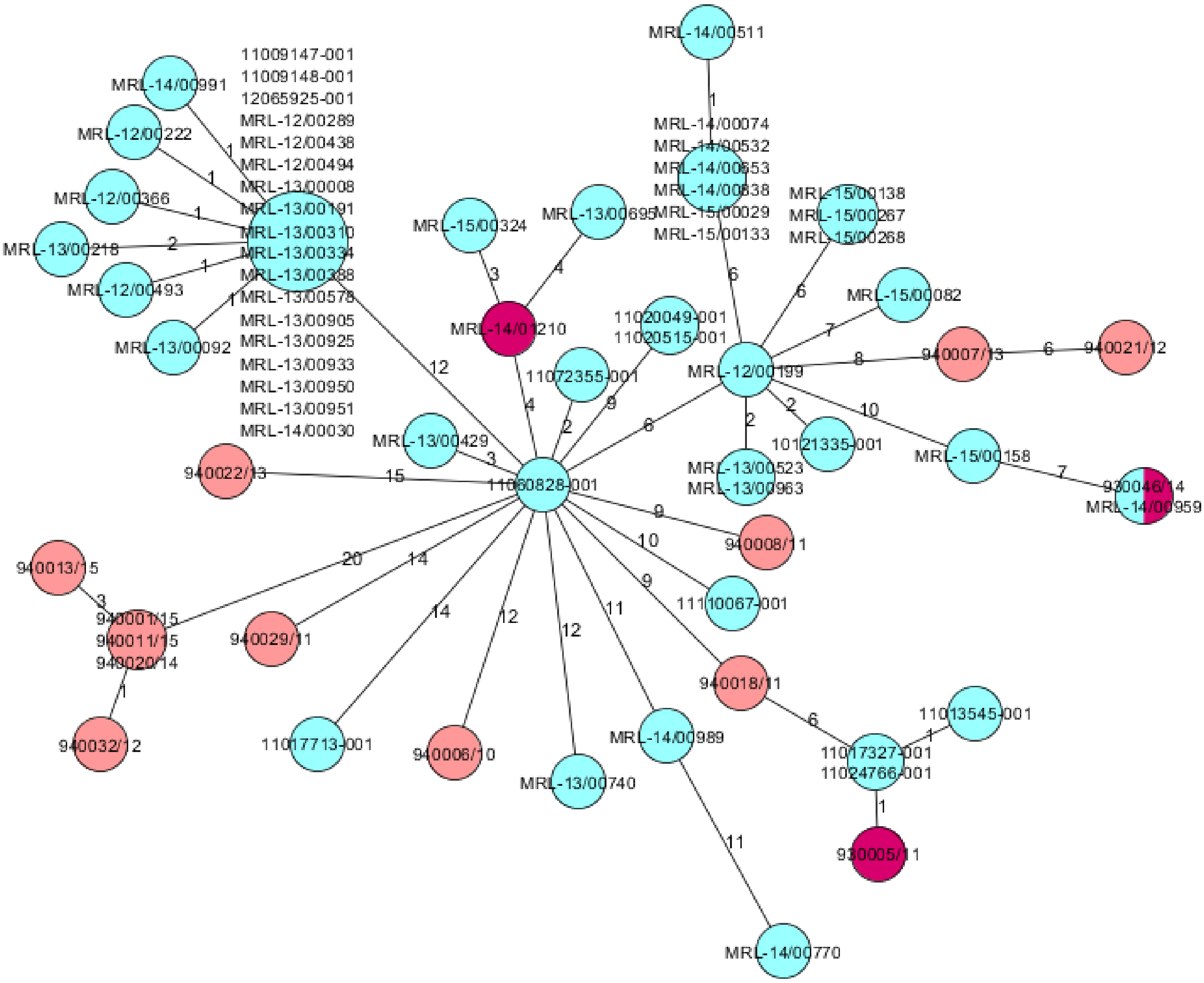

Figure 21: Minimum spanning tree of core genome MLST analysis of isolates with similar PFGE pattern as outbreak strains from Germany 2015. Blue nodes are food-related isolates, dark pink nodes are clinical isolates from Austria and light pink nodes are clinical isolates from Germany.

In the bottom left corner of figure 21, the four outbreak strains of 2014 and 2015 (940001/15, 940011/15, 940013/15 and 940020/14) are clustered together with an isolate from 2012 (940032/12). A large number of food isolates share the same cgMLST-type between the years 2011 and 2014 in the upper left side of figure 21. On the right side of figure 21, a food isolate (MRL-14/00959) and a clinical isolate from Austria (930046/14), both collected in 2014, are indistinguishable by cgMLST, while below that pair a few food isolates (11017327-001, 11024766-001 and 11013545-001) differ only in one and two alleles from an Austrian clinical isolate, respectively (930005/11), all collected in 2011.

To validate the results of cgMLST, a SNP-based analysis was performed on those isolates. The whole SNP-based circular phylogenetic tree is shown in figure 22.

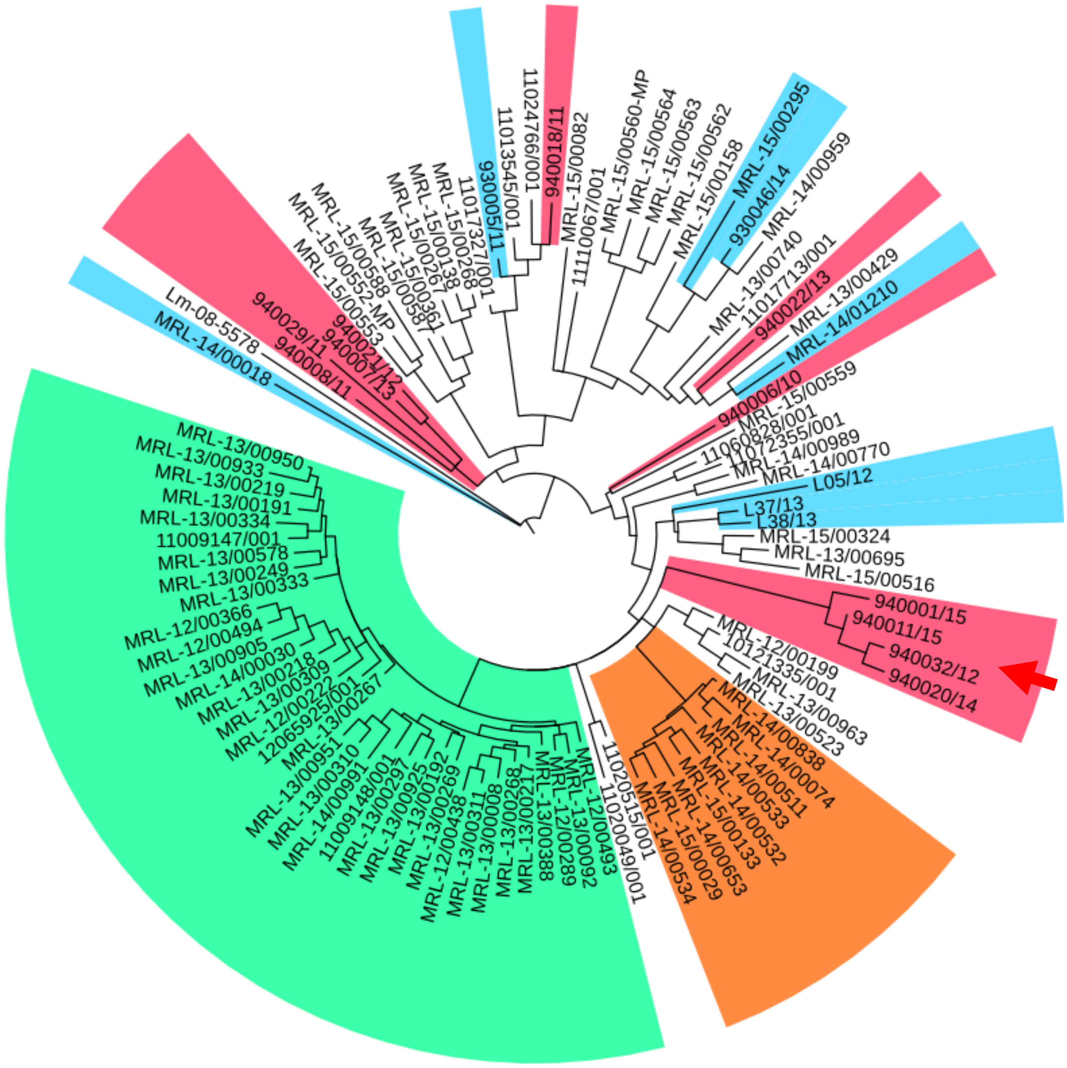

Figure 22: Phylogenetic circular tree calculated from whole-genome SNP based comparison of clinical isolates, food isolates and outbreak strains from Germany 2015. Pink color indicates clinical isolates from Germany, blue labels indicate clinical isolates from Austria and green and orange label clusters of food isolates which are highly similar in their cgMLST profile.

The large number of food isolates, which was indistinguishable in PFGE and cgMLST shows slight variations within their genomes on the SNP level (green color in figure 22). The Outbreak cluster on the right side of figure 22 (indicated by the red arrow) can be seen as closely related in the SNP-based tree as well as in the cgMLST based minimum spanning tree.

5. Discussion

5.1. Re-Cultivation of SLCC *Listeria monocytogenes* isolates

By re-cultivation of about 10 % subset of the whole SLCC, a total of 484 out of 647 strains (74.8 %) were successfully re-cultivated, while 163 strains could not be re-cultivated. The majority (90) did not grow in liquid media after incubation at 37 °C for at least 7 days. About half of that number (44) showed no hemolytic activity or showed untypical morphology like yellowish color or an irregular or filamentous form. The rest (30) showed no growth or no blue color on Rapid L'mono selective agar. Therefore 74 strains in total are unlikely to be *L. monocytogenes* or contaminated to a high degree with different microorganisms. In many cases the agar of stab cultures even was partially or completely covered with mold.

Compared to a previous study for re-cultivation of the SLCC (Haase *et al.*, 2011), a lower percentage (74.8 % vs. 85.8 %) of isolates was successfully reactivated. One reason is that the SLCC is not only composed by *L. monocytogenes* strains, but also a small fraction of other *Listeria* species, which is not always documented for each isolate. While the original study attempted to re-cultivate all of the SLCC isolates, my focus was set to *L. monocytogenes* exclusively. In addition there was a significant difference reported between re-cultivation of lyophilisates (96 %) and stab cultures (71 %). Haase *et al.* therefore focused on lyophilisates, while I used mainly stab cultures for re-cultivation. Considering these values the expected rate, 419 out of 591 stabs and 53 out of 56 lyophilisates should be cultivatable. In total 472 out of 648 isolates or 73 % would be expected, which is only slightly lower than the resuscitation performance in this thesis.

5.2. Serogroup determination by multiplex PCR

As shown in the results (figure 3), the majority of isolates of the SLCC fitted the expected serogroup in the multiplex PCR. While achieving 85 % seems satisfactory, the method was originally evaluated to identify 96-100 % of the isolates correctly (Doumith *et al.*, 2005). The source of the discrepancy is difficult to elicit. In respect of the age of the isolates, one can raise the question if it was possible for isolates stored in stab cultures over decades to change their serotype or serogroup. The authors of this PCR assay used the genes *lmo0737* and *lmo1118* from *L. monocytogenes* strain

EGD-e and *ORF2819* and *ORF2110* of *L. monocytogenes* strain CLIP 80459 (Doumith *et al.* 2004). While the EGD-e genes are specific for serogroups IIa and IIc, the CLIP 80459 genes are specific for serogroups IIb and IVb, which leads to the conclusion that it is highly unlikely that a strain can actually change its serogroup between IIa and IIb or IIc and IVb. The exchange in serogroup between IIa and IIc on the other hand seems rather sensible. The PCR result is IIc when both *lmo0737* and *lmo1118* are present, while IIa is *lmo0737* only. It seems possible that one gene can be lost over the time, notably there is a higher number changing from IIa to IIc than *vice versa*. The explanation can be found in the choice of EGD-e as reference strain, which is of serotype 1/2a but in serogroup IIc instead of IIa as would be expected. Moreover the serotype 1/2a is far more prominent than 1/2c which explains why the total number of isolates changing from the larger amount (1/2a) to the smaller amount (1/2c) is higher.

Other possibilities are the error made in serotype determination before the stab cultures were inoculated or cross-contamination of isolates. In the history of the Special Listeria Culture Collection (SLCC) each strain was recultivated in a series of inoculation of new stab cultures. Further “stab cultures” as preservation method closed with a rubber stop might not be the best choice to outlast for at least three decades. As the moldy surface of some agar cultures showed, contamination occurred in many ways. Also it is impossible to trace back if and how often a glas vial had been opened in the past. Cross-contamination is a serious problem in this strain collection.

In order to make sure that the strain is the same as documented by Seeliger, strains with different reported serotypes and experimentally observed serogroups were excluded from whole-genome sequencing. Although isolates with a serotype serogroup match might be as well cross contaminated, since the possibility to be contaminated with strains of the same group is quite high for four different groups.

On the other hand, a detailed whole-genome sequence analysis of serotype 1/2a isolates in serogroup IIc and serotype 1/2c isolates in serogroup IIa would be highly interesting and could reveal the genes determining serogroup 1/2c and 1/2a.

5.3. Quality assessment of read data

The quality of sequencing raw data can be estimated by different parameters. While in first generation sequencing each sequence was manually reviewed, in next generation sequencing monitoring quality of millions of reads is only possible in terms of statistics. Using FastQC software different figures are produced, one of them is shown in figure 4. The quality pictured in this figure shows an average samples quality distribution over sequence lengths. Without experience or reference it is impossible to state the level of quality. In principle this figure shows the need of quality trimming software to remove low quality bases from the 3’ end of each sequence. Apart from low quality bases, FastQC also shows remainders of technical sequences like sequencing adapters and indexes, which need to be removed (Bolger *et al.* 2014).

Following the instructions of the Nextera XT Library Preparation Kit size selection of the shattered fragments is done by magnetic beads in the “normalization” step. This resulted in insert sizes around 300 bp (294 bp on average), which was the sequenced length. Therefore over 50% of all sequences produced most certainly identical information for the forward and reverse sequence. Moreover, remaining sequences of adapters, primers or indices used for sequencing should be included on the 3’ end of most sequences. Hence, the FastQC results show different conditions (figure 6 and 7). The actual average size of sequences generated by the MiSeq for this isolate is about 251 bp. Also, a local maximum of the curve is around 250 bp. The low amount of adapter- and index-sequences indicates efficient trimming during read processing on the MiSeq software, otherwise large numbers of short sequences dramatic adapter content would be expected.

The sequences downloaded from the ENA had an average of 341 bp, while only 250 bp were sequenced. Considering that paired-end sequenced should mean sequencing of two not overlapping but closely distant reads, the true advantage of this technology is gone. Moreover, if the insert between sequencing indices and adapters is smaller than the read length, obviously the reads would contain adapter sequences which can cause difficulties in assembling or lead to artefacts in genome sequences. This so-called “read-through” seemed to be a minor problem on my data, even though sequencing length was clearly longer than the sequences obtained. Also, the average sequencing quality decreases rapidly after 250 bp. It is possible to generate 300 bp long reads, although it might be more sensible to sequence 250 bp instead, which would take less time.

Further, it seems that the average insert size is suboptimal. This could be altered by adopting the Illumina Nextera XT Library Preparation Kit by controlling the fragments size after indexing instead of the “normalization” size selection step using magnetic beads. Another option is to sequence less than 300 bp to guarantee high quality reads.

5.4. Contaminations

The contamination levels were determined in a two-step process. By collecting reads which were not mapped to *de-novo* assembled contigs, two assumptions were made. Possible contaminations would be represented by fragments in a dramatically lower concentration than the targeted sequences. Therefore foreign DNA sequences would not be able to assemble individual contigs. Only reads of the targeted sequence would be mapped to the contigs, and the contamination would not be able to align to contigs of *Listeria monocytogenes* origin.

By measuring only unmapped reads a large amount of reads is neglected. Hence, each contig was analyzed for possible contaminations in the second step using Tera-BLAST^{TM}. The reads of returned regions from a different genus with identities higher than 90 % were quantified. The results generated by this method consisted of two types: contaminations and false calls.

Many *de-novo* assemblies have short contigs which could be fully covered by single reads, especially with 300 bp long sequences. The majority of small contigs with less than 1,000 bp total size was successfully aligned to *Listeria monocytogenes* genomes. Nevertheless some contigs showed similarities to a different genus. In isolates L16-12 and L22-12 such small contigs are examples for contaminations in the assemblies. The contaminated contigs in those two assemblies have coverages of more than 1,000, which is at least ten times higher than other contigs.

Falsely called contaminations in this context are parts of a larger contig embedded in *Listeria* sequence with high similarities to different organisms. After revision of the Tera-BLAST^{TM} results those pseudo-contaminations are lower ranked hits. They cover only parts of the full sequence which itself is completely aligned to a *Listeria* reference sequence. In isolates MRL-14/00747, R-61951 and 930046/14 the *Enterococcus* hits are false calls. Those sequences highlight a general problem counting contaminated reads by using BLAST. Highly conserved genome regions or genes acquired via horizontal gene transfer are easily wrongly classified. As contaminations, such reads could be successfully mapped to similar regions in the target genome (Frazer *et al.*, 2003). For unmapped reads the BLAST result can be misinterpreted.

The results indicate that most isolates have only marginal amounts of contaminations. A vast majority of those reads which were regarded as potential contaminations, could not be identified by BLAST. They were neither assignable to *Listeria* nor to other species and therefore could not be proofed to be contaminations.

5.5. Comparison of different *de-novo* assembling tools using cgMLST

For core genome multi locus sequence typing (cgMLST), sequence data has to be assembled to draft genomes first. For this task, a large variety of tools and software is available today. The software SeqSphere$^+$, which was used for cgMLST analysis, comes with a reference mapper (BWA SW) and a *de-novo* assembler (Velvet). To find the best method for generating assemblies for cgMLST, two other assemblers were applied: SPAdes and MaSuRCA. In the GAGE-B comparison study, those two *de-novo* assemblers were the most accurate (Magoc *et al.* 2013). Since *Listeria monocytogenes* genomes can show large differences between serotypes, not only one single reference was chosen but 57 complete genomes were downloaded from the GenBank archive (table 4, material and methods). For each individual dataset, the best reference was determined before cgMLST analysis (listed in table 7). Also the newer algorithm BWA-MEM was used instead of BWA-SW which was reported to be faster and more accurate (Li, 2013).

Each method was tested on the same 90 isolates with subsequent core genome multi locus sequence typing (cgMLST): Surprisingly the allelic profile was not the same for all four assemblies of one paired-end read dataset. Discrepancies occurred in 10 out of 90 isolates. In eight datasets, one of four assemblies differed in one allele: six for BWA-MEM and two for MaSuRCA assemblies, one assembly created by BWA-MEM differed in two alleles. In one dataset MaSuRCA and BWA-MEM resulted in the same allelic profile, differing in one allele to the Velvet and SPAdes assemblies. Failed or missing genes were ignored for distance calculation as well as in SeqSphere$^+$.

More differences between assemblies were encountered where genes were classified as “failed” or “missing”. Genes which are not covered by reads over the whole consensus sequence were called “missing”, genes with frameshifts or different consensus lengths than the reference were classified as “failed”. The average number of genes “failed” or “missing” for MaSuRCA assemblies was 10, for BWA-MEM assemblies 8.2, for Velvet assemblies 7.3 and for SPAdes assemblies 6.9.

As already shown in the GAGE-B study (Magoc *et al.* 2013), assemblies created by different assemblers, sometimes bear large differences. The by far highest number of genes missing was found in MaSuRCA assemblies. This is surprising, as this assembler was found to work best for several other organisms compared on classical assembly comparison parameters (Magoc *et al.*, 2013). Admittedly the read pre-processing was unfavorable for MaSuRCA, according to the manual, which requires untrimmed reads and furthermore, unpaired reads could not be used by MaSuRCA. This might be the only reason why a remarkably large number of genes were missing. In fact other assembler comparison studies struggled with different requirements of each tool before (Jünemann *et al.*, 2014). However, both MaSuRCA and SPAdes support untrimmed reads and use their own error-correction algorithm. In the SPAdes assembler this is an optional process, while it is an important component of the algorithm creating super-reads in MaSuRCA. However, Velvet and BWA lack any read pre-processing, therefore quality trimming using Trimmomatic was chosen as initial task.

In general the averages of missing and/or failed genes are in a close range considering the total amount of 1701 genes tested. 10 defective genes in one assembly is still less than 0.6 %. Independent of the tools used for assembling the read libraries used in this thesis, cgMLST is a highly robust method to compare whole-genome data. As expected, the impact of different assemblers on assembly is insignificant which might be an implication of the high quality Illumina MiSeq reads. As shown by Jünemann *et al.* (2014), IonTorrent sequencing data is quite sensitive to the assembler used.

In contrast to missing genes or genes with frameshifts, false allele numbers have an impact on phylogenetic trees or minimum spanning trees, based on cgMLST data.

Even though some methods produce different allelic profiles in one or two positions, the impact on the result was marginal.

However, the *de-novo* assemblers Velvet and SPAdes seem to be the best choice for cgMLST analysis. This conclusion is valid for Illumina MiSeq reads of

L. monocytogenes with a length of 300 bp, since the choice of the right assembler strongly depends on the sequencing platform and organism (Magoc *et al.*, 2013). Due to significant advantages in computing time (Jünemann *et al.*, 2014) Velvet is preferred over SPAdes and was used as the standard assembler for cgMLST in this study.

Using BWA-MEM assemblies for cgMLST resulted in the highest number of genes classified as "failed" and assemblies with different allelic profiles than the *de-novo* assemblers. There is a relatively high chance to get wrong called alleles, which contribute to total distances between genomes. However, mapping reads to the *de-novo* assemblies is an important step where BWA-MEM is a good choice. Using read-mapped BAM files for cgMLST allows assessing the quality of genes found. When Velvet is used in SeqSphere^{+} read-mapping information is automatically included by creating the AMOS file, but if different *de-novo* assemblers are used read mapping is recommended.

5.6. Core genome multi locus sequence typing

Based on allelic profiles for 501 *Listeria monocytogenes* genomes a phylogenetic tree was calculated and exported from SeqSphere^{+} to visualize all isolates in one tree (figure 14).

The apparent higher diversity of the IIa clade could be a result of the biased core genome, which was created using EGD-e as reference, which belongs to the IIc serogroup and serotype 1/2a. Hence apart from the possibility that the lab strain EGD-e lost several genes, the core genome would not be dramatically different since a core genome by definition is the set of genes represented in all strains of a certain organism.

The average variance in alleles of the 1,701 genes belonging to the core genome was relatively high. The median of gene loci had 25 different alleles in 432 isolates. 69 isolates sequenced for a recent listeriosis outbreak were removed to eliminate a possible bias in counting the most frequent allele. Half of the genes had between 17 and 34 different alleles, while for two genes (*lmo2609*, *lmo2614*) only one sequence was found. The numbers show, that for a broad range of genes in the core genome similar mutation rates apply. Still the outliers are as interesting showing a few genes with high selection pressure and low genetic variance possible. The gene product behind the locus *lmo2609* is rpmJ, gene locus *lmo2614* encodes rpmD. Both are subunits to the 50S ribosomal protein. In the top 13 with a maximum of three different alleles 10 genes are encoding protein subunits of the 30S and the 50S ribosome. Further cspL, a *Listeria* cold shock protein and atpE, encoding subunit C of the ATP-synthase have only three different alleles.

This fits the expectation that ribosomal genes are the genes that are best conserved. The cold shock protein also seems to be a crucial factor for *L. monocytogenes*. As already mentioned *L. monocytogenes* is able to grow relatively rapid at low temperatures. This has been exhausted by cold enrichment techniques in the past (Curtis and Lee, 1995). Therefore a high conservation of this protein was no surprise either.

With only three different alleles and 77 % of isolates showing one single genotype also *lmo2689a* was one of the most conserved gene loci. The locus is annotated as “hypothetical protein” in GenBank. The function of the gene is currently unknown, but the protein’s existence was predicted by the annotation pipeline. The result, that only three different alleles were found for this gene supports the hypothesis that this gene exists. It is very unlikely that a gene locus without any functionality is conserved like this gene is. The gene might play an important role as well, which has not been discovered yet.

On the opposite end of table 7, which show the total numbers of different alleles, the genes have more diverse functions. Surprisingly, the gene encoding the enzyme lactose-dehydrogenase (*ldh*) had the most variants. This gene was chosen as one of seven housekeeping genes for classical MLST of *L. monocytogenes*.

5.7. Evolutionary distance between SLCC isolates and recent clinical isolates

As shown in figure 17, some historic SLCC strains are relatively close related to recent strains. Also this group of isolates is in distance to other isolates, which indicates that those relations are not random but their genomes do have many genes in common.

The number of SNPs and indels would result in the need of vast computation power to calculate differences. Still there is a chance for SNP analysis of the *L. monocytogenes* genome, as the majority of reads could be mapped, assuming that a small number of isolates is compared. At least a close related and trusted reference genome would be needed to minimize the total amount of SNPs and maximize the reads successfully mapped.

With complete genomes, which will be obtained from sequencing in future using long read sequences (Bashir *et al.*, 2012), a different approach than SNP calling would still be needed to compare distant *L. monocytogenes* strains, which evaluates re-arrangement in chromosomes, deletions, insertions and SNPs at the same time. The simplification to break down genetic variation to sequence types is a valid approach to compare distantly related strains or specimen of the same species.

5.8. Cluster investigation by cgMLST and single nucleotide polymorphism (SNP) based analysis in comparison

Whole-genome sequence data allows a large variety of analysis methods and several methods for subtyping of re-sequenced organisms. Single-nucleotide polymorphism or variants (SNP/SNV) based methods have been used to differentiate organisms on genus and species level in the past. The expansion of the method to whole-genome comparison bears difficulties, since high throughput sequencing produces in most cases only draft assemblies. Core genome MLST on the other hand is a relatively new approach for genotyping on a whole-genome basis. In order to validate the cgMLST principle SNP based comparison was performed on a small subset of the whole-genome sequence data of all samples.

SNP-based comparison provided a higher resolution for the samples to be compared with 1,042 variant positions, which was more than twice the 437 cgMLST genes, which varied. Assuming, that each gene variant differed in one SNP, at least 437 SNPs were accounted in protein coding regions, while 605 SNPs were located in non-coding or non-core genome coding genes. This is not what would be expected, since non-coding regions are said to have higher mutation rates. However, the SNP distribution was equal over the genome as shown in figure 20. It is possible that the intergenic regions have varied more than what could be successfully aligned by reference mapping, which would be an explanation for these numbers.

In addition 437 of 1,701 genes is about a quarter of all core genome genes (25.7 %), which points to a relatively large evolutionary distance between those isolates, while

1,042 variant positions are less than 0.1 % of the total number of nucleotides, which can be considered a rather small distance. It seems that those 1,042 positions are just a small amount of all variant positions and mutations that occurred in those compared 21 genomes. However, the 437 genes with different sequence types might also be just a subset of all genes, which could be compared in the same 21 genomes. For phylogenetic calculations, both the cgMLST and SNP based analysis resulted in similar trees and therefore the core genome MLST for *Listeria monocytogenes* in this study should be considered to be valid for subtyping of this species.

5.9. Outbreak investigation using cgMLST and SNP based phylogenetics

To detect an outbreak, a highly discriminative method is needed. Until today, PFGE is the standard analysis used for this. To identify the source(s) of an outbreak, food isolates have to be found, which have the same PFGE pattern and further receipts have to be collected which document that a certain product was consumed. A matching PFGE pattern alone is no proof. This was also described for the “acid curd cheese outbreak” in 2009 and 2010 in Austria (Ruppitsch *et al.*, 2015a). By using whole-genome sequencing and cgMLST, a single PFGE pattern group was divided in several clusters.

For the ongoing outbreak in Germany 2015 both PFGE and cgMLST where applied (Ruppitsch *et al.*, 2015b). Several clinical isolates from Austria and Germany and food isolates from Austria where sequenced for this outbreak. In addition to cgMLST the whole-genome sequencing data was used for SNP based analysis.

The food isolates sequenced in this study did not match the outbreak cluster, which is about 20 alleles away from the closest Austrian isolate, although they show the same PFGE pattern. Only two clinical isolates matched a food isolate in their cgMLST profile and in year of occurrence: 930005/11 was only one allele different to two food isolates. 930046/14 shares the cgMLST profile with a food isolate strain. It is possible, that those items were the sources of infection in both cases, but impossible to proof after the time period that has elapsed since then.

The results indicate that one Austrian food source is not involved in this outbreak, although there are isolates with the same PFGE pattern. However, for most single cases, a correlated food isolate and the source could not be found. Further, in several reported outbreaks a contaminated food source can bear different clonal types (Mead *et al.*, 2006). Therefore those sources could not be ultimately excluded as source of infection.

6. Conclusion

During this thesis, I was able to re-cultivate a small fraction of the historic Special Listeria Culture Collection (SLCC) of which 484 out of 647 isolates are now available in cryotubes and frozen at -80°C. I proved that for 74.8 % of those isolates it was still possible to cultivate those stab-cultures. 191 of those isolates were whole-genome sequenced using Illumina short read technology and the sequence data will be available for public access in the future. The raw assembly of the oldest isolate (SLCC208) has already been published (Accession No: NZ_LMXJ00000000.1, Hyden *et al.*, 2016). The sequence data and the preserved strains might be of use in future to analyse the recent past of *L. monocytogenes*.

A central part of this thesis was whole-genome sequence generation and genotyping of *L. monocytogenes* strains by core genome multi locus sequence typing (cgMLST). As a result, I aimed to find the optimal assembly strategy by comparing reference based assembly, MaSuRCA, SPAdes and Velvet with a dataset of 90 isolates. For *L. monocytogenes* and cgMLST SPAdes and Velvet produced the best results, while Velvet might be the method of choice, since it was reported to be faster. However, the best assembly engine is always a matter of sequence data available and might depend on the organism. The latest sequencing technologies promise to overcome problems of short read sequence assembly and might result in more complete genomes.

Another important aspect of this thesis was to use whole-genome sequencing with cgMLST and SNP analysis as genotyping methods. Further, a recent outbreak from Germany was in part investigated applying those methods. It was again proven, that next generation sequencing (NGS) is a more powerful and versatile tool than PFGE, which is the established gold standard technique for genotyping *L. monocytogenes*. Apart from the higher discriminative power, the greatest advantage of NGS is the almost unlimited possibilities for further analyses. Core genome MLST might replace PFGE and other typing methods in the near future for *L. monocytogenes* and other endemic or epidemic pathogens.

7. Literature

Achtman, M. (2008). Evolution, Population Structure, and Phylogeography of Genetically Monomorphic Bacterial Pathogens. Annual Review of Microbiology *62*, 53–70.

Aird, D., Ross, M.G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D.B., Nusbaum, C., and Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol. *12*, R18.

Allerberger, F., and Wagner, M. (2010). Listeriosis: a resurgent foodborne infection. Clin. Microbiol. Infect. *16*, 16–23.

Atkinson, M.R., Deutscher, M.P., Kornberg, A., Russell, A.F., and Moffatt, J.G. (1969). Enzymic synthesis of deoxyribonucleic acid. XXXIV. Termination of chain growth by a 2',3'-dideoxyribonucleotide. Biochemistry *8*, 4897–4904.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., et al. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol *19*, 455–477.

Bashir, A., Klammer, A.A., Robins, W.P., Chin, C.-S., Webster, D., Paxinos, E., Hsu, D., Ashby, M., Wang, S., Peluso, P., et al. (2012). A hybrid approach for the automated finishing of bacterial genomes. Nat Biotech *30*, 701–707.

Batzoglou, S., Jaffe, D.B., Stanley, K., Butler, J., Gnerre, S., Mauceli, E., Berger, B., Mesirov, J.P., and Lander, E.S. (2002). ARACHNE: a whole-genome shotgun assembler. Genome Res. *12*, 177–189.

Bayley, H. (2015). Nanopore sequencing: from imagination to reality. Clin Chem *61*, 25–31.

Bentley, D.R. (2006). Whole-genome re-sequencing. Curr. Opin. Genet. Dev. *16*, 545–552.

Bergey, D.H., Krieg, N.R., and Holt, J.G. (1984). Bergey's manual of systematic bacteriology (Baltimore, MD: Williams & Wilkins).

Berkrot, B. (2013). Thermo Fisher to buy Life Tech for $13.6 billion. Reuters.

Bille, J., Blanc, D.S., Schmid, H., Boubaker, K., Baumgartner, A., Siegrist, H.H., Tritten, M.L., Lienhard, R., Berner, D., Anderau, R., et al. (2006). Outbreak of human listeriosis associated with tomme cheese in northwest Switzerland, 2005. Euro Surveill. *11*, 91–93.

Bio-IT World (2013). Six Years After Acquisition, Roche Quietly Shutters 454 - Bio-IT World.

biomickwatson What does the PacBio Sequel mean for the future of sequencing?. Accessed on 22. Feb 2016 on http://www.opiniomics.org/what-does-the-pacbio-sequel-mean-for-the-future-of-sequencing/

Boisvert, S., Raymond, F., Godzaridis, É., Laviolette, F., and Corbeil, J. (2012). Ray Meta: scalable de novo metagenome assembly and profiling. Genome Biology *13*, R122.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

Bradnam, K.R., Fass, J.N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., Boisvert, S., Chapman, J.A., Chapuis, G., Chikhi, R., et al. (2013). Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. Gigascience *2*, 10.

Bragg, L.M., Stone, G., Butler, M.K., Hugenholtz, P., and Tyson, G.W. (2013). Shining a Light on Dark Sequencing: Characterising Errors in Ion Torrent PGM Data. PLoS Comput Biol *9*.

Branton, D., Deamer, D.W., Marziali, A., Bayley, H., Benner, S.A., Butler, T., Di Ventra, M., Garaj, S., Hibbs, A., Huang, X., et al. (2008). The potential and challenges of nanopore sequencing. Nat Biotech *26*, 1146–1153.

Carpentier, B., and Cerf, O. (2011). Review--Persistence of Listeria monocytogenes in food industry equipment and premises. Int. J. Food Microbiol. *145*, 1–8.

Centers for Disease Control and Prevention (CDC) (2011). Outbreak of invasive listeriosis associated with the consumption of hog head cheese--Louisiana, 2010. MMWR Morb. Mortal. Wkly. Rep. *60*, 401–405.

Centers for Disease Control and Prevention (CDC) (2015). Multistate Outbreak of Listeriosis Linked to Commercially Produced, Prepackaged Caramel Apples Made from Bidart Bros Apples (Final Update).

Chaisson, M., Pevzner, P., and Tang, H. (2004). Fragment assembly with short reads. Bioinformatics *20*, 2067–2074.

Compeau, P.E.C., Pevzner, P.A., and Tesler, G. (2011). How to apply de Bruijn graphs to genome assembly. Nat Biotech *29*, 987–991.

Curtis, G.D.W., and Lee, W.H. (1995). Culture media and methods for the isolation of Listeria monocytogenes. International Journal of Food Microbiology *26*, 1–13.

Disson, O., Grayo, S., Huillet, E., Nikitas, G., Langa-Vives, F., Dussurget, O., Ragon, M., Le Monnier, A., Babinet, C., Cossart, P., et al. (2008). Conjugated action of two species-specific invasion proteins for fetoplacental listeriosis. Nature *455*, 1114–1118.

Doijad, S.P., Barbuddhe, S.B., Garg, S., Poharkar, K.V., Kalorey, D.R., Kurkure, N.V., Rawool, D.B., and Chakraborty, T. (2015). Biofilm-Forming Abilities of Listeria monocytogenes Serotypes Isolated from Different Sources. PLoS ONE *10*, e0137046.

Donker-Voet, J. (1959). A serological study on some strains of Listeria monocytogenes isolated in Michigan. American Journal of Veterinary Research *20*, 176–179.

Doumith, M., Buchrieser, C., Glaser, P., Jacquet, C., and Martin, P. (2004). Differentiation of the major Listeria monocytogenes serovars by multiplex PCR. J. Clin. Microbiol. *42*, 3819–3822.

Doumith, M., Jacquet, C., Gerner-Smidt, P., Graves, L.M., Loncarevic, S., Mathisen, T., Morvan, A., Salcedo, C., Torpdahl, M., Vazquez, J.A., et al. (2005). Multicenter validation of a multiplex PCR assay for differentiating the major Listeria monocytogenes serovars 1/2a, 1/2b, 1/2c, and 4b: toward an international standard. J. Food Prot. *68*, 2648–2650.

Dussurget, O., Cabanes, D., Dehoux, P., Lecuit, M., Buchrieser, C., Glaser, P., Cossart, P., and European Listeria Genome Consortium (2002). Listeria monocytogenes bile salt hydrolase is a PrfA-regulated virulence factor involved in the intestinal and hepatic phases of listeriosis. Mol. Microbiol. *45*, 1095–1106.

Earl, D., Bradnam, K., St. John, J., Darling, A., Lin, D., Fass, J., Yu, H.O.K., Buffalo, V., Zerbino, D.R., Diekhans, M., et al. (2011). Assemblathon 1: A competitive assessment of de novo short read assembly methods. Genome Res *21*, 2224–2241.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., et al. (2009). Real-time DNA sequencing from single polymerase molecules. Science *323*, 133–138.

Emrich, C.A., Tian, H., Medintz, I.L., and Mathies, R.A. (2002). Microfabricated 384-Lane Capillary Array Electrophoresis Bioanalyzer for Ultrahigh-Throughput Genetic Analysis. Anal. Chem. *74*, 5076–5083.

Ewing, B., Hillier, L., Wendl, M.C., and Green, P. (1998). Base-Calling of Automated Sequencer Traces UsingPhred. I. Accuracy Assessment. Genome Res. *8*, 175–185.

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

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., and Merrick, J.M. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science *269*, 496–512.

Fleischmann, R.D., Alland, D., Eisen, J.A., Carpenter, L., White, O., Peterson, J., DeBoy, R., Dodson, R., Gwinn, M., Haft, D., et al. (2002). Whole-Genome Comparison of Mycobacterium tuberculosis Clinical and Laboratory Strains. J Bacteriol *184*, 5479–5490.

Forbes Analyst: The Better Desktop DNA Sequencer May Be Losing The Marketing War. Forbes.

Frazer, K.A., Elnitski, L., Church, D.M., Dubchak, I., and Hardison, R.C. (2003). Cross-Species Sequence Comparisons: A Review of Methods and Available Resources. Genome Res. *13*, 1–12.

Fretz, R., Pichler, J., Sagel, U., Much, P., Ruppitsch, W., Pietzka, A.T., Stöger, A., Huhulescu, S., Heuberger, S., Appl, G., et al. (2010). Update: Multinational listeriosis outbreak due to “Quargel”, a sour milk curd cheese, caused by two different L. monocytogenes serotype 1/2a strains, 2009-2010. Euro Surveill. *15*.

Fullwood, M.J., Wei, C.-L., Liu, E.T., and Ruan, Y. (2009). Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. Genome Res. *19*, 521–532.

Gandhi, M., and Chikindas, M.L. (2007). Listeria: A foodborne pathogen that knows how to survive. International Journal of Food Microbiology *113*, 1–15.

Gaulin, C., Ramsay, D., and Bekal, S. (2012). Widespread Listeriosis Outbreak Attributable to Pasteurized Cheese, Which Led to Extensive Cross-Contamination Affecting Cheese Retailers, Quebec, Canada, 2008. Journal of Food Protection *75*, 71–78.

GenomeWeb Website (2010). Life Technologies to Acquire Ion Torrent for up to $725M.

Giovannacci, I., Ragimbeau, C., Queguiner, S., Salvat, G., Vendeuvre, J.-L., Carlier, V., and Ermel, G. (1999). Listeria monocytogenes in pork slaughtering and cutting plants: use of RAPD, PFGE and PCR–REA for tracing and molecular epidemiology. International Journal of Food Microbiology *53*, 127–140.

Glaser, P., Frangeul, L., Buchrieser, C., Rusniok, C., Amend, A., Baquero, F., Berche, P., Bloecker, H., Brandt, P., Chakraborty, T., et al. (2001). Comparative genomics of Listeria species. Science *294*, 849–852.

Gnerre, S., MacCallum, I., Przybylski, D., Ribeiro, F.J., Burton, J.N., Walker, B.J., Sharpe, T., Hall, G., Shea, T.P., Sykes, S., et al. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci U S A *108*, 1513–1518.

Graves, L.M., and Swaminathan, B. (2001). PulseNet standardized protocol for subtyping Listeria monocytogenes by macrorestriction and pulsed-field gel electrophoresis. International Journal of Food Microbiology *65*, 55–62.

Haase, J.K., Murphy, R.A., Choudhury, K.R., and Achtman, M. (2011). Revival of Seeliger's historical "Special Listeria Culture Collection." Environ. Microbiol. *13*, 3163–3171.

Haase, J.K., Didelot, X., Lecuit, M., Korkeala, H., L. monocytogenes MLST Study Group, and Achtman, M. (2014). The ubiquitous nature of Listeria monocytogenes clones: a large-scale Multilocus Sequence Typing study. Environ. Microbiol. *16*, 405–416.

Hodkinson, B.P., and Grice, E.A. (2015). Next-Generation Sequencing: A Review of Technologies and Tools for Wound Microbiome Research. Adv Wound Care (New Rochelle) *4*, 50–58.

Hoffmann, S., Batz, M.B., and Morris, J. J. Glenn (2012). Annual Cost of Illness and Quality-Adjusted Life Year Losses in the United States Due to 14 Foodborne Pathogens. Journal of Food Protection *75*, 1292–1302.

http://www.fda.gov/Safety/Recalls/ (2015). Recalls, Market Withdrawals, & Safety Alerts.

Huhulescu, S. (2015). Listeriose Jahresbericht 2014 (Nationale Referenzzentrale für Listeriose).

Huse, S.M., Huber, J.A., Morrison, H.G., Sogin, M.L., and Welch, D.M. (2007). Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol *8*, R143.

Hyden, P., Pietzka, A., Allerberger, F., Springer, B., Sensen, C., and Ruppitsch, W. (2016). Draft Genome Sequence of a 94-Year-Old Listeria monocytogenes Isolate, SLCC208. Genome Announc *4*.

Jain, M., Fiddes, I.T., Miga, K.H., Olsen, H.E., Paten, B., and Akeson, M. (2015). Improved data analysis for the MinION nanopore sequencer. Nat. Methods *12*, 351–356.

Jünemann, S., Prior, K., Albersmeier, A., Albaum, S., Kalinowski, J., Goesmann, A., Stoye, J., and Harmsen, D. (2014). GABenchToB: A Genome Assembly Benchmark Tuned on Bacteria and Benchtop Sequencers. PLoS One *9*.

Karow, J. (2015). Survey Finds High Interest in Long-Range Genome Data, Nanopore Tech While Illumina Rules NGS Market.

Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L., and Wilson, R.K. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. *22*, 568–576.

Koren, S., and Phillippy, A.M. (2015). One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. Current Opinion in Microbiology *23*, 110–120.

Koren, S., Harhay, G.P., Smith, T.P., Bono, J.L., Harhay, D.M., Mcvey, S.D., Radune, D., Bergman, N.H., and Phillippy, A.M. (2013). Reducing assembly complexity of microbial genomes with single-molecule sequencing. Genome Biology *14,* R101.

Korlach, J., Marks, P.J., Cicero, R.L., Gray, J.J., Murphy, D.L., Roitman, D.B., Pham, T.T., Otto, G.A., Foquet, M., and Turner, S.W. (2008). Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. Proc Natl Acad Sci U S A *105*, 1176–1181.

Letunic, I., and Bork, P. (2011). Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. Nucl. Acids Res. *39,* W475–W478.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics *27*, 2987–2993.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997 [q-Bio].

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

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., et al. (2010). De novo assembly of human genomes with massively parallel short read sequencing. Genome Res *20*, 265–272.

Liao, Y.-C., Lin, S.-H., and Lin, H.-H. (2015). Completing bacterial genome assemblies: strategy and performance comparisons. Scientific Reports *5*, 8747.

Lienau, E.K., Strain, E., Wang, C., Zheng, J., Ottesen, A.R., Keys, C.E., Hammack, T.S., Musser, S.M., Brown, E.W., Allard, M.W., et al. (2011). Identification of a Salmonellosis Outbreak by Means of Molecular Sequencing. New England Journal of Medicine *364*, 981–982.

Loman, N.J., Misra, R.V., Dallman, T.J., Constantinidou, C., Gharbia, S.E., Wain, J., and Pallen, M.J. (2012). Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotech *30*, 434–439.

MacDonald, P.D.M., Whitwam, R.E., Boggs, J.D., MacCormack, J.N., Anderson, K.L., Reardon, J.W., Saah, J.R., Graves, L.M., Hunter, S.B., and Sobel, J. (2005). Outbreak of listeriosis among Mexican immigrants as a result of consumption of illicitly produced Mexican-style cheese. Clin. Infect. Dis. *40*, 677–682.

Magoc, T., Pabinger, S., Canzar, S., Liu, X., Su, Q., Puiu, D., Tallon, L.J., and Salzberg, S.L. (2013). GAGE-B: an evaluation of genome assemblers for bacterial organisms. Bioinformatics *29*, 1718–1725.

Maiden, M.C.J., Bygraves, J.A., Feil, E., Morelli, G., Russell, J.E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D.A., et al. (1998). Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci U S A *95*, 3140–3145.

Mangen, M.-J.J., Bouwknegt, M., Friesema, I.H.M., Haagsma, J.A., Kortbeek, L.M., Tariq, L., Wilson, M., van Pelt, W., and Havelaar, A.H. (2015). Cost-of-illness and disease burden of food-related pathogens in the Netherlands, 2011. International Journal of Food Microbiology *196*, 84–93.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-J., Chen, Z., et al. (2005). Genome Sequencing in Open Microfabricated High Density Picoliter Reactors. Nature *437*, 376–380.

McCollum, J.T., Cronquist, A.B., Silk, B.J., Jackson, K.A., O'Connor, K.A., Cosgrove, S., Gossack, J.P., Parachini, S.S., Jain, N.S., Ettestad, P., et al. (2013). Multistate outbreak of listeriosis associated with cantaloupe. N. Engl. J. Med. *369*, 944–953.

Mead, P.S., Dunne, E.F., Graves, L., Wiedmann, M., Patrick, M., Hunter, S., Salehi, E., Mostashari, F., Craig, A., Mshar, P., et al. (2006). Nationwide outbreak of listeriosis due to contaminated meat. Epidemiol Infect *134*, 744–751.

Metzker, M.L. (2010). Sequencing technologies — the next generation. Nat Rev Genet *11*, 31–46.

Michelon, D., Félix, B., Vingadassalon, N., Mariet, J.-F., Larsson, J.T., Møller-Nielsen, E., and Roussel, S. (2015). PFGE Standard Operating Procedures for Listeria monocytogenes: Harmonizing the Typing of Food and Clinical Strains in Europe. Foodborne Pathogens and Disease *12*, 244–252.

Mikheyev, A.S., and Tin, M.M.Y. (2014). A first look at the Oxford Nanopore MinION sequencer. Mol Ecol Resour *14*, 1097–1102.

Miller, J.R., Delcher, A.L., Koren, S., Venter, E., Walenz, B.P., Brownley, A., Johnson, J., Li, K., Mobarry, C., and Sutton, G. (2008). Aggressive assembly of pyrosequencing reads with mates. Bioinformatics *24*, 2818–2824.

Miller, J.R., Koren, S., and Sutton, G. (2010). Assembly Algorithms for Next-Generation Sequencing Data. Genomics *95*, 315–327.

Moorhead, S.M., Dykes, G.A., and Cursons, R.T. (2003). An SNP-based PCR assay to differentiate between Listeria monocytogenes lineages derived from phylogenetic analysis of the sigB gene. Journal of Microbiological Methods *55*, 425–432.

Morey, M., Fernández-Marmiesse, A., Castiñeiras, D., Fraga, J.M., Couce, M.L., and Cocho, J.A. (2013). A glimpse into past, present, and future DNA sequencing. Mol. Genet. Metab. *110*, 3–24.

Mullapudi, S., Siletzky, R.M., and Kathariou, S. (2008). Heavy-metal and benzalkonium chloride resistance of Listeria monocytogenes isolates from the environment of turkey-processing plants. Appl. Environ. Microbiol. *74*, 1464–1468.

nanoporetech.com (2015). Oxford Nanopore Technologies.

Online-News "The Foreigner.no" (2014). Sweden listeria outbreak shows decline.

Online-News "thelocal.dk" (2014). Listeria outbreak claims 15th victim.

Ontario Ministry of Health and Long-Term Care (2009). 2008 listeriosis outbreak in Ontario: epidemiologic summary (Ministry of Health and Long-Term Care).

Orsi, R.H., Bakker, H.C. den, and Wiedmann, M. (2011). Listeria monocytogenes lineages: Genomics, evolution, ecology, and phenotypic characteristics. International Journal of Medical Microbiology *301*, 79–96.

Paterson, J.S. (1940). The antigenic structure of organisms of the genusListerella. The Journal of Pathology and Bacteriology *51*, 427–436.

Pevzner, P.A., Tang, H., and Waterman, M.S. (2001). An Eulerian path approach to DNA fragment assembly. Proc Natl Acad Sci U S A *98*, 9748–9753.

Pollack, A. (2011). Rothberg Seeks to Make DNA Sequencing Common. The New York Times.

Press release EFSA, and ECDC Campylobacteriosis cases stable, listeriosis cases continue to rise, say EFSA and ECDC.

Ragon, M., Wirth, T., Hollandt, F., Lavenir, R., Lecuit, M., Le Monnier, A., and Brisse, S. (2008). A new perspective on Listeria monocytogenes evolution. PLoS Pathog. *4*, e1000146.

Rhoads, A., and Au, K.F. (2015). PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics *13*, 278–289.

Ribeiro, F.J., Przybylski, D., Yin, S., Sharpe, T., Gnerre, S., Abouelleil, A., Berlin, A.M., Montmayeur, A., Shea, T.P., Walker, B.J., et al. (2012). Finished bacterial genomes from shotgun sequence data. Genome Res *22*, 2270–2277.

Roche Website (2015). Products : 454 Life Sciences, a Roche Company.

Ronaghi, M., Uhlén, M., and Nyrén, P. (1998). A Sequencing Method Based on Real-Time Pyrophosphate. Science *281*, 363–365.

Rothberg, J.M., Hinz, W., Rearick, T.M., Schultz, J., Mileski, W., Davey, M., Leamon, J.H., Johnson, K., Milgrew, M.J., Edwards, M., et al. (2011). An integrated semiconductor device enabling non-optical genome sequencing. Nature *475*, 348–352.

Ruppitsch, W., Pietzka, A., Prior, K., Bletz, S., Fernandez, H.L., Allerberger, F., Harmsen, D., and Mellmann, A. (2015a). Defining and Evaluating a Core Genome Multilocus Sequence Typing Scheme for Whole-Genome Sequence-Based Typing of Listeria monocytogenes. J. Clin. Microbiol. *53*, 2869–2876.

Ruppitsch, W., Prager, R., Halbedel, S., Hyden, P., Pietzka, A., Huhulescu, S., Lohr, D., Schönberger, K., Aichinger, E., Hauri, A., et al. (2015b). Ongoing outbreak of invasive listeriosis, Germany, 2012 to 2015. Euro Surveill. *20*.

Rusk, N. (2011). Torrents of sequence. Nat Meth *8*, 44–44.

Salcedo, C., Arreaza, L., Alcalá, B., de la Fuente, L., and Vázquez, J.A. (2003). Development of a Multilocus Sequence Typing Method for Analysis of Listeria monocytogenes Clones. J Clin Microbiol *41*, 757–762.

Salipante, S.J., SenGupta, D.J., Cummings, L.A., Land, T.A., Hoogestraat, D.R., and Cookson, B.T. (2015). Application of Whole-Genome Sequencing for Bacterial Strain Typing in Molecular Epidemiology. J Clin Microbiol *53*, 1072–1079.

Salzberg, S.L., Phillippy, A.M., Zimin, A., Puiu, D., Magoc, T., Koren, S., Treangen, T.J., Schatz, M.C., Delcher, A.L., Roberts, M., et al. (2012). GAGE: A critical evaluation of genome assemblies and assembly algorithms. Genome Res *22*, 557–567.

San Francisco Business Times (2008). Applied Biosystems, Invitrogen complete $6.7 billion merger.

Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. PNAS *74*, 5463–5467.

Sanger, F., Coulson, A.R., Hong, G.F., Hill, D.F., and Petersen, G.B. (1982). Nucleotide sequence of bacteriophage λ DNA. Journal of Molecular Biology *162*, 729–773.

Seeliger, H.P.., and Höhne, K. (1979). Chapter II Serotyping ofListeria monocytogenesand Related Species.

Sovic, I., Krizanovic, K., Skala, K., and Sikic, M. (2015). Evaluation of hybrid and non-hybrid methods for de novo assembly of nanopore reads. bioRxiv 030437.

Staden, R. (1979). A strategy of DNA sequencing employing computer programs. Nucleic Acids Res *6*, 2601–2610.

Stamatakis, A. (2015). Using RAxML to Infer Phylogenies. Curr Protoc Bioinformatics *51*, 6.14.1–6.14.14.

Stein, R.A. (2008). Next-Generation Sequencing Update. GEN *28*.

Steinbock, L.J., and Radenovic, A. (2015). The emergence of nanopores in next-generation sequencing. Nanotechnology *26*, 074003.

Sutton, G.G., White, O., Adams, M.D., and Kerlavage, A.R. (1995). TIGR Assembler: A New Tool for Assembling Large Shotgun Sequencing Projects. Genome Science and Technology *1*, 9–19.

Tham, W., Ericsson, H., Loncarevic, S., Unnerstad, H., and Danielsson-Tham, M.-L. (2000). Lessons from an outbreak of listeriosis related to vacuum-packed gravad and cold-smoked fish. International Journal of Food Microbiology *62*, 173–175.

Thermo Fisher Website (2015a). Ion PGM^{TM} Sequencer Specifications.

Thermo Fisher Website (2015b). SOLiD® Next-Generation Sequencing Systems & Accessories.

Todd, E.C.D., and Notermans, S. (2011). Surveillance of listeriosis and its causative pathogen, Listeria monocytogenes. Food Control *22*, 1484–1490.

Trapnell, C., and Salzberg, S.L. (2009). How to map billions of short reads onto genomes. Nat Biotech *27*, 455–457.

Treangen, T.J., Abraham, A.-L., Touchon, M., and Rocha, E.P.C. (2009). Genesis, effects and fates of repeats in prokaryotic genomes. FEMS Microbiology Reviews *33*, 539–571.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001). The sequence of the human genome. Science *291*, 1304–1351.

Walsh, P.S., Metzger, D.A., and Higuchi, R. (1991). Chelex 100 as a medium for simple extraction of DNA for PCR-based typing from forensic material. BioTechniques *10*, 506–513.

Ward, T.J., Ducey, T.F., Usgaard, T., Dunn, K.A., and Bielawski, J.P. (2008). Multilocus Genotyping Assays for Single Nucleotide Polymorphism-Based Subtyping of Listeria monocytogenes Isolates. Appl. Environ. Microbiol. *74*, 7629–7642.

Wetterstrand, K. (2015). DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP).

Zerbino, D.R., and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. *18*, 821–829.

Zhang, W., Jayarao, B.M., and Knabel, S.J. (2004). Multi-virulence-locus sequence typing of Listeria monocytogenes. Appl. Environ. Microbiol. *70*, 913–920.

Zimin, A.V., Marcais, G., Puiu, D., Roberts, M., Salzberg, S.L., and Yorke, J.A. (2013). The MaSuRCA genome assembler. Bioinformatics *29*, 2669–2677.

8. List of figures

9. Appendix

9.1. Listing of all strains sequenced

Sample ID	Country of Isolation	Origin	Collection Year	serogroup by PCR
SLCC0174	Canada	human	1953	IVb
SLCC0175	Canada	human	1953	IIa
SLCC0178	Unknown	human	1954	IVb
SLCC0179	Germany	human	1954	IIa
SLCC0180	Austria	human	1954	IIa
SLCC0181	Germany	human	1954	IIa
SLCC0182	Japan	animal	1948	L. spp
SLCC0183	Canada	human	1954	IIa
SLCC0185	Germany	human	1954	IIa
SLCC0186	Germany	human	1954	IVb
SLCC0187	Germany	human	1954	IIa
SLCC0189	Germany	human	1954	IIa
SLCC0190	Germany	human	1954	IIa
SLCC0191	Germany	human	1954	IIa
SLCC0192	Germany	human	1954	IIc
SLCC0193	Germany	human	1954	IIa
SLCC0195	Germany	human	1954	IIa
SLCC0196	Germany	human	1954	IIb
SLCC0197	Germany	human	1954	IIb
SLCC0200	Germany	human	1954	IIa
SLCC0204	France	human	1950	IVb
SLCC0207	France	animal	1954	IVb
SLCC0208	France	human	1921	IVb
SLCC0209	France	animal	1952	IIa
SLCC0212	Germany	human	1954	IIa
SLCC0213	France	human	1953	IVb
SLCC0222	Germany	human	1955	IIc
SLCC0247	Germany	human	1953	IIa
SLCC0268	USA	human	1951	L. spp
SLCC0273	USA	human	1951	IIa
SLCC0291	USA	animal	1951	IVb
SLCC0307	Germany	human	1953	IIa
SLCC0308	Germany	human	1953	IIa
SLCC0309	Germany	human	1953	IIa
SLCC0518	Germany	human	1957	IIa
SLCC0519	USA	human	1956	IVb
SLCC0520	USA	human	1956	IVb
SLCC0522	USA	human	1956	IVb
SLCC0524	USA	human	1956	IIa
SLCC0526	USA	human	1956	IVb

SLCC0531	Germany	animal	1957	IVb
SLCC0532	Germany	animal	1957	IVb
SLCC0534	Germany	animal	1956	IIa
SLCC0535	Germany	animal	1956	IIa
SLCC0536	Canada	human	1957	IVb
SLCC0538	Canada	human	1957	IIa
SLCC0542	France	human	1957	IIa
SLCC0546	France	human	1957	IIb
SLCC0549	USA	unknown	1957	IVb
SLCC0563	France	human	1957	IIa
SLCC0582	Netherlands	animal	1957	IIc
SLCC0589	Germany	human	1957	IVb
SLCC0590	Germany	human	1957	IIa
SLCC0597	USA	human	1957	L. spp
SLCC0615	Germany	animal	1958	IVb
SLCC0616	Germany	animal	1958	IIa
SLCC0617	Germany	human	1958	IIa
SLCC0623	Germany	human	1958	IIa
SLCC0624	unknown	unknown	Unknown	IIa
SLCC0627	Sweden	animal	1958	IIa
SLCC0628	Sweden	animal	1958	IIa
SLCC0629	Sweden	animal	1958	IIa
SLCC0631	Sweden	animal	1957	IIa
SLCC0632	Sweden	animal	1957	IIa
SLCC0634	Germany	human	1958	IIa
SLCC0635	Germany	human	1958	IIa
SLCC0636	USA	human	1956	IVb
SLCC0638	USA	human	1958	IVb
SLCC0640	USA	human	1958	IIb
SLCC0645	Germany	human	1958	IIa
SLCC0646	Germany	animal	1958	IVb
SLCC0647	Germany	human	1958	IVb
SLCC0661	Denmark	human	1958	IIb
SLCC0664	Germany	human	1958	IIa
SLCC0665	Canada	human	1958	IVb
SLCC0666	Canada	human	1958	IVb
SLCC0667	Canada	human	1958	IVb
SLCC0668	Canada	human	1958	IVb
SLCC0669	Canada	human	1958	IVb
SLCC0671	Canada	human	1958	IVb
SLCC0674	Canada	human	1958	IVb
SLCC0675	Canada	human	1958	IVb
SLCC0676	Canada	human	1958	IVb
SLCC0682	USA	human	1958	IVb
SLCC0683	Germany	human	1958	IIa
SLCC0684	Germany	animal	1957	IVb

SLCC0685	Germany	animal	1957	IIa
SLCC0687	Germany	animal	1958	IIa
SLCC0688	Germany	human	1958	IIa
SLCC0689	Germany	animal	1958	IVb
SLCC0693	Germany	human	1959	IVb
SLCC0694	Germany	human	1959	IIc
SLCC0695	France	animal	1957	IIa
SLCC0708	France	animal	1956	IIa
SLCC0709	France	animal	1958	IVb
SLCC0710	France	animal	1958	IIa
SLCC0757	Denmark	animal	1951	IIb
SLCC0805	Germany	human	1959	IVb
SLCC0853	Germany	human	1959	IIa
SLCC0887	unknown	unknown	1938	L. spp
SLCC0993	Canada	animal	1953	IIa
SLCC0995	Canada	animal	1953	IIa
SLCC1003	Canada	animal	1952	IVb
SLCC1008	Canada	animal	1951	IIa
SLCC1069	Germany	human	1960	IVb
SLCC1070	Germany	human	1960	IVb
SLCC1082	Germany	human	1960	IIc
SLCC1092	Germany	human	1960	IVb
SLCC1144	Germany	human	1960	IVb
SLCC1231	Germany	human	1961	IIa
SLCC1260	Germany	human	1961	IVb
SLCC1272	Germany	human	1961	IVb
SLCC1306	Germany	human	1961	IVb
SLCC1321	Germany	human	1961	IVb
SLCC1324	Germany	human	1961	IVb
SLCC1332	Germany	human	1961	IVb
SLCC1439	Germany	human	1962	IIa
SLCC1440	Germany	human	1962	IVb
SLCC1441	Germany	human	1962	IIa
SLCC1444	Germany	human	1962	IVb
SLCC1446	Germany	human	1962	IVb
SLCC1451	Germany	human	1962	IVb
SLCC1456	Germany	human	1962	IVb
SLCC1476	Germany	human	1962	IIa
SLCC1480	Germany	human	1962	IIb
SLCC1494	Germany	human	1962	IVb
SLCC1499	Germany	human	1962	IVb
SLCC1564	Germany	human	1963	IVb
SLCC1572	Germany	human	1962	IVb
SLCC1573	Germany	human	1962	IIa
SLCC1574	Germany	human	1962	IVb
SLCC1575	Germany	human	1962	IVb

SLCC1577	Germany	human	1963	IIa
SLCC1588	Germany	human	1963	IVb
SLCC1589	Germany	human	1963	IIa
SLCC1598	Germany	human	1963	IVb
SLCC1607	Germany	human	1963	IVb
SLCC1610	Germany	human	1963	IVb
SLCC1615	Germany	human	1963	IVb
SLCC1617	Germany	human	1963	IVb
SLCC1623	Germany	human	1963	IVb
SLCC1624	Germany	human	1963	IVb
SLCC1627	Germany	human	1963	IVb
SLCC1631	Germany	human	1963	IVb
SLCC1633	Germany	human	1963	IIa
SLCC1639	Germany	human	1963	IIa
SLCC1640	Germany	human	1963	IIa
SLCC1645	Germany	human	1963	IVb
SLCC1663	United Kingdom	animal	1937	IVb
SLCC1669	Germany	human	1963	IVb
SLCC1672	Germany	human	1963	IIc
SLCC1673	Germany	human	1963	IIc
SLCC1675	Germany	human	1963	IIa
SLCC1676	Germany	human	1963	IVb
SLCC1682	Germany	human	1963	IIa
SLCC1747	Germany	human	1964	IVb
SLCC1756	Germany	human	1964	IVb
SLCC1784	Germany	human	1964	IIa
SLCC1876	Germany	human	1964	IVb
SLCC1955	Germany	human	1964	IVb
SLCC2075	Germany	human	1964	IIa
SLCC2129	Germany	human	1965	IVb
SLCC2130	Germany	human	1965	IIa
SLCC2181	Germany	human	1965	IIb
SLCC2209	Germany	human	1965	IVb
SLCC2261	Germany	human	1965	IIa
SLCC2295	Germany	human	1965	IVb
SLCC2432	Germany	human	1966	IIa
SLCC2470	Germany	human	1966	IIb
SLCC2478	Germany	human	1966	IVb
SLCC2499	Germany	human	1966	IIb
SLCC2529	Germany	human	1966	IIb
SLCC2545	Germany	human	1966	IIb
SLCC2588	Germany	human	1967	IIb
SLCC2645	Germany	human	1967	IIb
SLCC2653	Germany	human	1967	IIa
SLCC2655	Germany	human	1967	IIc
SLCC2781	Germany	human	1968	IIa

SLCC2815	Germany	human	1968	IIb
SLCC2818	Germany	human	1968	IIb
SLCC2830	Germany	human	1968	IIa
SLCC2831	Germany	human	1968	IVb
SLCC2859	Germany	human	1968	IVb
SLCC2896	Germany	human	1969	IVb
SLCC2958	Germany	human	1969	IVb
SLCC4826	United Kingdom	animal	1924	IIa
SLCC5211	Germany	human	1978	IVb
SLCC5299	Germany	human	1978	IVb
SLCC5405	Germany	human	1980	IVb
SLCC5406	Germany	human	1980	IVb
SLCC5489	Germany	human	1981	IVb

Table 9: Summary of all Austrian clinical isolates sequenced. % Good cgMLST Targets refers to 1701 core genome targets in total.

Sample ID	Country of Isolation	Origin	Collection Year	% Good CGMLST Targets	Serogroup
L01/12	Austria	Human	2012	99.4	
L02/12	Austria	Human	2012	99.5	
L04/12	Austria	Human	2012	99.5	
L05/12	Austria	Human	2012	99.9	
L06/12	Austria	Human	2012	99.5	
L07/12	Austria	Human	2012	99.3	
L08/12	Austria	Human	2012	99.5	
L09/12	Austria	Human	2012	99.5	
L10/12	Austria	Human	2012	99.7	
L11/12	Austria	Human	2012	99.4	
L12/12	Austria	Human	2012	99.5	
L14/12	Austria	Human	2012	99.4	
L15/12	Austria	Human	2012	99.5	IIb
L16/12	Austria	Human	2012	99.5	IVb
L17/12	Austria	Human	2012	99.5	IVb
L22/12	Austria	Human	2012	99.9	IIa
L23/12	Austria	Human	2012	99.5	IVb
L24/12	Austria	Human	2012	99.8	IIa
L25/12	Austria	Human	2012	100	IIa
L26/12	Austria	Human	2012	99.6	IVb
L27/12	Austria	Human	2012	99.5	IVb
L28/12	Austria	Human	2012	99.9	IIa
L29/12	Austria	Human	2012	99.3	IVb
L30/12	Austria	Human	2012	99.6	IVb
L31/12	Austria	Human	2012	99.5	IVb
L32/12	Austria	Human	2012	99.5	IVb

L33/12	Austria	Human	2012	99.5	IVb
L34/12	Austria	Human	2012	99.8	IIa
L35/12	Austria	Human	2012	99.9	IIc
L36/12	Austria	Human	2012	99.4	IVb
L37/12	Austria	Human	2012	99.2	IVb
L38/12	Austria	Human	2012	99.4	IVb
L40/12	Austria	Human	2012	99.5	IVb
L41/12	Austria	Human	2012	99.9	IIa
L01/13	Austria	Human	2013	99.4	IVb
L02/13	Austria	Human	2013	99.9	IIa
L03/13	Austria	Human	2013	99.2	
L06/13	Austria	Human	2013	99.6	IIa
L07/13	Austria	Human	2013	99.9	IIa
L08/13	Austria	Human	2013	99.6	IIa
L09/13	Austria	Human	2013	99.7	IIa
L12/13	Austria	Human	2013	99.5	IVb
L13/13	Austria	Human	2013	99.9	IIa
L14/13	Austria	Human	2013	99.8	IIa
L15/13	Austria	Human	2013	99.6	IIa
L16/13	Austria	Human	2013	99.6	IVb
L18/13	Austria	Human	2013	99.7	IIa
L20/13	Austria	Human	2013	99.6	IIa
L22/13	Austria	Human	2013	99.9	IIa
L23/13	Austria	Human	2013	99.6	IIa
L24/13	Austria	Human	2013	99.3	IVb
L25/13	Austria	Human	2013	99.5	IVb
L26/13	Austria	Human	2013	99.5	IIa
L27/13	Austria	Human	2013	99.8	IIa
L28/13	Austria	Human	2013	99.8	IIa
L29/13	Austria	Human	2013	99.8	IIa
L30/13	Austria	Human	2013	99.4	IIb
L31/13	Austria	Human	2013	99.4	IVb
MRL-13/00815	Austria	Human	2013	99.4	IVb
MRL-13/00816	Austria	Human	2013	99.4	IVb
L35/13	Austria	Human	2013	99.8	IIa
L37/13	Austria	Human	2013	99.6	
L38/13	Austria	Human	2013	99.8	
MRL-14/00018	Austria	Human	2014	100	
MRL-14/00102	Austria	Human	2014	99.9	
MRL-14/00271	Austria	Human	2014	99.9	
MRL-14/00406	Austria	Human	2014	99.6	
MRL-14/00459	Austria	Human	2014	99.6	
MRL-14/00460	Austria	Human	2014	99.9	
MRL-14/00615	Austria	Human	2014	99.4	
MRL-14/00616	Austria	Human	2014	99.5	
MRL-14/00617	Austria	Human	2014	98.9	

MRL-14/00618	Austria	Human	2014	99.1	
MRL-14/00636	Austria	Human	2014	97.5	
MRL-14/00658	Austria	Human	2014	97.9	
MRL-14/00682	Austria	Human	2014	99.2	
MRL-14/00983	Austria	Human	2014	100	
MRL-14/00747	Austria	Human	2014	99.2	
MRL-14/00748	Austria	Human	2014	99.7	
MRL-14/00761	Austria	Human	2014	99.5	
MRL-14/00762	Austria	Human	2014	99.5	
MRL-14/00817	Austria	Human	2014	99.1	
MRL-14/00819	Austria	Human	2014	99.8	
MRL-14/00850	Austria	Human	2014	99.4	
MRL-14/00912	Austria	Human	2014	99.2	
MRL-14/00954	Austria	Human	2014	99.4	
MRL-14/00997	Austria	Human	2014	99.1	
MRL-14/01054	Austria	Human	2014	98.9	
MRL-14/01154	Austria	Human	2014	99.5	
MRL-14/01208	Austria	Human	2014	99.2	
MRL-14/01209	Austria	Human	2014	99.6	
MRL-14/01210	Austria	Human	2014	100	
MRL-14/01314	Austria	Human	2014	99.8	
MRL-14/01315	Austria	Human	2014	99.6	
MRL-14/01358	Austria	Human	2014	99.9	
MRL-14/01359	Austria	Human	2014	100	
MRL-14/01360	Austria	Human	2014	99.9	
MRL-14/01361	Austria	Human	2014	99.4	
MRL-14/01399	Austria	Human	2014	99.5	
MRL-15/00014	Austria	Human	2014	99.5	IVb
MRL-15/00015	Austria	Human	2014	99.4	IVb
MRL-15/00016	Austria	Human	2015	99.6	IIb
MRL-15/00032	Austria	Human	2015	100	IIa
MRL-15/00033	Austria	Human	2015	99.6	IIa
MRL-15/00034	Austria	Human	2015	99.6	
MRL-15/00035	Austria	Human	2015	99.9	IIa
MRL-15/00063	Austria	Human	2015	99.6	IIa
MRL-15/00085	Austria	Human	2015	97.9	IVb
MRL-15/00093	Austria	Human	2015	99.4	IVb
MRL-15/00126	Austria	Human	2015	97.9	
MRL-15/00127	Austria	Human	2015	99.9	IIa
MRL-15/00128	Austria	Human	2015	99.6	IIa
MRL-15/00145	Austria	Human	2015	99.9	IIa
MRL-15/00294	Austria	Human	2015	99.5	IVb
MRL-15/00295	Austria	Human	2015	99.9	IIa
MRL-15/00398	Austria	Human	2015	99.6	IVb
MRL-15/00441	Austria	Human	2015	99.4	IVb

9.2. Paired end read data downloaded from the ENA

Accession Nr:	Isolate ID
ERR538090	12025641
ERR538091	12025647
ERR538092	16132
ERR538093	2010-00770
ERR538094	3230TP3
ERR538095	3230TP5
ERR538096	4548TP4
ERR538098	ATCC15313
ERR538099	CIP104794
ERR538101	CIP105448
ERR538102	CIP105449
ERR538103	CIP105457
ERR538104	CIP105458
ERR538105	CIP105459
ERR538106	CIP59-53
ERR538107	CIP78-34
ERR538108	CIP78-35
ERR538109	CIP78-36
ERR538110	CIP78-39
ERR538112	CIP78-43
ERR538114	K70-10
ERR538115	L10-10
ERR538116	L14-10
ERR538117	L16-10
ERR538118	L17-10
ERR538119	L18-10
ERR538120	L19-10
ERR538122	L20-09
ERR538123	L20-10
ERR538124	L21-09
ERR538125	L23-09
ERR538126	L27-09
ERR538127	L29-09
ERR538128	L30-10
ERR538129	L31-09
ERR538130	L32-09
ERR538132	L32-10
ERR538133	L33-09
ERR538134	L33-10
ERR538135	L34-09
ERR538136	L35-09
ERR538137	L38-11
ERR538138	L4-10

ERR538139	L42-10
ERR538140	L68-09
ERR538141	L71-09
ERR538142	L75-09
ERR538143	L9-10
ERR538145	LD12-10
ERR538146	LD27-12
ERR538148	MRL-13-00230
ERR538150	Ro-015
ERR664374	L2708
ERR664375	L3308
ERR664376	L3808
ERR664377	L3908
ERR664378	L4008
ERR664379	L4508
ERR664380	L6808
ERR664381	L7508
ERR664382	W9508
ERR664383	W9608
ERR664384	W9708
ERR664394	L6708
ERR664778	SLCC0717
ERR664779	SLCC0759
ERR664780	SLCC1042
ERR664781	SLCC3280
ERR664782	SLCC3287
ERR664783	SLCC3961
ERR664784	SLCC4163
ERR664785	SLCC6263
ERR664786	SLCC4771