

APPROVED: 13 December 2016

doi:10.2903/sp.efsa.2017.EN-1151

Closing gaps for performing a risk assessment on *Listeria monocytogenes* in ready-to-eat (RTE) foods: activity 3, the comparison of isolates from different compartments along the food chain, and from humans using whole genome sequencing (WGS) analysis

Eva Møller Nielsen¹, Jonas T. Björkman¹, Kristoffer Kiil¹, Kathie Grant², Tim Dallman², Anaïs Painset², Corinne Amar², Sophie Roussel³, Laurent Guillier³, Benjamin Félix³, Ovidiu Rotariu⁴, Francisco Perez-Reche⁴, Ken Forbes⁴, Norval Strachan⁴

¹Statens Serum Institut, Copenhagen, Denmark; ²Public Health England, Colindale, UK; ³Anses, Maisons-Alfort, France; ⁴University of Aberdeen, UK

Abstract

This report presents the results of the project “Closing gaps for performing a risk assessment on *Listeria monocytogenes* in ready-to-eat (RTE) foods: activity 3, the comparison of isolates from different compartments along the food chain, and from humans using whole genome sequencing (WGS) analysis”. The main objective was to compare *L. monocytogenes* isolates collected in the EU from ready-to-eat (RTE) foods, compartments along the food chain and from human cases by the use of WGS. A total of 1,143 *L. monocytogenes* isolates were selected for the study, including 333 human clinical isolates and 810 isolates from the food chain. The isolates were whole genome sequenced. The phylogeny showed a clear delineation between *L. monocytogenes* lineages and between clonal complexes within lineages. A range of typing methods were applied to the sequence data, providing the framework to answer questions on genetic diversity and epidemiological relationships. Retrospective analysis of nine outbreaks showed that WGS is a powerful tool in national and international outbreak investigations as WGS can accurately rule isolates in or out of outbreaks. Source attribution models showed bovine reservoir to be the main source of human disease although other sources also contributed and generally confidence intervals were high. Numerous consistent genetic linkages between *a priori* unlinked strains were identified, some of which involved isolates from multiple countries. The presence of putative markers conferring the potential to survive/multiply in the food chain and/or cause disease in humans was explored by detecting the presence of putative virulence genes, AMR genes and factors conferring the ability to persist in the food processing chain. This study has demonstrated one of the major benefits of WGS, which is the ability to address a wide range of questions including those on virulence, antimicrobial resistance, source attribution, surveillance and outbreak detection and investigation, in a single experiment.

© European Food Safety Authority, 2017

Key words: *Listeria monocytogenes*, Whole Genome Sequencing, Genetic Diversity, Phylogeny, Food, Human

Question number: EFSA-Q-2014-00026

Correspondence: biocontam@efsa.europa.eu

Disclaimer: The present document has been produced and adopted by the bodies identified above as author(s). This task has been carried out exclusively by the author(s) in the context of a contract between the European Food Safety Authority and the author(s), awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the authors.

Acknowledgements:

We would like to thank all the persons and institutes that have provided the project with isolates and accompanying information. Without them, this project would not have been possible.

Lin Cathrine T. Brandal, Norwegian Institute of Public Health, Norway
Julio Vázquez Moreno and Raquel Abad Torreblanca, Instituto de Salud Carlos III, Spain
Marc Lecuit, Institut Pasteur, France
Alexandre Leclercq, Institut Pasteur, France
Iva Hristova, National Center of Infectious and Parasitic Diseases, Bulgaria
Marija Trkov, National Laboratory of Health, Environment and Food, Slovenia
Cecilia Jernberg, Public Health Agency of Sweden, Sweden
Ariane Pietzka, Austrian Agency for Health and Food Safety, Austria
Eelco Franz and Ingrid Friesema, RIVM, The Netherlands
Carlo Spanu, University of Sassari Sardinia
Ifip, French Institute for Pig and Pork Industry, Maisons-Alfort, France
All the NRLs for providing the isolates from the EU baseline study

Special thanks to Sylvain Brisse and Alexandra Moura, Institut Pasteur, France, for providing cgMLST data.

The authors would also like to thank the EFSA staff members: Maria Teresa da Silva Felicio, Beatriz Guerra, Ernesto Liebana and Valentina Rizzi as well as the members of the Working Group on *Listeria monocytogenes* contamination of ready-to-eat foods: Kostas Koutsoumanis, Roland Lindqvist, Moez Sanaa, Panagiotis Skandamis, Niko Speybroek, Johanna Takkinen and Martin Wagner for the support, revisions and suggestions during the development of the present procurement activity and report.

Suggested citation: Eva Møller Nielsen, Jonas T. Björkman, Kristoffer Kiil, Kathie Grant, Tim Dallman, Anaïs Painset, Corinne Amar, Sophie Roussel, Laurent Guillier, Benjamin Félix, Ovidiu Rotariu, Francisco Perez-Reche, Ken Forbes and Norval Strachan, 2017. Closing gaps for performing a risk assessment on *Listeria monocytogenes* in ready-to-eat (RTE) foods: activity 3, the comparison of isolates from different compartments along the food chain, and from humans using whole genome sequencing (WGS) analysis. EFSA supporting publication 2017:EN-1151. 170 pp. doi:10.2903/sp.efsa.2017.EN-1151

ISSN: 2397-8325

© European Food Safety Authority, 2017

Reproduction is authorised provided the source is acknowledged.

Summary

This report presents the results of the project "Closing gaps for performing a risk assessment on *Listeria monocytogenes* in ready-to-eat (RTE) foods: activity 3, the comparison of isolates from different compartments along the food chain, and from humans using whole genome sequencing (WGS) analysis". The project acronym, LISEQ, will be used in this report.

The main objective of the study was to compare *L. monocytogenes* isolates collected in the EU from ready-to-eat (RTE) foods, compartments along the food chain and from human cases by the use of whole genome sequencing (WGS).

A total of 1,143 *L. monocytogenes* isolates were selected for the study and these included 333 human clinical isolates and 810 isolates from the food chain. The food chain isolates were acquired as part of the EU-wide Baseline survey (BLS) on ready-to-eat food conducted in 2010-2011 (353 isolates), obtained as part of national surveys, control programmes or research projects (423 isolates) or in connection to outbreak investigations (34 isolates). The human clinical isolates were supplied by national public health laboratories and represented sporadic cases (262 isolates) and outbreak-related isolates (71 isolates) from eleven European countries, mainly in the years 2010-2011.

A database was constructed with the available metadata for the isolates with links to the genome sequences. The isolates were whole genome sequenced (WGS) at Public Health England's sequencing facility. To achieve the goals of the specific objectives we applied a range of microbial typing methods to the sequence data providing the framework to answer questions on genetic diversity and epidemiological relationships. Three allelic based typing methodologies were used, multi-locus sequence typing (MLST), core genome MLST (cgMLST) and ribosomal MLST (rMLST) as well as Single Nucleotide Polymorphisms (SNPs). SNPs inference and analysis was performed using software developed at PHE with each *L. monocytogenes* clonal complex (CC) having a separate database instance. Short read sequences from strains selected in the strains of the study were mapped against an appropriate reference genome. The resultant sequence alignment maps were processed and high quality variants extracted. From 1,143 isolates sequenced, 42 different CC and 13 singleton sequence types (STs) were identified. One isolate had a novel ST and could not be assigned to any CC. Ten clonal complexes accounted for 70% of the isolates.

It was a specific objective to perform a retrospective analysis of outbreak strains to investigate the suitability of WGS as a tool in outbreak investigations. Isolates from nine food-borne outbreaks, representing a range of different characteristics, were selected. The relationship between the human clinical isolates, isolates available from the suspected sources as well as similar background isolates were characterised by SNP and cgMLST analyses. Most of the outbreaks were tightly clustered; five out of nine had a maximum pairwise distance of <5 SNP, four outbreaks had a maximum SNP difference between 8 and 21. The cgMLST showed for the most part concordant results with the SNP analysis. In 8 outbreaks the median and maximum sizes of branches within a whole CC were shorter compared to SNP branches.

To explore the genetic diversity of *L. monocytogenes* within and between the different sources and of human origin, different indices and measures were applied to characterise and describe the variation of isolates in the collection. Simpson's index for humans and the five sources exhibited high diversity (>0.8) for both MLST and rMLST. Simpson's index of diversity between each of the sources was indistinguishable. Rarefaction curves showed for both MLST and rMLST that all of the genotypes had not been sampled. Bovine and human isolates had the highest number of new STs per isolate. Nei's genetic distance showed that there were significant differences between human and all sources at all levels of molecular analysis explored, but the distance between humans and bovine was the smallest.

Two approaches were used to assess the epidemiological relationship of *L. monocytogenes* from the different sources and of human origin considering the genomic information and the metadata available for each isolate. The first approach was using the method of source attribution, i.e., partitioning of the human disease burden of listeriosis to specific sources. Because of the relatively small number of isolates, all of the isolates along the food chain that originate from a particular

reservoir were combined. Human clinical cases were attributed to these sources by utilising five different mathematical models and the genomic typing data that was generated. Source attribution was applied utilising 5 models (Hald, Dutch, STRUCTURE, Asymmetric Island and Aberdeen) for 5 sources [fish, swine, ovine, bovine and poultry] and four sources [removing poultry]). All of the models showed bovine as the main source of human disease (32-64% for 5 sources and 33-61% for 4 sources) but for a number of the models there were broad confidence intervals.

The second approach to assess the possible epidemiological relationship between strains was to identify clusters of clinical and food isolates based on SNP differences. The WGS data was analysed along with the epidemiological information of the food and clinical isolate to assess, retrospectively, relationships between circulating strains of *L. monocytogenes* in EU within 2010-2012 period. The retrospective analysis showed that numerous consistent genetic linkages, between a priori unlinked strains, can be established with WGS. By the use of SNP pairwise distances, 124 "clusters" of isolates were identified and 27 of these included both human and food isolates, potentially relating sporadic human cases to contemporary food isolates that circulate in EU. All three categories of RTE food products were involved, but most of the clusters were related to smoked fish.

Another specific objective of the study was to identify the presence of putative markers conferring the potential to survive/multiply in the food chain and/or cause disease in humans (e.g. virulence and antimicrobial resistance). We analysed the WGS data for the presence of 115 putative markers of virulence. More than 80% of markers were present in greater than 95% of the isolates suggesting that most putative markers described in the literature are ubiquitous across *L. monocytogenes* lineages I and II. The majority of markers not present in all isolates were over-represented in food and/or lineage II isolates with markers associated with stress survival or cell wall modification being particularly enriched. Conversely, the recently discovered *Listeria* pathogenicity island 3 and the surface protein VIP were more likely to be found in clinical and/or lineage I isolates. Similar to the virulence genes, the presence of genes related to resistance to antimicrobials and detergents were searched in the LISEQ collection. There was found remarkable low resistance to tetracycline (<0.1%) and penicillin (1%). Resistance to detergents and antiseptics via efflux activity was significant with mechanisms detected at a prevalence approaching 20%. Some studies have suggested that *L. monocytogenes* strains that are able to persist in the food production environment are genetically distinct from "transient strains" that do not have this capability. The presence or absence of genes thought to promote persistence was not found to be pertinent for predicting persistent phenotype. However, it was shown that WGS SNP-based analysis is well suited and valuable for investigating persistence and contamination routes within food processing facilities and within the food chain.

In conclusion, this study carried out WGS of a large unique collection of *L. monocytogenes* isolates from foods, food processing environments and clinical cases from a large number of European countries. This study has demonstrated one of the major benefits of WGS, which is the ability to address a wide range of questions including those on virulence, antimicrobial resistance, source attribution, surveillance and outbreak detection and investigation, in a single experiment. This study illustrates one of the major strengths of WGS in comparison to conventional molecular typing methods, which is its ability to provide high quality, unambiguous data. WGS analysis such as cgMLST and cgSNP-based typing approaches have been shown to have unparalleled strain typing resolution and it has been demonstrated here how WGS is able to link previously undetected cases to outbreaks and detect clusters of cases that were previously undetected. It has also been shown, however, that knowledge of the accessory genome can contribute to the interpretation of strain relatedness. The limitations of WGS have less to do with the actual sequencing and the analyses themselves but more dependent on representative sampling of isolates and requirement for good epidemiological data to further investigate genetically linked by WGS. This study supports the use of WGS for *L. monocytogenes* outbreak investigations although experience from more complex outbreaks would be valuable. However, is difficult to recreate outbreak investigations accurately retrospectively and in order to maximise the advantages of using WGS for outbreak detection it would be highly valuable to use WGS prospectively for the surveillance of listeriosis across Europe.

Table of Contents

Abstract.....	1
Acknowledgements:	2
Summary	3
1. Introduction.....	8
1.1. Background and Terms of Reference as provided by the requestor	10
2. Isolate collection	12
2.1. Selection of strains to fulfil the main and specific objectives of the study.....	12
2.2. Strains from the baseline survey (BLS)	14
2.2.1. Selection criteria BLS 1. Availability of the strains.....	14
2.2.2. Selection criterion BLS 2: Isolates from same origin (food and Member State)	14
2.2.3. Selection criterion BLS 3. Multiple isolates from a sample (all categories)	15
2.3. Strains from other foods (OF)	15
2.3.1. Selection criterion OF 1. Food origin	15
2.3.2. Selection criterion OF 2. Temporal criterion	15
2.3.3. Selection criterion OF 3. Further selection.....	15
2.3.4. Fruit and vegetables.....	15
2.4. Strains from food chain production stages	16
2.5. Strains from sporadic human clinical cases	16
2.5.1. Selection criterion C1. Availability of the strains.	16
2.5.2. Selection criterion C2. Temporal and geographical criterion	16
2.5.3. Selection criterion C3. Incidence	16
2.5.4. Selection criterion C4. Subtype information.....	17
2.6. Strains from outbreaks	17
2.7. Strain selection summary.....	17
2.8. Database.....	21
2.8.1. Database structure.....	21
2.8.2. Core information on strains.....	22
2.8.3. Data export	22
3. Methodologies	23
4. Sequencing and Phylogentic Analysis	25
4.1. Methods	25
4.1.1. DNA extraction	25
4.1.2. DNA sequencing and validation.....	25
4.1.3. Assembly and annotation.....	26
4.1.4. Gene by gene based typing	26
4.1.5. Phylogenetic analysis	26
4.1.6. Genetic distance: SNP address.....	28
4.1.7. Genetic distance: SNP address.....	28
4.2. Results	29
4.2.1. Processing of isolates	29
4.2.2. Gene by gene based typing: MLST and clonal complex assignment	30
4.2.3. Developing the Framework for Phylogenetic Analysis.....	32
4.2.4. Phylogenetic analysis of major clonal complexes	34
4.3. Conclusion.....	38
5. Retrospective analysis of outbreaks.....	39
5.1. Methods	39
5.2. Results	40
5.2.1. Outbreak 1 – CC155.....	40
5.2.2. Outbreak 2 – CC1	42
5.2.3. Outbreak 3 – CC7	44
5.2.4. Outbreak 4 – CC59.....	45

5.2.5. Outbreak 5 – CC415	47
5.2.6. Outbreak 6 – CC398	48
5.2.7. Outbreak 7 – CC87	49
5.2.8. Outbreak 8 – ST14	52
5.2.9. Outbreak 9 – CC4	54
5.3. Conclusions	56
6. Genetic diversity	57
6.1. Methods	57
6.1.1. Simpson's Diversity Index	57
6.1.2. Rarefaction	58
6.1.3. Nei's genetic distance	58
6.1.4. Graphical visualisation and cluster analysis	58
6.1.5. Analyses	58
6.1.6. Selection of Genomes for analysis suitable for genetic diversity and source attribution analysis	59
6.2. Results and Discussion	60
6.2.1. 7 locus MLST	60
6.2.2. 30 locus rMLST	62
6.2.3. 1,748 locus cgMLST	65
6.2.4. 39,529 cgSNP	66
6.2.5. Graphical Visualisation	66
6.3. Conclusions	69
7. Epidemiological relationship: Source attribution	69
7.1. Methods	70
7.1.1. Source Attribution Methods	70
7.1.2. Self-Attribution	74
7.1.3. Analyses	74
7.2. Results and Discussion	74
7.2.1. Source Attribution of 5 sources	74
7.2.2. Source Attribution of 4 Sources (Excluding Poultry)	81
7.2.3. Discussion	86
7.3. Conclusions	86
8. Epidemiological relationship – linking of genetically related isolates	87
8.1. Methods	87
8.1.1. Definition of genetically clustered strains	87
8.1.2. R packages and software	88
8.2. Results	89
8.2.1. Epidemiological analysis of genetically clustered strains: link between human sporadic strains and potential relation with food strains	91
8.2.2. Geographical and temporal widespread of genetically clustered strains	94
8.2.3. Consistency of clusters established	95
8.3. Conclusion	96
9. Putative markers	96
9.1. Methods	97
9.1.1. Antibiotic resistance genes	97
9.1.2. Published virulence factors	97
9.1.3. Genes implicated in persistence	97
9.1.4. Markers of host association	98
9.2. Results	98
9.2.1. Antimicrobial resistance	98
9.2.2. Published virulence factors	99
9.2.3. Genes implicated in persistence	101
9.2.4. Markers of host association	109
9.3. Conclusion	109
10. Conclusions	110

11. Additional supporting information.....	115
References.....	116
Glossary	124
LISEQ database glossary.....	124
List of abbreviations used in the report.....	125
Appendix 1: Isolates from the EU-wide baseline survey on prevalence of <i>L. monocytogenes</i> in certain RTE foods conducted in 2010-2012	126
Appendix 2: Isolates other food, ready-to-eat meat and cheese	139
Appendix 3: Isolates from other food, fruits and vegetables.....	147
Appendix 4: Isolates from the food production chain	148
Appendix 5: Isolates from sporadic clinical cases.....	156
Appendix 6: Isolates from outbreaks	163
Appendix 7: Rarefaction and Simpson's diversity index of 7 locus MLST clinical data stratified by age.....	166
Appendix 8: Links to Attribution model software.....	168

1. Introduction

Listeria monocytogenes causes a range of clinical illnesses from mild diarrhoea to severe invasive infection (listeriosis) including bacteraemia, meningitis, encephalitis, abortion and stillbirth. In EU in 2014, a total number of 2,161 confirmed human cases were reported by 27 MS, corresponding to an EU notification rate of 0.52 cases per 100,000 population. The highest notification rates were observed in Denmark, Sweden, Finland and Spain (1.64, 1.30, 1.19 and 1.15 cases per 100,000 population respectively) (EFSA and ECDC, 2015). Listeriosis generally affects individuals who have a weakened immune system including the elderly, those who are immunosuppressed due to existing medical conditions or their treatment, pregnant women and neonates. Whilst listeriosis is a relatively rare disease, it has a high fatality rate of 20-30% and the burden of disease is high. The majority of cases appear to be sporadic, although outbreaks are not uncommon. Listeriosis is almost exclusively transmitted by contaminated food with ready-to-eat meat and fish products and soft and semi-soft cheeses often identified as sources of infection. Due to its ability to survive under conditions of stress, *L. monocytogenes* has the capacity to persist in food processing environments, sometimes for years and often this is the route by which ready-to-eat food becomes contaminated. Identifying the food vehicle and tracing the origin of contamination are paramount in developing and implementing effective control and preventative measures and the typing of *L. monocytogenes* isolates continues to play a crucial role in such investigations.

A number of phenotypic and genotypic methods are used for typing *L. monocytogenes*. Traditionally, serotyping, based on the agglutination of somatic (O) and flagellar (H) antigens, classifying *L. monocytogenes* into at least 13 serotypes, has been the first level of subtype discrimination. More recently a multiplex PCR scheme based on the amplification of four specific marker genes (*Imo0737*; *ORF2110*; *Imo1118* and *ORF2819*), has been used to distinguish *L. monocytogenes* into four molecular groups that correlate well with known *L. monocytogenes* serotypes (1/2a-3a; 1/2b-3b-7; 1/2c-3c and 4b-3b-7) (Doumith et al., 2004). Since at least 95% of isolates from food and clinical cases are of serotypes 1/2a, 1/2b, 1/2c and 4b this robust, reproducible PCR-based method has been adopted by many reference laboratories as a rapid alternative that overcomes the many drawbacks of serotyping. However, because both traditional serotyping and serogrouping by multiplex PCR separate *L. monocytogenes* into only four groups, they lack discriminatory power which has led to the development of a variety of molecular typing techniques for higher resolution subtyping for outbreak detection, for linking human cases to food sources and for tracking strains along the food chain. The most widely used method in reference laboratories is pulsed-field gel electrophoresis (PFGE), but other methods including multi-locus variable number of tandem-repeat analysis (MLVA) and fAFLP are also used (recently reviewed (Camargo et al., 2016)). PFGE using two restriction enzymes is the standard method in PulseNet laboratories in North America, PulseNet International, as well as in the European surveillance system for human infections (TESSy). PFGE is also the first method to be included in the on-going project - coordinated by ECDC and EFSA - that aims to include molecular typing data from humans and food/animal/environment into a joint molecular typing database (EFSA, 2014).

The molecular typing methods commonly in use such as PFGE suffer a number of practical limitations including labour intensiveness, reproducibility and inter laboratory comparability. Whilst such methods have been extremely valuable in assisting epidemiological investigations of listeriosis, they also lack the ability to inform evolutionary relationships which would provide valuable insight for source tracking and source attribution (Orsi et al., 2011). Such information is available through single nucleotide polymorphism (SNP)-based typing approaches including multilocus sequencing typing (MLST) and whole genome sequencing (WGS) analysis. MLST is a well-established sequencing-based method whereby the unique variation in specific fragments, of a set of housekeeping genes (usually 7) are assigned an allele number and the alleles at each loci provide an allelic profile or sequence

type. This technique has been used to study and describe the population structure and phylogeny of many bacterial pathogens and has shown that *L. monocytogenes* forms a structured population consisting of four divergent lineages (I-IV) (Ragon et al., 2008). Each lineage comprises specific serotypes, with Lineage I containing serotypes: 1/2b, 3b, 4b, 4e and 7; lineage II: serotypes 1/2a, 1/2c, 3a and 3c; lineage III: serotypes 4b, 1/2a, 4a and 4c and lineage IV: 4a and 4c. The genetic lineages have different, although at times overlapping, genetic, phenotypic and epidemiological characteristics with the majority human illness caused by strains in lineages I and II. Thus 7-locus MLST, in common with other current molecular typing techniques, in focusing on only a small portion of the bacterial genome, provides insufficient strain resolution for detailed epidemiological investigations and the use of other molecular typing techniques are also required.

With the advent of next generation sequencing technologies, entire bacterial genomes are now readily available for analysis affording the highest level of strain discrimination, the ability to infer phylogenetic relationships and access to a wealth of additional information such as virulence and resistance markers. It is anticipated that in the near future whole genome sequencing (WGS) will replace the currently used typing methods for foodborne pathogens, as it is now possible to obtain a multitude of different characteristics based on WGS data in real time at a reasonable cost. The value of WGS as a bacterial typing tool has already been assessed in a number of specific settings and WGS is increasingly used for outbreak investigations and source tracing of *L. monocytogenes*. Thus, WGS was used in several recent national studies for outbreak detection and investigations, e.g. in Austria (Rychli et al., 2014), Australia (Kwong et al., 2016), USA (Jackson et al., 2016), Denmark (Kvistholm Jensen et al., 2016). The improved resolution obtained by WGS enabled linking of isolates and more robust case definitions enabling isolates to be ruled in or out of outbreaks. Furthermore, WGS makes it possible to recognise extended time-period outbreaks and link clinical cases to food products and food production facilities (Gillesberg Lassen et al., 2016). Such studies demonstrate the advantages of using WGS analysis for national surveillance and outbreak detection and investigation. There have also been studies where WGS analysis has been used to investigate the strains that persist in the food processing environment and shown to distinguish those that are persistent from ones that are repeatedly reintroduced (Stasiewicz et al., 2015).

However, the experience of using WGS for international comparison of isolates and linking human cases to food products or food production facilities across borders is still very limited. This study aimed at supporting the future European-wide use of WGS for improved food safety. The main objective of the study was to compare *L. monocytogenes* isolates collected in the EU from ready-to-eat (RTE) foods, along the food chain and from human cases. The isolates were whole genome sequenced and the data were used for a number of analyses aiming at describing the phylogeny and diversity of isolates from foods and humans in Europe, to evaluate the use of WGS for outbreak investigations, to evaluate the possible epidemiological relationship between isolates, and to identify possible markers of virulence/survival. More than 1,100 isolates were selected for the study (Section 2). The sequences were analysed with a variety of methods in order to fulfil the defined objectives of the study. This has facilitated the comparison of a variety of WGS analysis methods using a large data set from different European countries. This type of study has not been performed previously but is essential to ensure the adoption at the European level of robust, harmonised WGS analysis methods for national and international surveillance, outbreak detection and investigation.

1.1. Background and Terms of Reference as provided by the requestor

In the European Union (EU), listeriosis continues to be a serious food-borne illness, with high morbidity, hospitalisation and mortality in vulnerable populations. For example in 2012, 1,642 confirmed human cases of listeriosis were reported including 198 deaths.¹ The trend in reported human listeriosis cases has been gradually increasing over the past four years.

The main route of transmission to humans is through consumption of contaminated food. The bacterium can be found in raw foods and in processed foods that are contaminated during and/or after processing. Because *L. monocytogenes* is able to multiply at low temperatures (2 to 4°C), ready-to-eat (RTE) foods with a relatively long shelf-life (such as fishery products, meat products and cheese) are of particular public health concern.

An EU-wide baseline survey (BLS) was conducted in 2010 and 2011 to estimate the prevalence and contamination levels in three RTE foods at retail in accordance with Decision 2010/678/EU: packaged (not frozen) smoked or gravad fish (3,053 samples), packaged heat-treated meat products (3,530 samples) and soft or semi-soft cheeses (3,452 samples). The Part A report (prevalence estimates) was published in 2013.² The EU prevalence of fish samples at the time of sampling was 10.4 % and at the end of shelf-life 10.3 %, while for meat and cheese samples at the end of shelf-life these prevalences were 2.07 % and 0.47 %, respectively. The Terms of Reference of the subsequent Part B report are (a) the analysis of (risk) factors related to the prevalence of contaminated foods, (b) the development of predictive models for the microbial growth of *L. monocytogenes* under various storage conditions, and (c) the development of predictive models for compliance with *L. monocytogenes* food safety criteria in RTE foods. Publication of this report is expected in June 2014.

In a self-task mandate by the BIOHAZ Panel, information on current and prospective molecular sub-typing methods for food-borne pathogens (among which *L. monocytogenes*) has been reviewed in terms of discriminatory capability, reproducibility, and capability for international harmonisation. The opinion was published at the end of 2013.³ Molecular approaches to characterise isolates, specifically using sequence-based approaches as those based on whole genome sequence (WGS) analyses, provide the means of describing and characterising the variation of bacterial populations with the highest resolution possible, enhancing substantially our ability to understand and trace the sources and spread of the diseases that they may cause.

Main objective:

The main objective of the contract resulting from the present procurement procedure is to compare *L. monocytogenes* isolates collected in the EU from RTE foods, compartments along the food chain and humans using whole genome sequencing (WGS) analysis.

The specific objectives are as follows:

¹ EFSA and ECDC, 2014. The European Union Summary Report on Trends and Sources of Zoonoses, Zoonotic Agents and Food-borne Outbreaks in 2012. EFSA Journal 2014;12(2):3547, 312 pp. doi:10.2903/j.efsa.2014.3547

² European Food Safety Authority, 2013. Analysis of the baseline survey on the prevalence of *Listeria monocytogenes* in certain ready-to-eat (RTE) foods in the EU, 2010-2011 Part A: *Listeria monocytogenes* prevalence estimates. EFSA Journal 2013;11(6):3241, 75 pp. doi:10.2903/j.efsa.2013.3241

³ EFSA BIOHAZ Panel (EFSA Panel on Biological Hazards), 2013. Scientific Opinion on the evaluation of molecular typing methods for major food-borne microbiological hazards and their use for attribution modelling, outbreak investigation and scanning surveillance: Part 1 (evaluation of methods and applications). EFSA Journal 2013;11(12):3502, 84 pp. doi:10.2903/j.efsa.2013.3502

Specific objective 1: to carry out the molecular characterisation of a selection of *L. monocytogenes* isolates from different sources, i.e. RTE foods, compartments along the food chain (e.g. food producing animals, food processing environment), and humans employing WGS analysis.

The tenderer is expected, as part of the project, to present first a list of the *L. monocytogenes* isolates from different sources, i.e. RTE foods, compartments along the food chain (e.g. food producing animals, food processing environment), and humans that are accessible for WGS analysis. Isolates from RTE foods shall at least consist of *L. monocytogenes* isolates collected from the BLS in RTE foods. In the BLS 134 packaged smoked or gravad fish samples were found contaminated with *L. monocytogenes* at the time of sampling, 133 at the end-of-shelf-life, and 176 at both stages. Also, 72 and 16 positive samples were detected in packaged heat-treated meat products and soft and semi-soft cheeses, respectively. These isolates are stored at the EU Reference Laboratory for *L. monocytogenes* and/or individual National Reference Laboratories for *L. monocytogenes*.⁴ EFSA will ensure access for the successful contractor to these isolates, however, all the costs related to the preparation and transport of these isolates to their own premises shall be covered by the contractor. The successful contractor will not be authorised to use these isolates for any purpose outside the remit of this project and will have to destroy them after completion of the project. An example of "Agreement for the transfer of Materials" to be signed between the Contractor and the EURL and/or the NRL before the shipment of the isolates can be found in Annex 7 of these tender specifications. Further, *L. monocytogenes* isolates from other EU sources should also be included (e.g. RTE foods, food producing animals, food processing environment) as well as human isolates from the same period of the baseline survey (i.e. 2010-2011). If necessary, the isolate collection can be complemented with other isolates from recent years (2012-2014). For each isolate relevant metadata should be available in order to fulfil the objective. The tenderer is responsible for the identification of the origin (e.g. geography, laboratory) of these isolates. Consideration of the sources, including the number of isolates and related metadata, should be made with the overall aim of maximising the outcome of the objectives of the study.

Then the criteria employed by the tenderer for the selection of the isolates for WGS analysis should be clearly described in the offer. The minimum and/or maximum number of isolates to be included is left to the discretion of the tenderer, with an estimate of this number to be provided in the offer. The total number of isolates proposed should soundly represent isolates from different sources as stated above and humans as well as from various geographical regions in the EU. Both the selection criteria and the number of isolates proposed should be indicated in the offer and applied with the overall aim of maximising the outcome of the objectives of the study.

Although the primary responsibility to list and select *L. monocytogenes* isolates is with the tenderer, EFSA may provide, during the implementation of the contract, information on potential available sources of isolates in particular from humans. The successful contractor should consult and agree with EFSA on the final selection of the set of isolates to be typed. The WGS typing should be carried out with state-of-the-art equipment and methodologies which conform with current laboratory standards and that can be referred to or reported in a clear and concise manner. Robust annotation pipelines for the WGS data generated should be designed and implemented with the aim of getting a harmonised framework for subsequent data analysis.

Specific objective 2: to analyse the WGS typing data of the selected *L. monocytogenes* isolates with three goals:

⁴ No objections were raised at the Meeting of the SCoFAH (Section Biological Safety of the Food Chain), held in Brussels on 16 October 2013 for accessing the isolates from the EU BLS and its epidemiological data for this activity. Storage of isolates by the NRLs or EU RL *Lm* beyond the minimum duration of 2 years has been requested.

- i. to explore the genetic diversity of *L. monocytogenes* within and between the different sources and human origin;
- ii. to assess the epidemiological relationship of *L. monocytogenes* from the different sources and of human origin considering the genomic information and the metadata available for each isolate;
- iii. to identify the presence of putative markers conferring the potential to survive/multiply in the food chain and/or cause disease in humans (e.g. virulence and antimicrobial resistance).

Specific objective 3: to perform a retrospective analysis of outbreak strains (i.e. using a subset of epidemiologically linked human and food isolates) to investigate the suitability of WGS as a tool in outbreak investigations:

Strains from known food-borne outbreaks of human listeriosis should be characterised employing WGS methods and analysed following the methodological frame employed under objective 1 above. Next, the available WGS data should be analysed for establishing and/or supporting links between the different strains. The outcome of this analysis should provide an evaluation on the advantages and limitations of employing WGS data for investigating outbreaks of food-borne listeriosis.

This contract was awarded by EFSA to:

Contractor: Statens Serum Institut (SSI), French Agency for Food, Environmental and Occupational Health & Safety (ANSES), Public Health England (PHE), University of Aberdeen (UA)

Contract: Closing gaps for performing a risk assessment on *Listeria monocytogenes* in ready-to-eat (RTE) foods: activity 3, the comparison of isolates from different compartments along the food chain, and from humans using whole genome sequencing (WGS) analysis.

Contract number: OC/EFSA/BIOCONTAM/2014/01-CT 1

2. Isolate collection

2.1. Selection of strains to fulfil the main and specific objectives of the study

A geographically wide selection of isolates from human cases and RTE foods that are likely to be the direct source of human *Listeria monocytogenes* infections forms the main basis for the analyses to fulfil objectives 2i and 2ii (genetic diversity and epidemiological relationship). High priority was given to this selection by including 262 and 610 isolates from sporadic human cases and retail food samples, respectively, to enable meaningful epidemiological linking as well as robust source attribution modelling.

As stated in the tender specifications, isolates from the EU-wide baseline survey (BLS) conducted in 2010 and 2011 (EFSA, 2013) should be included in this project. Therefore, the selection of RTE food isolates was based on the availability of isolates from the BLS, which included three types of RTE food: smoked and gravad fish, packaged heat-treated meat products, soft and semi-soft cheeses. These types of food are considered important sources of *L. monocytogenes* causing human illness (Greig and Ravel, 2009; Batz et al., 2012). A total of 353 isolates from the BLS were selected for this project according to their availability and the criteria described in the following sections. However, most of the available BLS isolates were from fish products whereas a much smaller portion was from meat products and cheeses (EFSA, 2013). Therefore, additional isolates from RTE meat products and cheeses were obtained from as many different EU Member States as possible to restore the collection of RTE food isolates to a more balanced set of the three RTE food categories. The choice of focusing

on the food categories represented in the BLS will strengthen the conclusions in relation to these specific and highly relevant sources, but the potential of making strong conclusions on the significance of other potential food sources is limited. A few isolates from fruits and vegetables were also selected.

It was anticipated that a substantial number of isolates from live animals would be needed to fully represent live animals as a source in the epidemiological analyses, especially considering that the genetic variation in the bacterial population is likely to decline with transfer to the next stage in the food chain (environment – animal – production facility – food – human). Since the study was limited to approximately 1,000 isolates to be sequenced within the project it was decided not to include isolates from live animals as in source attribution strains isolated at production facility stage or in foods are generally used (Pires et al., 2009). An additional reason to avoid the animal samples is that it is very hard to find animal samples with proven epidemiological links to illness or contaminated products (e.g. foods). An exception was made for samples obtained from specific fisheries.

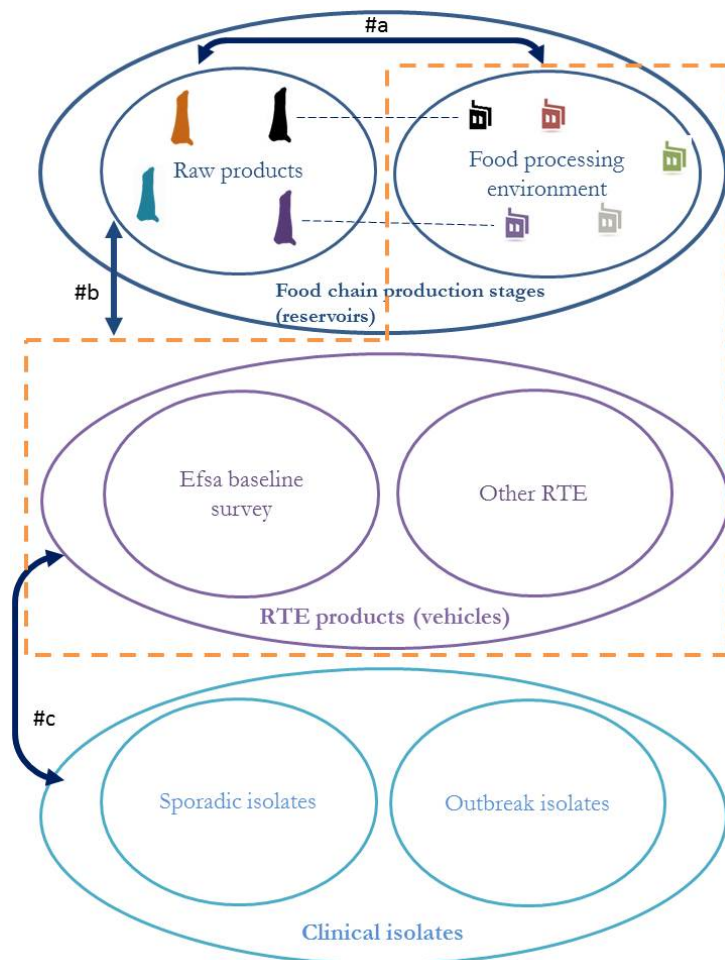
Source attribution modelling was based on the host animal (e.g. ovine, bovine, fish etc.). Since information is available on the animal origin of the food matrix (e.g. the origin of the milk is known, bovine/sheep/goat, for milk and milk products) the different foods can be linked to their host animal.

Isolates from raw food sampled at production sites of meat, milk and fish production as well as environmental isolates from such production sites (potentially persistent strains) are included in this project (Figure 2.1 #a). It was not possible in all instances to obtain raw products and environmental isolates taken at the same production plant. This was mainly due to raw products and end products not being produced in a single factory as it occurs for example with cheese.

Data for the specific objective 2iii was obtained by comparing isolates from human infections and RTE foods (Figure 2.1 #c) by looking for genes/alleles overrepresented in those from human cases compared to RTE food as well as specific animal reservoirs as the bovine, ovine, etc. (putative markers conferring the potential to cause disease in humans). Likewise, isolates from raw food were compared to the isolates from potentially persistent strains and RTE food (Figure 2.1 #b) to look for putative markers conferring a better potential to survive/multiply in the food chain. The chance of identifying types of *L. monocytogenes* overrepresented in human cases compared to the presence in the main groups of RTE food is considered reasonably high (at least giving the possibility of identifying potential markers for human disease). Identifying markers for potential to survive/multiply in the food chain is less straight forward due to the expected high diversity of strains entering the food chain (e.g. (Rückerl et al., 2014)) – and most likely different strains of *Listeria* are continuously entering a specific facility. In this study, data is analysed on a general population level based on a sample of isolates representing the different compartments. It is not possible, in this study, to investigate the dynamics and circulation of strains within a specific factory or food processing chain (Figure 2.1, dashed lines). This would require very large data sets and a large complete study involving sampling, see e.g. (Muhterem-Uyar et al., 2015), and this is not part of the scope of the present study.

To fulfil specific objective 3, isolates from nine retrospective outbreaks, including human cases and related food isolates, were selected to represent different sources of outbreaks in different geographical regions. The results of the analysis of the nine epidemiologically described outbreaks show the degree of genetic variation between strains associated with a single outbreak and give indications on the usefulness of WGS in outbreak investigations.

The specific criteria applied for the selection of isolates are described in the following sections.



The links between groups of strains #a, #b and #c will help to fulfil specific objectives (see text).

Figure 2.1. Representation of the different categories of isolates from different stages of food chain and isolates implicated in clinical stages selected in the study

2.2. Strains from the baseline survey (BLS)

The isolates that have been collected in the baseline survey were the target priority for sequencing. Not all the isolates were sequenced. The selection criteria of the isolates are described in Sections 2.2.1, 2.2.2 and 2.2.3.

2.2.1. Selection criteria BLS 1. Availability of the strains

Some strains in the BLS were not sent to the EURL and therefore not available for this study. The available strains came from 22 MSs and 1 non-MS out of the 24 participants in the BLS.

2.2.2. Selection criterion BLS 2: Isolates from same origin (food and Member State)

2.2.2.1. Selection criterion BLS 2a: Isolates from paired samples

Analyses of *L. monocytogenes* were made at the end of shelf-life for all three types of the surveyed RTE foods and, also, at the time of sampling for the fish samples. Isolates from paired fishery product samples with the same molecular profile (PFGE) at initial sampling and at end of shelf life were usually considered to be the same strain and therefore sequencing both isolates would not provide additional

useful data. The strains isolated from the product at the end of shelf life were the ones selected for sequencing.

2.2.2.2. Selection criterion BLS 2b: Isolates from same origin (food and Member State)

It is also likely that the same strains are represented amongst isolates provided by the same MS from the same type of food, e.g. in the BLS, out of a total of 12 isolates from country B fishery product there were only 3 different PFGE profiles. This information on diversity of PFGE types amongst BSL isolates indicated that the same product (that is coming from a unique factory) was sampled in a unique retailer. Therefore, not all of these isolates were selected.

2.2.3. Selection criterion BLS 3. Multiple isolates from a sample (all categories)

Although one isolate per positive food sample should have been selected, several laboratories collected more than one isolate (2 to 6). It was decided to include different isolates that originated from the same sample as long as typing information confirmed that these isolates were different (e.g. based on their PFGE profile). This was the case for nine samples. Accordingly, when typing data was unable to distinguish them, one of the isolates was selected randomly.

The selected BLS isolates consist of 353 isolates coming from 22 MSs + 1 non-MS. The 297, 49 and 7 strains were respectively isolated from RTE fishery products, meat products, and soft and semi-soft cheeses. The complete list of isolates selected after applying selection criteria is given in Appendix 1.

2.3. Strains from other foods (OF)

The BLS provides numerous isolates for fishery products but far less for cheese and meat products. In order to increase the number of strains from cheese and meat, eight Member States provided isolates from these food products. The following criteria were applied to the available strains.

2.3.1. Selection criterion OF 1. Food origin

Strains from meat and dairy products were selected. Only isolates from RTE foods were considered.

2.3.2. Selection criterion OF 2. Temporal criterion

In order to match with BLS and clinical isolates, these were selected from the 2010-2012 period.

2.3.3. Selection criterion OF 3. Further selection

As more than 500 strains were available for RTE meat products and more than 300 for cheese products after application of selection criteria OF 1 and OF 2, an additional selection was needed. For MSs and non-MS where subtyping information (PFGE, AFLP, agglutination/molecular serotype) was available, the most prevalent groups were selected. For the other MSs a random sampling (each isolate has the same probability to be sampled) has been carried out. From the country Q, 12 sequenced isolates from soft and semi-soft cheese were already available.

2.3.4. Fruit and vegetables

Few strains were available from fruits and vegetables. Fruits and vegetables are usually not routinely tested for *L. monocytogenes*. Five isolates from the country B that were isolated within the time frame of the BLS were included see Appendix 3.

The category of isolates constituting "other foods" is composed of 218 isolates (including 12 already sequenced) coming from 8 MSs (A, B, C, G, Q, V, X and Z) with a 126 and 80 isolates from RTE meat products and cheese, respectively. In addition, 12 strains were assigned to the combined food

category. The complete list of isolates selected after applying selection criteria is given in Appendix 2. The six isolates from fruits and vegetables are listed in Appendix 3.

2.4. Strains from food chain production stages

In order to compare isolates from both the baseline survey and clinical cases to isolates persisting or circulating in food processing chains, we chose four sets of strains corresponding to the three types of RTE food in the BLS. In order to match with the BLS and clinical isolates, these were preferentially selected from the 2010-2012 period. The period was sequentially extended until the desired number of isolates was reached. The strains are presented in Appendix 4.

The first set constitutes 62 strains isolated from pork meat cuts and meat products and from food processing environments (between 2003 and 2014) in various regions of country C and from food processing environments in country B (2010-2011). The list of strains is given in Appendix 4. The selection of pork production isolates from country C includes isolates that will be sequenced in the context of another study and made available for this study: 6 isolates from pork processing environment (potentially persistent strains) and 2 from raw products.

The second set constitutes 21 strains isolated from raw milk and cheese and semi-soft cheese production environments in the country B.

The third set consists of 100 strains isolated from cheese and semi-soft cheese factories in the country Q. These isolates were already sequenced and the genome sequences were made available for this project.

The fourth set constitutes 29 strains isolated from raw fish and smoked salmon production environments (e.g. environmental swab samples). These strains come from the country B.

The category of food chain production isolates consists of a total of 200 isolates coming from 3 MSs (countries B, C and Q). Of these, 142 were isolated from environmental samples. The complete list of isolates is given in Appendix 4. Ninety-six of these isolates were already sequenced before this project.

2.5. Strains from sporadic human clinical cases

Clinical isolates from assumed sporadic human cases were included in the study. Isolates were from the baseline survey period, 2010-2011, and priority was given according to the criteria stated below. It has not been possible to obtain clinical isolates from all relevant MSs, as some laboratories/MSs were not willing to contribute to the project.

2.5.1. Selection criterion C1. Availability of the strains.

Isolates from the country A and B were directly available for the project since the national clinical reference laboratories from these MSs were partners in the project (NPHLs of country A and B, respectively). Furthermore, we got positive response from nine MSs that were willing to contribute clinical isolates: country C, D, F, Q, T, W, X, Y, Z.

2.5.2. Selection criterion C2. Temporal and geographical criterion

In order to match with BLS and other food isolates, the 2010-2011 period was preferred. The selection strategy was to encompass a wide geographical distribution. From some MSs (F, Q, W), only a small number of isolates from these years were available and all of these were therefore included. For MSs and one non-MS with a smaller population size (A, D, T, Y, Z), 20 isolates were selected for sequencing, while from MSs with large populations (B, C, X) - and a correspondingly higher number of listeriosis cases - 35 isolates were selected from each. In addition, a number of isolates that have already been sequenced were included.

2.5.3. Selection criterion C3. Incidence

An effort was made to get not only MSs with large areas/populations represented in the project, but also those with high incidence. According to the ECDC “Annual epidemiological report 2014 - food- and waterborne diseases and zoonoses” the Nordic MSs and non-MS have among the highest incidences in Europe (five of the six highest), and in this study three are included. These high incidences could be an artefact of different surveillance systems in the Nordic vs other MSs and non-MSs, but at the European level, this publication was the most reliable estimate available.

2.5.4. Selection criterion C4. Subtype information.

For MSs where subtyping information (PFGE, AFLP, agglutination/molecular serotype) was available, the most prevalent groups were selected. For the other MSs a random sampling (each isolate has the same probability to be sampled) has been carried out.

The clinical isolates constitute 262 isolates from 10 MSs + 1 non-MS. A total of 250 isolates were sequenced in this project whilst whole genome sequences were already available for 16 isolates. The complete list of isolates selected after applying selection criteria is given in Appendix 5.

2.6. Strains from outbreaks

To investigate the suitability of WGS as a tool in outbreak investigations, isolates from epidemiologically confirmed retrospective outbreaks were selected. Isolates from nine well-described outbreaks in three MSs + one non-MS are available for the project. Clinical isolates as well as isolates from suspected or confirmed sources are included. From some outbreaks, all available isolates were included. From large outbreaks, a random selection of up to 25 isolates was made.

Table 2.1: Number of human and food isolates in each of the nine outbreaks

	Country	Human	Food	Vehicle of infection
Outbreak 1	B	5	10	Beef
Outbreak 2	B	5	3	Crab meat
Outbreak 3	B	5	4	Sandwiches
Outbreak 4	B	2	2	Ox tongue
Outbreak 5	B	9	1	Unknown
Outbreak 6	T	4	1	Rakfisk
Outbreak 7	X	13	6	Foie gras
Outbreak 8	X	4	9	Cheese
Outbreak 9	C	25	0	Brie cheese

The set of outbreak isolates (Table 2.1) consists of 105 isolates from nine outbreaks; from eight of these, both clinical and food isolates were available. The sample size per outbreak ranged from 5 to 25 isolates. Whole genome sequences were already available for 13 isolates. The remaining isolates have been sequenced in the project. The complete list of isolates is given in Appendix 6.

2.7. Strain selection summary

A summary of the final set of strains included in the project is given in Table 2.2. In total, 676 “food-related” strains were selected for sequencing within the project: BLS, “other foods” and “food chain production stages”. PHE contributed with 100 additional already sequenced strains from cheese procession plants in country Q (including isolates from 12 food samples and 88 environmental samples; listed in Appendix 4). In total, 262 assumed sporadic human clinical isolates were selected

for the study - 16 of these are already sequenced. Nine epidemiologically confirmed outbreaks were represented by 105 isolates from either human infections or food samples related to the outbreak (13 of these were already sequenced). The strain collection was sent to PHE for sequencing in batches. The final complete strain collection was held at PHE during the project.

Table 2.2.: Repartition of the 1,143 isolates according to country and context of isolation

Country	Baseline Study Appendix 1	Other foods Appendix 2+3	Food production chain Appendix 4	Clinical, sporadic Appendix 5	Outbreaks Appendix 6	Total
A	7	29		35		71
B	4	28	68	31	43	174
C	35	83	32	35	25	210
D	4			20		24
E	6					6
F	15			8		23
G	4	4				8
H	5					5
J	10					10
K	14					14
L	54					54
M	2					2
N	9					9
P	3	4				7
Q	33		100	23		156
R	4					4
S	4					4
T	4			20	5	29
U	62					62
V	6	28				34
W	7			15		22
X	38	34		35	32	139
Y	8			20		28
Z	15	13		20		48
Total	353	223	200	262	105	1,143

Table 2.3 presents in detail the repartition of the 776 strains corresponding to Appendices 1, 2, and 3.

Table 2.3.: Repartition of the 776 food isolates according to country, food matrix and source

Country code	Food matrix	Source	Number of isolates	Total per country
A	Elaborated food products combining several food categories	Mixed sources	1	
	Fish and fishery products	Fish	6	
	Meat and meat products	Bovine	2	

Country code	Food matrix	Source	Number of isolates	Total per country
	Meat and meat products	<i>Gallus gallus</i> (fowl)	3	
	Meat and meat products	Sheep	1	
	Meat and meat products	Swine	8	
	Meat and meat products	Unspecified	14	
	Milk and milk products	Bovine	1	36
B	Elaborated food products combining several food categories	Mixed sources	5	
	Fish and fishery products	Fish	33	
	Fruit, vegetables, cereals and herbs	Vegetal	5	
	Meat and meat products	Bovine	4	
	Meat and meat products	<i>Gallus gallus</i> (fowl)	1	
	Meat and meat products	Mixed animal source	3	
	Meat and meat products	Poultry not specified	1	
	Meat and meat products	Swine	6	
	Meat and meat products	Unspecified	12	
	Milk and milk products	Bovine	5	
	Milk and milk products	Goat	1	
	Milk and milk products	Unspecified	24	100
	C	Fish and fishery products	Fish	31
Meat and meat products		Ducks	3	
Meat and meat products		<i>Gallus gallus</i> (fowl)	2	
Meat and meat products		Mixed sources	3	
Meat and meat products		Poultry not specified	1	
Meat and meat products		Swine	51	
Meat and meat products		Unspecified	20	
Milk and milk products		Bovine	35	
Milk and milk products		Goat	1	
Milk and milk products		Unspecified	3	150
D	Fish and fishery products	Fish	3	
	Milk and milk products	Bovine	1	4
E	Fish and fishery products	Fish	6	6
F	Fish and fishery products	Fish	15	15
G	Fish and fishery products	Fish	3	
	Meat and meat products	Swine	5	8
H	Fish and fishery products	Fish	3	
	Milk and milk products	Unspecified	2	5
J	Fish and fishery products	Fish	9	
	Meat and meat products	<i>Gallus gallus</i> (fowl)	1	10
K	Fish and fishery products	Fish	14	14
L	Fish and fishery products	Fish	42	
	Meat and meat products	Mixed sources	2	

Country code	Food matrix	Source	Number of isolates	Total per country
M	Meat and meat products	Swine	9	54
	Milk and milk products	Bovine	1	
M	Meat and meat products	Fish	1	2
	Meat and meat products	Swine	1	
N	Fish and fishery products	Fish	9	9
P	Fish and fishery products	Fish	2	7
	Meat and meat products	Swine	1	
	Milk and milk products	Bovine	4	
Q	Fish and fishery products	Fish	30	133
	Meat and meat products	Swine	1	
	Milk and milk products	Sheep	100	
	Milk and milk products	Unspecified	2	
R	Fish and fishery products	Fish	1	4
	Meat and meat products	Unspecified	3	
S	Fish and fishery products	Fish	4	4
T	Fish and fishery products	Fish	4	4
U	Fish and fishery products	Fish	55	62
	Meat and meat products	<i>Gallus gallus</i> (fowl)	1	
	Meat and meat products	Mixed sources	1	
	Meat and meat products	Swine	2	
	Meat and meat products	Turkeys	2	
	Milk and milk products	Unspecified	1	
V	Fish and fishery products	Fish	6	34
	Meat and meat products	Unspecified	5	
	Milk and milk products	Sheep	16	
	Milk and milk products	Unspecified	7	
W	Fish and fishery products	Fish	7	7
X	Elaborated food products combining several food categories	Mixed sources	9	72
	Fish and fishery products	Fish	21	
	Meat and meat products	Bovine	1	
	Meat and meat products	Ducks	1	
	Meat and meat products	<i>Gallus gallus</i> (fowl)	5	
	Meat and meat products	Geese	1	
	Meat and meat products	Mixed sources	1	
	Meat and meat products	Swine	24	
	Meat and meat products	Turkeys	1	
	Meat and meat products	Unspecified	2	
	Milk and milk products	Goat	1	
	Milk and milk products	Unspecified	5	
Y	Fish and fishery products	Fish	8	8

Country code	Food matrix	Source	Number of isolates	Total per country
Z	Elaborated food products combining several food categories	Mixed sources	2	
	Fish and fishery products	Fish	11	
	Meat and meat products	Bovine	7	
	Meat and meat products	<i>Gallus gallus</i> (fowl)	2	
	Meat and meat products	Swine	5	
	Meat and meat products	Unspecified	1	28

2.8. Database

All available information on isolate characteristics and associated descriptive epidemiological information have been collected from isolate providers and organised in a database. This database links the WGS typing data and metadata associated for each strain. The database is specific to this project and will be used only for this study. The database was developed and managed by Bionumerics software (version 7, Applied Maths, Sint-Martens-Latem, Belgium). Supplementary Excel file contains all the information included in the database (Annex A).

2.8.1. Database structure

The database has a structure similar to the EURL *Lm* DB.

Most of the fields contain predefined pick lists to avoid errors in reporting. The food matrices and the description of the food nature listed within the LISEQ database respected the EFSA standard sample description code (SSD2) (Félix et al., 2014). Food product description followed three level hierarchical information. Food products were first classified according to the food matrix type (see Figure 2.2). Then for each food matrix, food products are listed. Finally, information on process is provided. Figure 2 describes the information fields of the project database.

The database (both in Excel and in Bionumerics format) also includes strain-typing-results extracted from the genome analysis, such as MLST, Clonal complex, SNP address and cgMLST (only descriptive fields are shown in Figure 2.2).

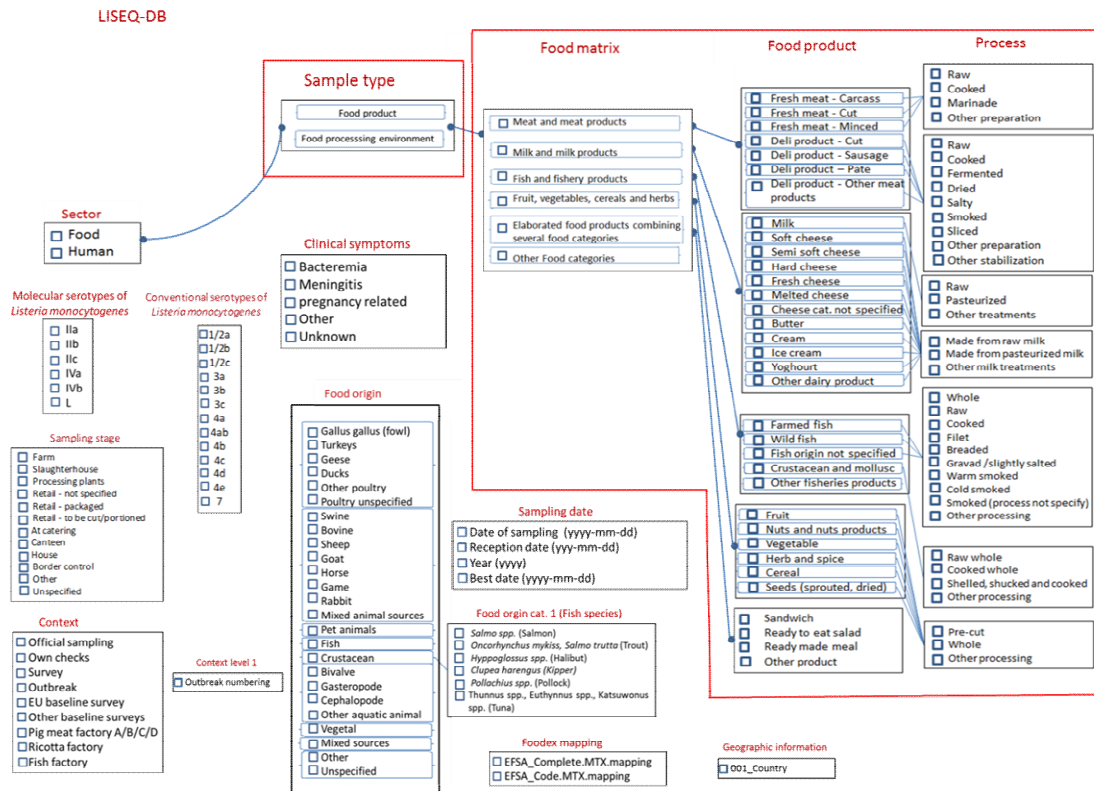


Figure 2.2.: Description of the information fields in the LISEQ-DB

2.8.2. Core information on strains

All the strains (Table 2.2) do not have the same degree of associated information. For some strains full information is available with three levels information on food – Food matrix/Food products/Process (e.g. Meat and Meat Products/Deli products-pate/sliced) – for other strains the information is less detailed. Yet, the following core information is at least present for the food strains:

- Sample type
- Geographical information (at least Country)
- Sampling date (at least year)
- Food matrix/Food products
- Food origin

For strains from sporadic human listeriosis cases, the information is limited to sampling date (at least year) and geographical information. Food outbreak strains shared the same degree of information as other food strains. For all the clinical strains, an additional information field contains the “clinical symptoms” data with five options (bacteremia, meningitis, pregnancy related, other, unknown).

2.8.3. Data export

For each specific objective (epidemiological relationship and source attribution modelling), outputs with the required information field were exported in appropriate format (csv, etc.).

3. Methodologies

In order to fulfil the main objective of the contract, i.e. to *compare L. monocytogenes isolates collected in the EU from RTE foods, compartments along the food chain and humans using whole genome sequencing (WGS) analysis*, a range of methodologies were employed. This Section provides an overview of the methods and the rationale for choosing these.

The first specific objective was to *carry out the molecular characterisation of a selection of L. monocytogenes isolates from different sources, i.e. RTE foods, compartments along the food chain (e.g. food producing animals, food processing environment), and humans employing WGS analysis*. As a first step towards this goal, DNA was extracted from the selected isolates and subjected to WGS. The platform employed for sequencing was the Illumina HiSeq, which is the most commonly employed cost effective, rapid method for sequencing high numbers of bacterial genomes and thus as close to a standardized sequencing method as possible. Additionally, sequence data obtained by Illumina sequencing contains less sequencing errors as compared to several other sequencing platforms. The current sequencing methods rely on massive parallel sequencing, meaning that instead of sequencing the genomes from one end to the other, each genome is fragmented and these small fragments are sequenced simultaneously, in parallel. Thus, the sequence data produced from this type of setup consists of millions of sequence reads per isolate genome, with each read typically of a size around 100-200 nucleotides. Subsequently, the millions of reads are pieced together thereby assembling an almost complete genome. The procedures for sample preparation, sequencing, quality control and assembly are described in Section 4.

The WGS data generated was analysed by different bioinformatics procedures (described in Section 4) to explore the phylogeny and to produce data sets (typing data) that could form the basis for the further analysis and interpretation of data for the specific objectives 2 and 3.

In order to perform molecular characterisation of isolates several gene-by-gene approaches were employed in this project. Firstly, 7-locus MLST as defined in Ragon et al. (2008) was extracted from the WGS data, and although now based on WGS this way of performing MLST is completely comparable with the conventional form of MLST based on Sanger sequencing of the seven loci. Sequence types (STs) were defined on basis of the allelic sequences of the seven loci and employed to assign isolates to clonal complexes (CC's).

Listeria monocytogenes is a very diverse species and contains four divergent lineages with a high number of variants, and therefore to be able to perform fine resolution phylogenies. Single Nucleotide Polymorphism (SNP) analysis was performed separately for distinct clonal complexes, employing the different reference genomes for each clonal complex, i.e. the best suitable for each clonal complex. This was done in order to obtain the maximal phylogenetic resolution. In brief, SNP analysis is a method in which phylogeny is inferred on the basis of sequence variations between isolates, across the parts of the genome that are shared by all isolates included in the analysis, i.e. the core genome. SNPs are assigned by comparing (mapping) sequence reads for each isolate against a common reference genome, here a reference genome specific for each clonal complex, and subsequently documenting the difference between any two isolates.

SNP analysis is at present the most widely used method for WGS-based discrimination for public health surveillance and outbreak detection, but at present the method is not standardised, which results in SNP analysis being variable between laboratories and results difficult to directly compare. The number of SNPs is dependent on sequencing technology, sequence data quality and reference genome employed, and even varies depending on the number of isolates included in the analysis.

Due to the SNP analysis not yet being standardised, gene-by-gene methods such as ribosomal MLST (rMLST) and core-genome MLST (cgMLST) may be more suited to public health surveillance and outbreak investigation of gastrointestinal bacterial pathogens, since they can be carried out in a more standardised way and especially because the results are easier to communicate. cgMLST is a gene-by-gene approach similar to the conventional MLST but instead of being based on seven loci it is based on the core genome of the species, with the present cgMLST scheme including 1748 loci (Moura et al.,

2016). In cgMLST allelic variations within each of these loci make up the final type, and is used to differentiate between clonal isolates. Typically, cgMLST (and rMLST) are performed on assembled genome data, in contrast to SNP analysis.

Further analysis and interpretation of data for specific objectives 2 and 3 were based on the WGS typing data sets produced under objective 1 described above. For Specific Objective 3, *to perform a retrospective analysis of outbreak strains to investigate the suitability of WGS as a tool in outbreak investigations*, isolates from nine food-borne outbreaks of human listeriosis were selected. These outbreaks represented a range of different characteristics with respect to a number of factors such as food source, time span, geography, number of cases, etc. For each outbreak, the relation between the human clinical isolates, isolates available from the suspected sources as well as similar background isolates were characterised (Section 5). For outbreak investigations, high-discriminatory methods are desirable and thus SNP and cgMLST analyses were employed for characterising the outbreaks. The SNP and allele differences (cgMLST) seen for epidemiologically related and un-related isolates in this retrospective analysis of outbreaks provided valuable input to the analysis parameters used for assessing the epidemiological relationship between *L. monocytogenes* isolates of human origin and those from different sources (Objective 2). Therefore, the analysis of retrospective outbreaks (Objective 3) is presented before the analyses related to Objective 2.

For the first goal of Specific Objective 2, *to explore the genetic diversity of L. monocytogenes within and between the different sources and human origin*, different indices and measures were applied to characterise and describe the variation of isolates in the collection (presented in Section 6). These included diversity index, rarefaction curves and Nei's genetic distance. These measures can give an overall understanding of the diversity within and between sources/reservoirs and differences can be statistically tested. To give useful description of diversity, diversity index and rarefaction curves must be based on typing methods producing a limited number of distinct types in the data set. We chose MLST (7 loci) as well as rMLST (30 loci) as these are both well-established typing methods with an internationally recognised nomenclature. In addition, cgMLST and SNP data was also used for assessing the genetic distance between populations.

We employed two different approaches for the second goal of Specific Objective 2, *to assess the epidemiological relationship of L. monocytogenes from the different sources and of human origin considering the genomic information and the metadata available for each isolate*. The first approach was using the method of source attribution (Section 7, i.e., partitioning of the human disease burden of listeriosis to specific sources, where the term source includes animal reservoirs and vehicles (e.g. foods). Because of the relatively small number of isolates, all of the isolates along the food chain that originate from a particular reservoir were combined. This enabled the following sources of isolates and their respective genomes to be determined: bovine, ovine, swine, fish and poultry. Human clinical cases were attributed to these sources by utilising five different mathematical models and the genomic typing data that was generated as described in Section 4.

The second approach to assess the possible epidemiological relationship between strains was to identify clusters of clinical and food isolates based on SNP differences (Section 8). The WGS data was analysed along with the epidemiological information of the food and clinical isolate to assess, retrospectively, relationships between circulating strains of *L. monocytogenes* in EU within 2010-2012 period. Clusters of interest were further investigated by focusing on metadata associated to each strain.

The third goal of Specific Objective 2 was *to identify the presence of putative markers conferring the potential to survive/multiply in the food chain and/or cause disease in humans (e.g. virulence and antimicrobial resistance)*. In recent years, numerous *L. monocytogenes* virulence factors have been suggested. We analysed the genomes for the presence/absence of a large set of putative virulence genes and compared the representation of these in clinical and food chain isolates as well as the representation according to phylogeny, i.e. lineage, clonal complex and sequence type (Section 9). Although the majority of *L. monocytogenes* isolates are generally susceptible to antimicrobials, a small proportion are found to demonstrate resistance to certain clinically relevant antimicrobials and the antimicrobial resistance determinants have been described genetically. Similar to the virulence genes,

the presence of antimicrobial resistance genes was searched in the LISEQ collection. Some studies have suggested that *L. monocytogenes* strains that are able to persist in the food production environment are genetically distinct from “transient strains” that do not have this capability. Recently, a number of genes potentially involved in persistence have been suggested. To explore this hypothesis, we first tested the ability of WGS to differentiate potential persistent strains from other strains collected in a cheese plant. Secondly, we compared the presence/absence of potential “persistence genes” in strains isolated from food processing environment (potentially persistent strains) to those in strains isolated in raw product (potentially non persistent, or transient strains) (Section 9). Along with the study of putative markers involved for virulence and persistence, investigation of potential host specific markers was carried out (Section 9).

The methods, results and conclusions of each subject and specific objective are described in detail in the Sections 5-9. An overall conclusion is given in Section 10.

4. Sequencing and Phylogentic Analysis

This section details the methodology undertaken to produce high quality whole genome sequencing data from the *Listeria monocytogenes* isolates selected in this study. To achieve the goals of specific objective 2 we then applied a range of microbial typing methods to the sequence data providing the framework to answer questions on genetic diversity and epidemiological relationships.

4.1. Methods

4.1.1. DNA extraction

DNA extraction of all isolates was performed at PHE using a pre-lysis procedure optimised for *L. monocytogenes* followed by automated DNA extraction. In brief, bacterial growth is harvested into a 96 deep well processing plate and treated with lysozyme at 37°C for 1 h followed by Proteinase K overnight at 56°C with gentle shaking. Lysates are then heated to 95-100°C for 10 minutes to ensure any unlysed organisms are killed and enzymes are destroyed. Samples are then treated with Ribonuclease A for 15 minutes at 37°C, centrifuged and the supernatants transferred to an automated nucleic acid extraction platform, presently the QiaSymphony. The yield and purity of extracted DNA is assessed using the Life Technologies® Quant-iT™ high sensitivity 96-well assay and the GloMax® Multi+ Detection and LabChip® DX Systems. DNA is diluted to 10-30ng/µl and submitted for whole genome sequencing to the PHE Genomic Development and Services Unit.

The Pasteur Institute provided DNA from clinical and outbreak isolates for sequencing at PHE. The extracted DNA complied with PHE's quality criteria (10-30ng/µl and OD_{260/280} = 1.8-2.0).

4.1.2. DNA sequencing and validation

Paired-end libraries were generated using the Illumina Nextera XT sample preparation kit. Assessment of fragment sizes was performed on the Perkin Elmer Labchip GX after fragmentation and clean-up. After normalisation, samples were pooled by hand and library quantification was performed using the KAPA library quantification kit for Illumina sequencing, on an ABI Vii7. Paired-end sequencing was performed on the Illumina HiSeq 2500 instrument using the TruSeq Rapid SBS kit (200 cycle) and TruSeq Paired-end rapid cluster kit. The following cycle parameters were used for sequencing: Read 1: 101, Index read 1: 8, Index read 2: 8 and Read 2: 101. RTA version 1.17.21.3 was used for generation of base call files.

FASTQ creation and de-multiplexing via CASAVA was performed on a dedicated high performance cluster (HPC). FASTQ reads were quality trimmed using Trimomatic (Bolger et al., 2014) with bases removed from the trailing end that fall below a PHRED score of 30. If the read length post trimming was less than 50 the read and its pair were discarded. If the post trimmed yield was less than 150 megabases the sample was discarded. A kmer (a short string of DNA of length k) based approach

was used (<https://github.com/phe-bioinformatics/kmerid>) to confirm the identity of the sample and to ensure the sequence was free from contamination. If any non-*Listeria* kmers were identified in the FASTQ reads the sample was discarded.

4.1.3. Assembly and annotation

Short reads were assembled using appropriate *de novo* assembly tools (e.g. SPADES). Spades assembly (version 3.5.0) run with Kmer 21,33,55,77,83, and the only-assembler option (Bankevich et al., 2012).

Assembled genomes were annotated in terms of protein coding features and RNA features. Prokka software (version 1.11) (Seemann, 2014) was used to annotate all the isolates. Preselected kingdom Bacteria and genus *Listeria* was performed to insure the accuracy of the annotation.

4.1.4. Gene by gene based typing

Gene by gene, or Allelic based methods have long been used to characterise microbial organisms and provide the opportunity for a common nomenclature based on the presence of specific allele types. The correlation of three different gene based typing methodologies with the underlying phylogeny elucidated above was carried out.

Multi-locus Sequence Typing (MLST)

The international MLST database for *L. monocytogenes* is maintained by Institut Pasteur (<http://bigsd.b.pasteur.fr/listeria/listeria.html>). The database holds 2,944 isolates representing 739 STs (as of 24 Apr 2015) and represents a globally representative collection from some 70 countries. More than 80 different sample sources are included of which more than 900 isolates are human-associated with the remainder from food, animal and environmental sources. The population genetic structure is characterised by major clonal groups. The top five clonal groups comprise one third of the described STs.

The MLST sequence type as defined by the Pasteur Scheme (Ragon et al., 2008) was extracted from each sequences using MOST (<https://github.com/phe-bioinformatics/MOST>) (Tewolde et al., 2016) and assigned a clonal complex in accordance with the Institut Pasteur international MLST database for *L. monocytogenes* (<http://bigsd.b.pasteur.fr/listeria/listeria.html>) designation.

Core genome MLST (cgMLST)

A cgMLST scheme extends the concept of MLST to the core loci present in majority of genomes from a representative collection of isolates from a species/family. We employed a recently developed Lm cgMLST method by Moura and colleagues (Moura et al., 2016) to assign an allelic designation to the 1,748 loci in the scheme for each of the genomes sequenced. The scheme was implemented using BIGSdb (Jolley and Maiden, 2010).

Ribosomal MLST (rMLST)

Ribosomal Multilocus Sequence Typing (rMLST) is an approach that indexes variation of the 53 genes encoding the bacterial ribosome protein subunits (*rps* genes) as a means of integrating microbial taxonomy and typing. The rMLST allelic variants were extracted as a subset of the cgMLST set based on the loci names, resulting in a total set of 30 ribosomal genes (Annex A).

4.1.5. Phylogenetic analysis

Phylogenetic methods exploiting nucleotide resolution variation (Single Nucleotide Polymorphisms (SNPs)) between bacterial isolates can be used to elucidate the relatedness and ancestry of strains under robust evolutionary models and provide a framework to explore the genetic diversity. SNPs inference and analysis was performed using software developed at PHE: SnapperDB (<https://github.com/phe-bioinformatics/snapperdb>), with each *L.monocytogenes* clonal complex having a separate database instance. In summary short read sequences from strains selected in this

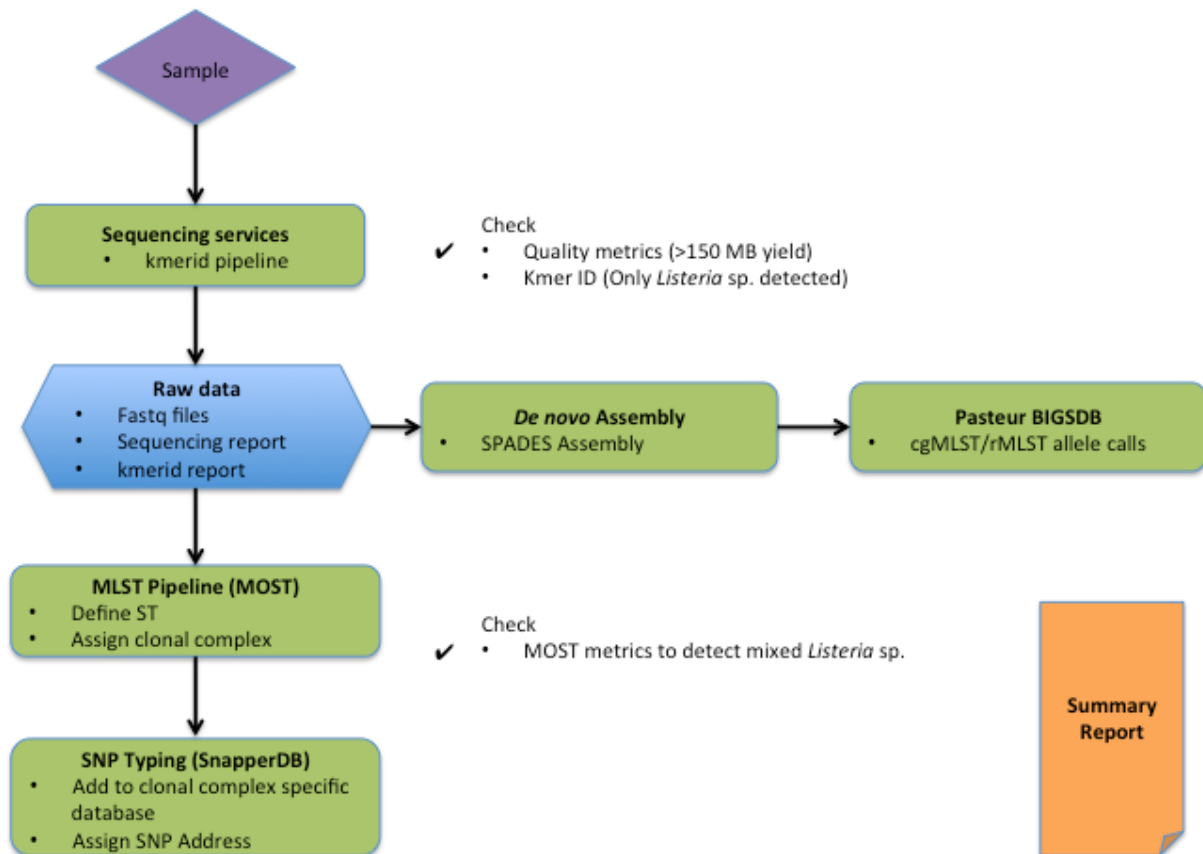
study were mapped against an appropriate reference genome of *L. monocytogenes* (see Table 4.1) using BWA-MEM (Li and Durbin, 2010). The resultant sequence alignment maps (SAMs) are processed and high quality variants (MQ>30, AD>0.9, DP>10) extracted using GATK2 (McKenna et al., 2010). Variants and uncertain positions are stored for further analysis. Maximum likelihood phylogenies were produced using RAxML-8.17 (Stamatakis, 2014) using the GTRGAMMA model (<https://github.com/lguillier/LISEQ-codes#chapter-4>).

Recombination allows for rapid introduction of new genetic material between strains and such evolutionary events can impact on the ancestral inference provided by phylogenetic methods. Gubbins (Croucher et al., 2014) was used to identify recombinant regions of the genome within clonal groups. These positions can then be filtered as appropriate when inferring the ancestry of strains.

These robust phylogenetic representations of the population structure of strains provided the framework to assess the diversity of *L. monocytogenes* within and between the different sources and human origin at the lineage, clonal complex and strain level resolution.

Table 4.1.: Complete genome sequences that were used as reference genomes for each corresponding clonal complex

Clonal Complex	<i>L. monocytogenes</i> strain	Genbank Accession
CC1	F2365	AE017262
CC2	J1-220	CP006047
CC3	R2-502	CP006594
CC4	Clip80459	FM242711
CC6	J1816	CP006046
CC7	10403S	CP002002
CC8	88-0478	CP006862
CC9	SLCC2479	FR733649
CC5	J2-064	CP006592
CC11	J0161	CP002001
CC14	SLCC7179	FR733650
CC69	HCC23	CP001175
CC121	6179	HG813249
CC131	SLCC2376	FR733651
CC155	1998	CP002004
CC361	J2-031	CP006593



Samples are sequenced and checked if they are contaminant free and have a FASTQ yield greater than 150 MB. The FASTQ reads are assembled and submitted to the Pasteur BIGSDB instance where cgMLST and rMLST allele calls assigned. The 7 locus MLST is extracted using MOST and if unambiguous the reads are mapped to the appropriate clonal complex specific reference genome and a SNP address assigned. A summary report containing the metrics and results of this process is produced. Samples are sequenced and checked they are contaminant free and have a FASTQ yield greater than 150 MB.

Figure 4.1: Schematic diagram of sequence analysis pipeline

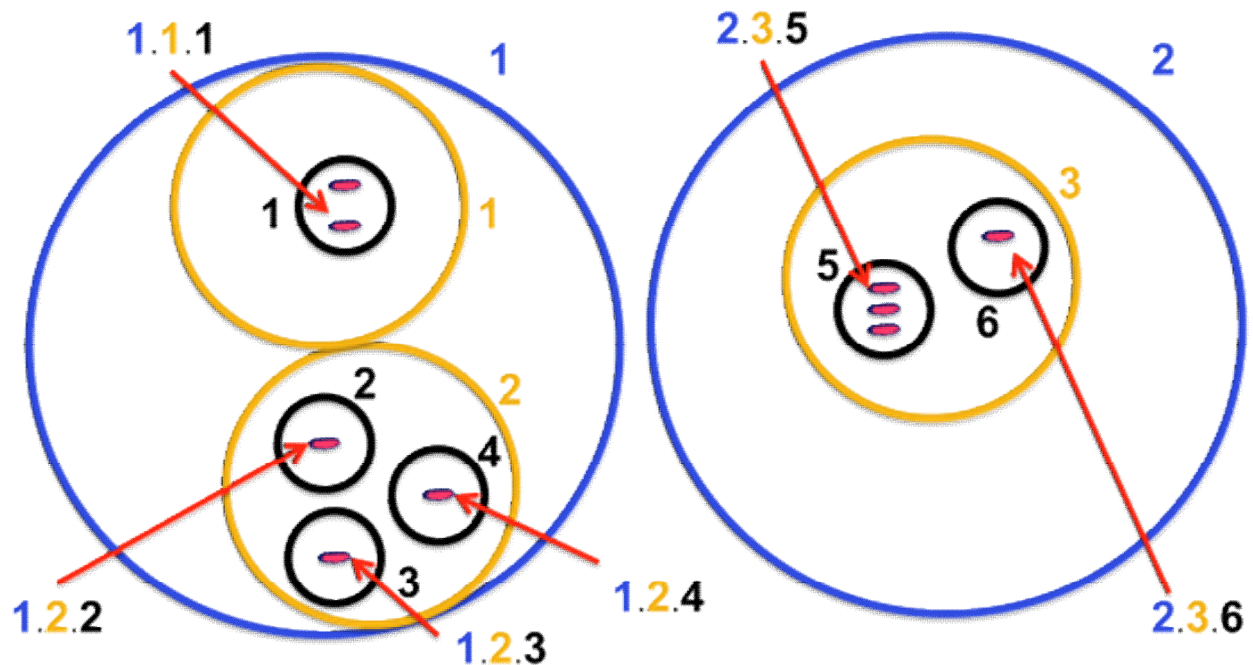
4.1.6. Genetic distance: SNP address

As new strains are added to SnapperDB they are compared to the database and a distance matrix maintained of all pairwise SNP distances. The distance represents those positions that differ between a pair of isolates with respect to the reference genome. Single linkage clustering of genetic distance is an effective method of describing phylogenetic groups as it is inclusive of clonal expansion events. Using hierarchical single linkage clustering of the pairwise SNP distances we are able to define a nomenclature that corresponds to the tree structure from deep branches through to clades through to identical strains. This enables an isolate level nomenclature to be derived for each genome sequence allowing efficient searching of the population studied as well as automated cluster detection. Single linkage clustering was performed at 7 descending SNP distance thresholds ($\Delta 250$, $\Delta 100$, $\Delta 50$, $\Delta 25$, $\Delta 10$, $\Delta 5$, $\Delta 0$) which generates a 7 digit address where each number represents a unique cluster at that threshold.

4.1.7. Genetic distance: SNP address

As new strains are added to SnapperDB they are compared to the database and a distance matrix maintained of all pairwise SNP distances. The distance represents those positions that differ between a pair of isolates with respect to the reference genome. Single linkage clustering of genetic distance is an effective method of describing phylogenetic groups as it is inclusive of clonal expansion events.

Using hierarchical single linkage clustering of the pairwise SNP distances we are able to define a nomenclature that corresponds to the tree structure from deep branches through to clades through to identical strains. This enables an isolate level nomenclature to be derived for each genome sequence allowing efficient searching of the population studied as well as automated cluster detection. Single linkage clustering was performed at 7 descending SNP distance thresholds ($\Delta 250$, $\Delta 100$, $\Delta 50$, $\Delta 25$, $\Delta 10$, $\Delta 5$, $\Delta 0$) which generates a 7 digit address where each number represents a unique cluster at that threshold.



The red rods represent bacterial genomes and the distance between them represents genetic distance. In this example we have a three levels of clustering for similarity, the outer blue rings corresponds to 10 SNP differences, the inner yellow rings corresponds to 5 SNP differences and the inner black rings represent 0 SNP differences. If two samples are identical they have the same SNP address, i.e. two isolates that have the address 1.1.1 have 0 SNP differences between them. The isolates with the addresses 1.2.2 and 1.2.3 have a different final digit indicating that they are not identical but both have matching 1st and 2nd digits so are within the same 10 SNP cluster and the same 5 SNP cluster. When we compare the isolate 1.2.2 against 2.3.5, as the first digit is different, they are different by greater than 10 SNPs.

Figure 4.2.: Schematic diagram representing hierarchical clustering

4.2. Results

4.2.1. Processing of isolates

PHE received 1,048 *L. monocytogenes* isolates from the ANSES, the SSI, University of Aberdeen and PHE. In addition, 108 isolates were previously sequenced (Italian cheese and cheese factories, n= 100 and UK outbreak, n=8). All isolates received by PHE were sequenced as described. Sequence data for 13 isolates were not accepted for further analysis due to poor quality or unresolvable strain contamination.

The final dataset is composed of 1,143 sequences with a high quality of sequencing.

4.2.2. Gene by gene based typing: MLST and clonal complex assignment

From 1,143 isolates sequenced, 42 different CC and 13 singleton STs (unassigned CC) were identified. One isolate could not be assigned to any ST or CC. 10 clonal complexes accounted for 70% of the samples (Table 4.2, Figures 4.3 and 4.4). The population structure of the isolates in the study are further described as a minimum spanning tree (Figure 4.5).

Table 4.2.: The clonal complexes identified and the number of isolates by isolation context as defined in Section 2 and listed in the strain selection appendices

Clonal Complex	Lineage	RTE food Appendix 1-3	Food chain processing Appendix 4	Clinical, sporadic Appendix 5	Outbreak-related Appendix 6	Total
CC121	II	144	37	6	0	187
CC9	II	81	15	14	0	110
CC8	II	69	5	24	0	98
CC1	I	10	4	50	8	72
CC2	I	19	29	20	0	68
CC101	II	10	41	16	0	67
CC6	I	30	3	28	0	61
CC155	II	32	1	8	13	54
CC7	II	16	4	16	8	44
CC14	II	13	2	9	13	37
CC4	I	1	1	10	24	36
CC87	I	10	0	4	19	33
CC31	II	24	7	1	0	32
CC3	I	18	7	6	0	31
CC37	II	9	15	5	0	29
CC204	II	17	3	1	0	21
CC59	I	10	0	4	4	18
CC5	I	7	6	4	0	17
CC21	II	13	0	2	0	15
CC20	II	8	2	2	0	12
CC415	II	0	2	0	9	11
CC18	II	0	6	4	0	10
Minor CCs	LI=31 LII=48 L.innocua=1	35	10	28	7	80
Total		576	200	262	105	1,143

Note: Minor CCs included CC398, CC11, CC193, CC224, CC403, CC54, CC177, CC19, CC220, CC29, CC77, CC217, CC26, CC379, CC207, CC218, CC388, CC475, CC88, CC89, ST184, ST200, ST32, ST382, ST392, ST560, ST570, ST602, ST736, ST773, ST839 (ordered according to occurrence).

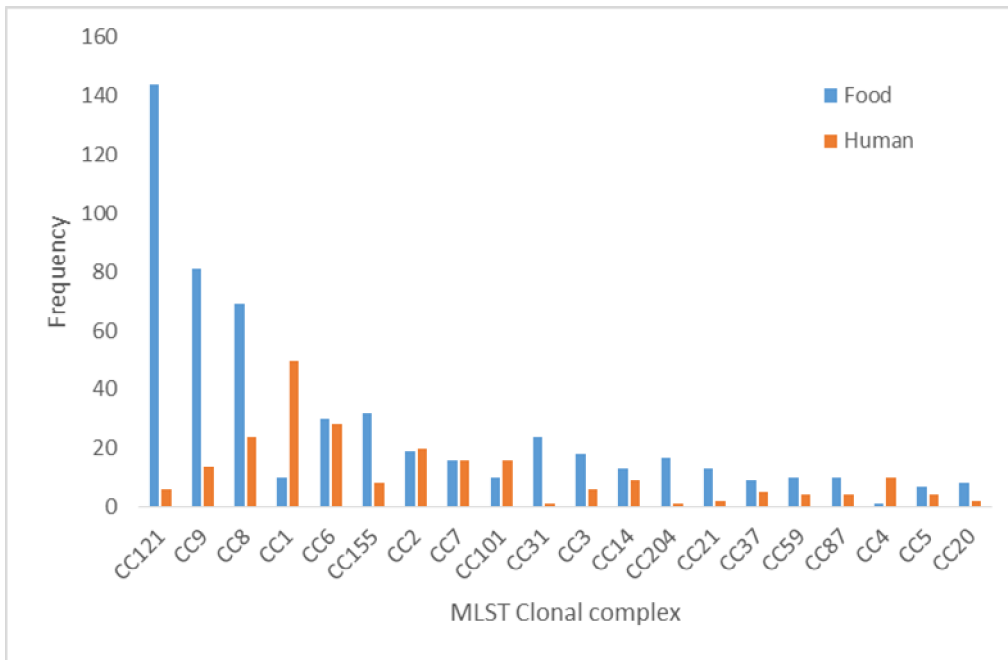


Figure 4.3.: Distribution of clonal complexes in ready-to-eat food (isolates in Appendices 1, 2 and 3) and from sporadic human clinical infections (isolates in Appendix 5)

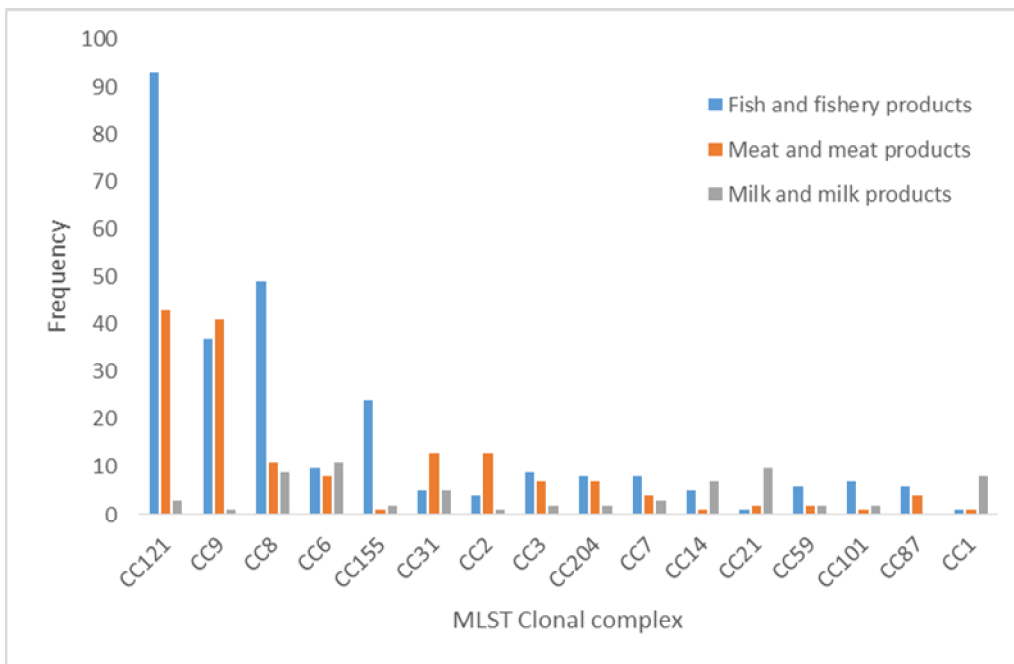
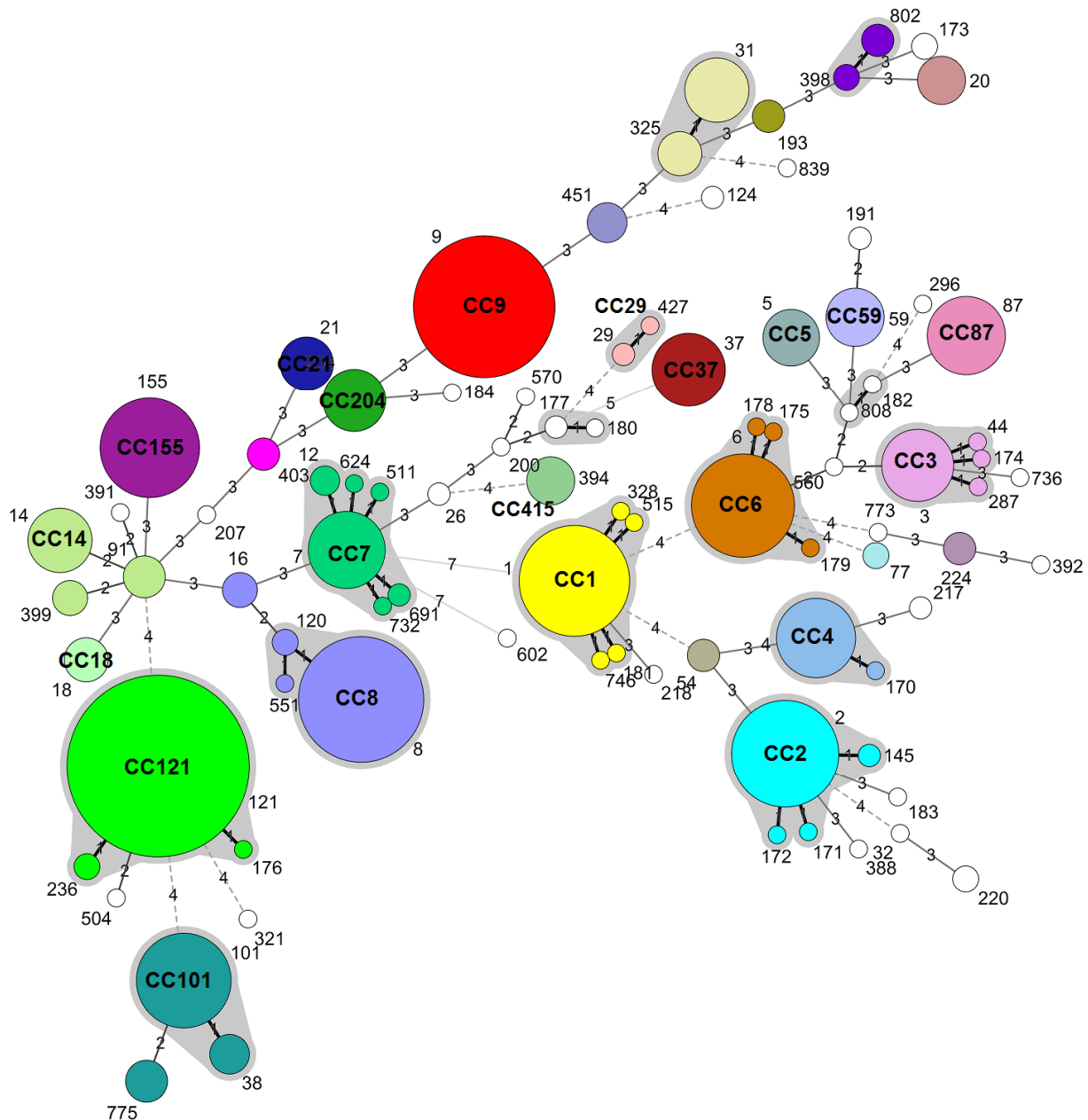


Figure 4.4.: Distribution of clonal complexes from the three major food product categories (isolates in Appendices 1-3)



Each circle represents a single sequence type (ST) that is numbered on the tree. Clonal complexes (CC) defined by single locus variants are shaded in grey. The number of loci that differ between STs is labelled on the branches.

Figure 4.5.: Minimum spanning tree of the isolates included in this study as described by 7 locus MLST

4.2.3. Developing the Framework for Phylogenetic Analysis

L. monocytogenes contains 4 divergent lineages with greater than 100,000 variants across the population. This study contained isolates from lineage I & II and the population structure based on whole genome SNPs is displayed in the phylogenetic tree in Figure 4.6. From the phylogeny it can be seen that there is a clear delineation between lineages and clonal complex's within lineages.

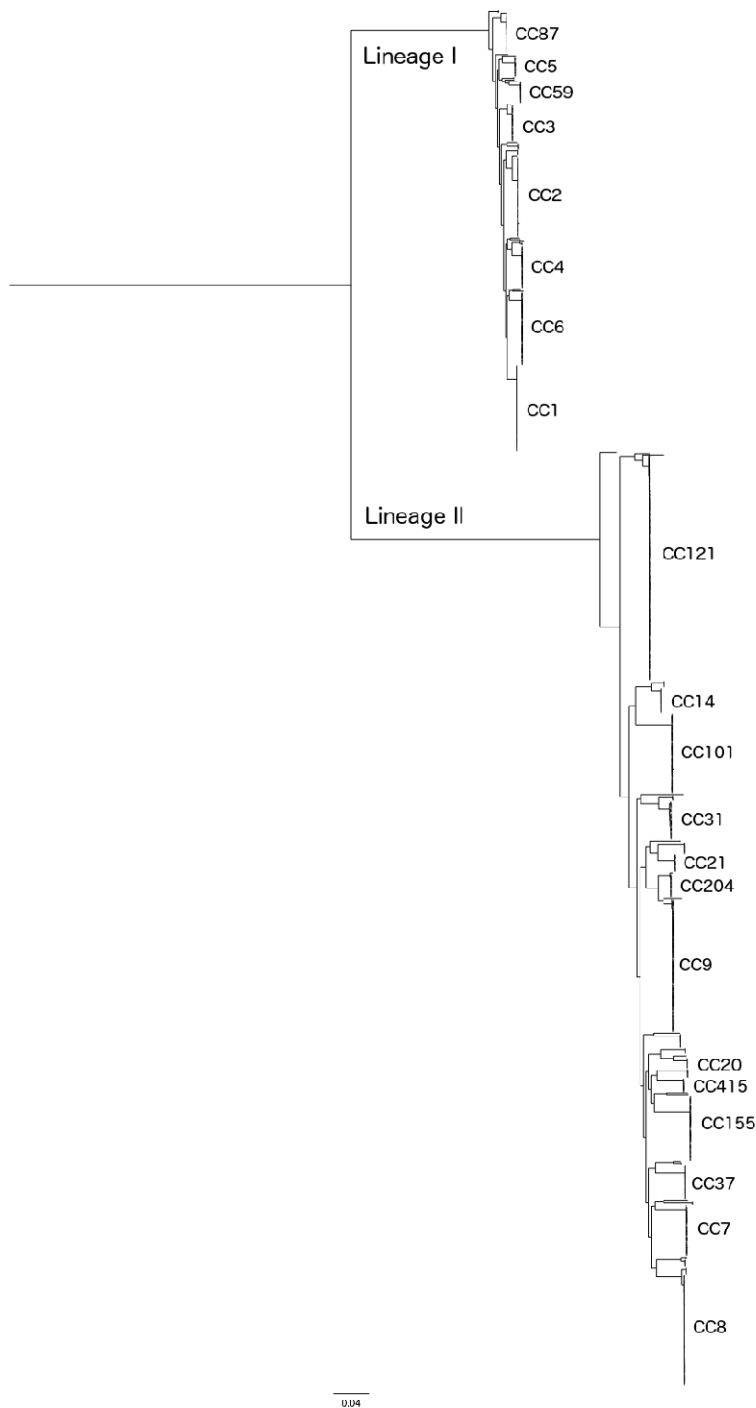


Figure 4.6.: Whole genome SNP maximum likelihood phylogeny of *L. monocytogenes* genome sequences with the clades annotated by 7 loci MLST Clonal Complex (CC) generated using Parsnp (<https://github.com/marbl/parsnp>) with EGE-e (NC_003210.1) as the reference genome

In order to elucidate the fine phylogenetic relations between isolates and to develop strain level nomenclature for further analysis, strains were further analysed within each clonal complex using the clonal complex specific reference mapping approach described in 4.1.5. Isolates not belonging to a clonal complex were not assigned the SNP address nomenclature.

4.2.4. Phylogenetic analysis of major clonal complexes

Clonal Complex 121

This clonal complex has the most number of isolates from this study with a 187 in total, with 165 food product isolates, 16 from the food-processing environment and 6 human clinical.

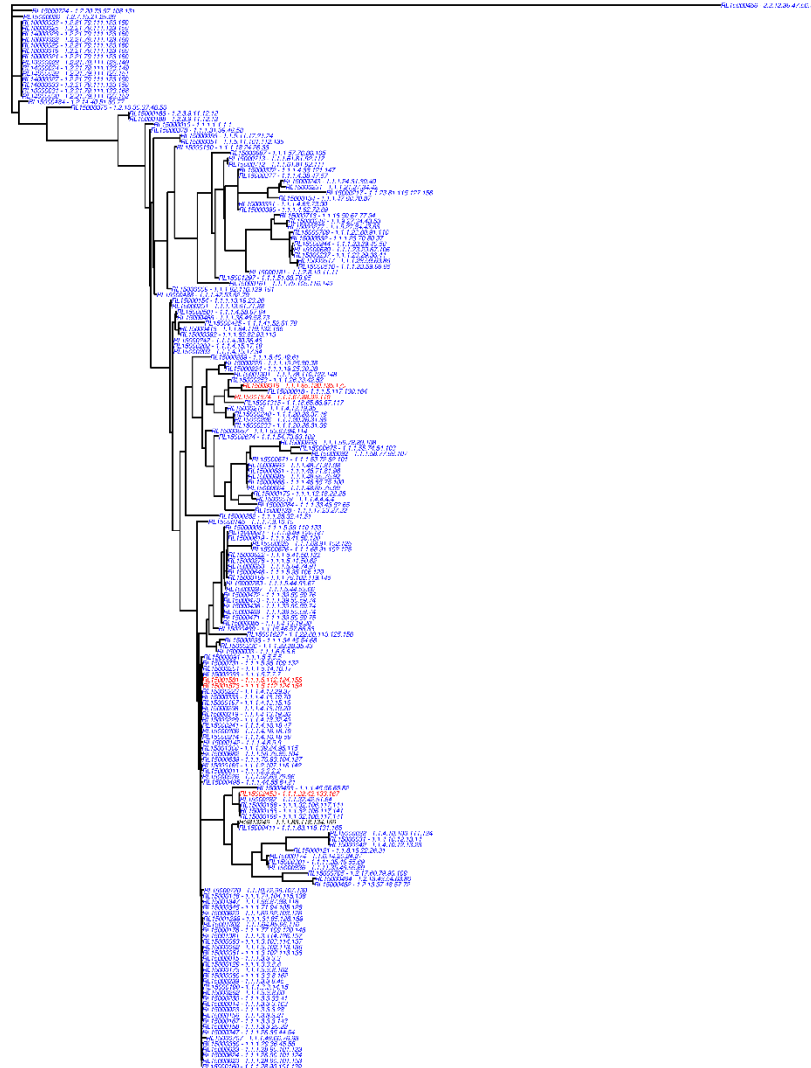


Figure 4.7.: Whole genome maximum likelihood phylogeny constructed using RAxML with the GTRGAMMA model of Clonal Complex 121. Clinical cases are coloured red and non-clinical isolates coloured blue. Taxa are labelled by strain identity and SNP address. High-resolution figure is available in full size at <https://github.com/lguillier/LISEQ-codes/blob/master/Chapter4/Chap4-5Trees.7z>

Clonal Complex 9

Clonal complex 9 contains 110 isolates, 13 from food processing environment, 83 from food products, 14 from human clinical.

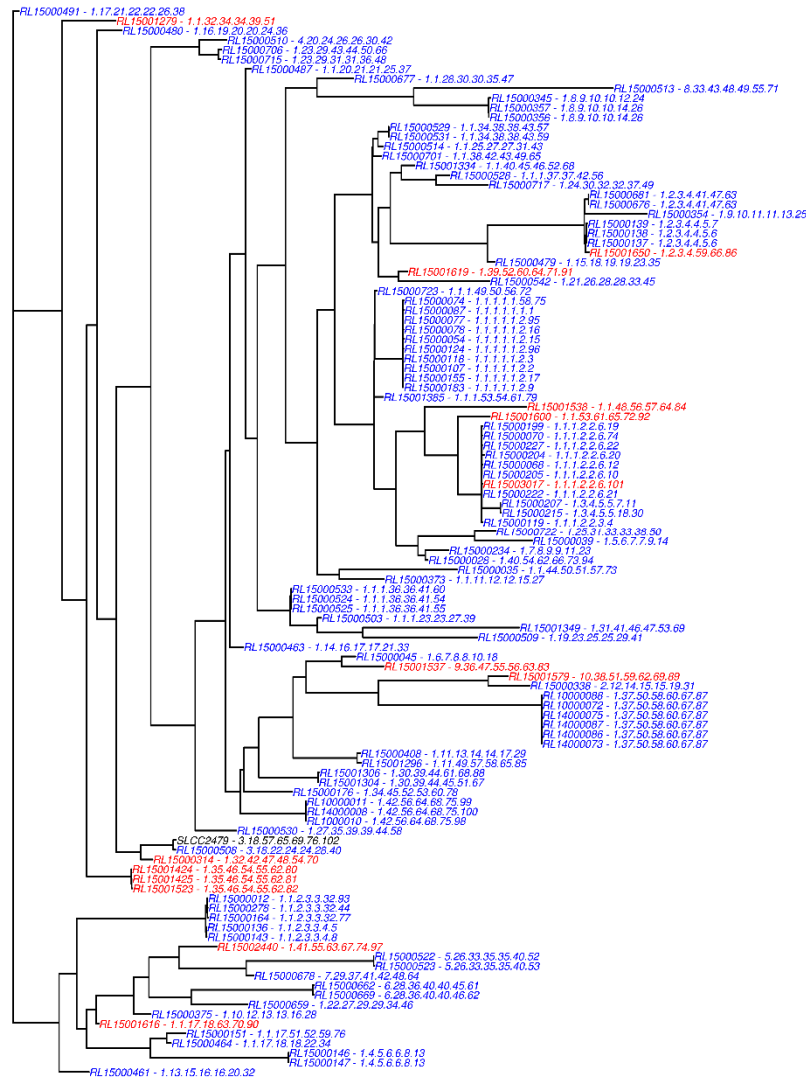


Figure 4.8.: Whole genome maximum likelihood phylogeny constructed using RAxML with the GTRGAMMA model of Clonal Complex 9. Clinical cases are coloured red and non-clinical isolates coloured blue. Taxa are labelled by strain identity and SNP address. High-resolution figure is available in full size at <https://github.com/lguillier/LISEQ-codes/blob/master/Chapter4/Chap4-5Trees.7z>

Clonal Complex 8

The second most populated cluster is clonal complex 8. This group contains 98 isolates, 4 from food processing environment, 70 from food products and 24 from human clinical.

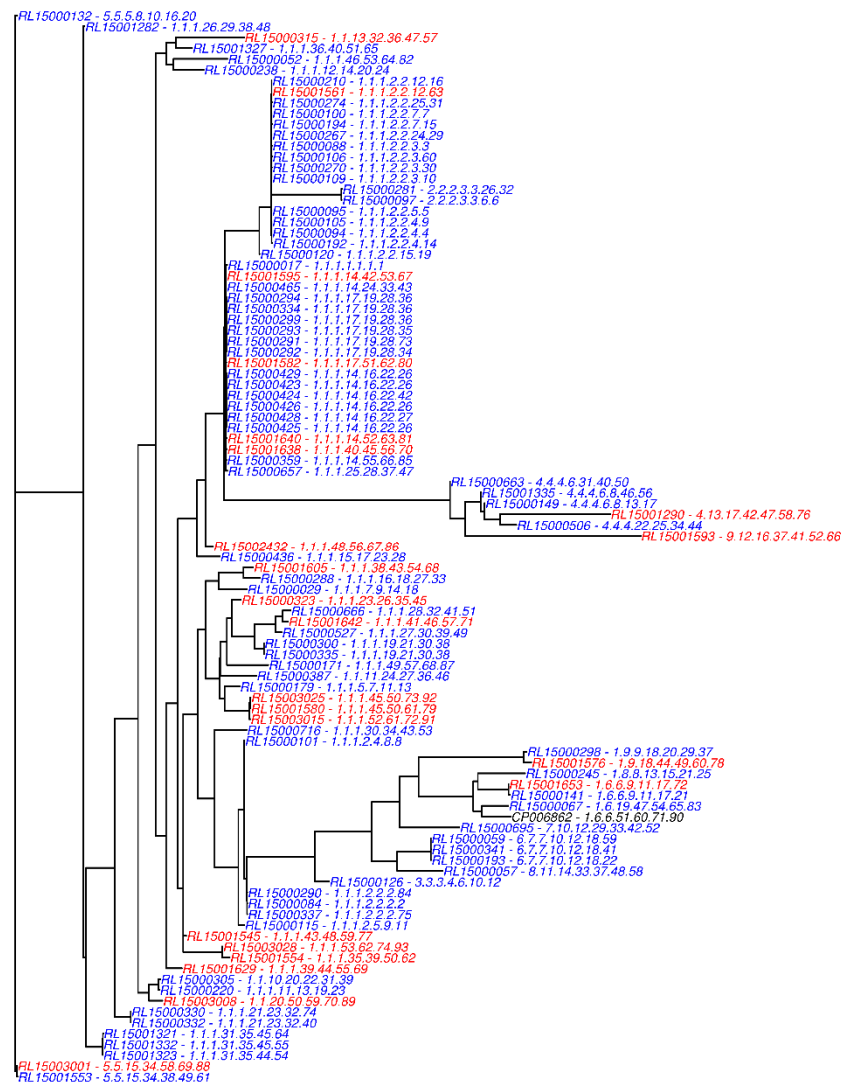


Figure 4.9.: Whole genome maximum likelihood phylogeny constructed using RAxML with the GTRGAMMA model of Clonal Complex 8. Clinical cases are coloured red and non-clinical isolates coloured blue. Taxa are labelled by strain identity and SNP address. High-resolution figure is available in full size at <https://github.com/Iguillier/LISEQ-codes/blob/master/Chapter4/Chap4-5Trees.7z>

Clonal Complex 101

Clonal complex 101 contains 67 strains, 35 are from food processing environment, 16 from food products and 16 from human clinical. There is currently no complete genome from the CC101 complex and therefore the isolate with the best assembly was used as a reference genome.

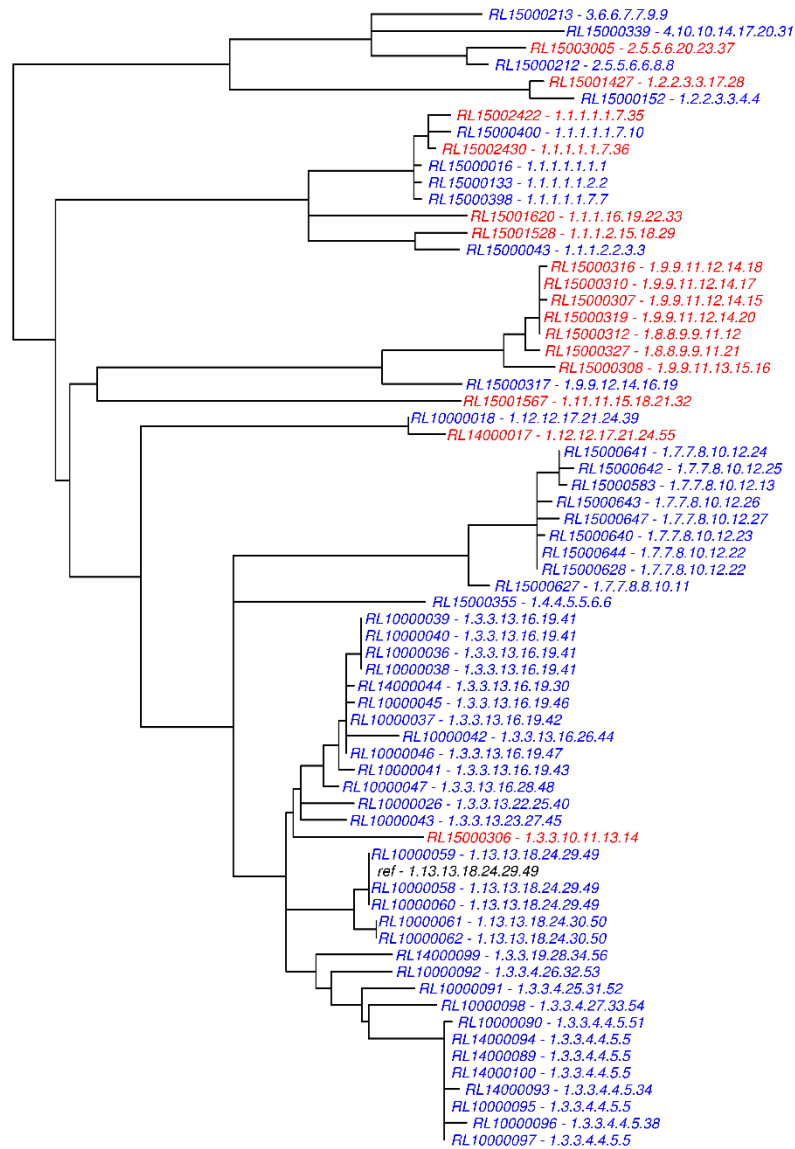


Figure 4.10.: Whole genome maximum likelihood phylogeny constructed using RAXML with the GTRGAMMA model of Clonal Complex 101. Clinical cases are coloured red and non-clinical isolates coloured blue. Taxa are labelled by strain identity and SNP address. High-resolution figure is available available in full size at <https://github.com/lguillier/LISEQ-codes/blob/master/Chapter4/Chap4-5Trees.7z>

Clonal Complex 4

This clonal complex of 36 isolates is mainly composed of human clinical isolates (n=34) and 2 food products. It has been shown as a hypervirulent cluster by a recent publication (Maury et al., 2016).

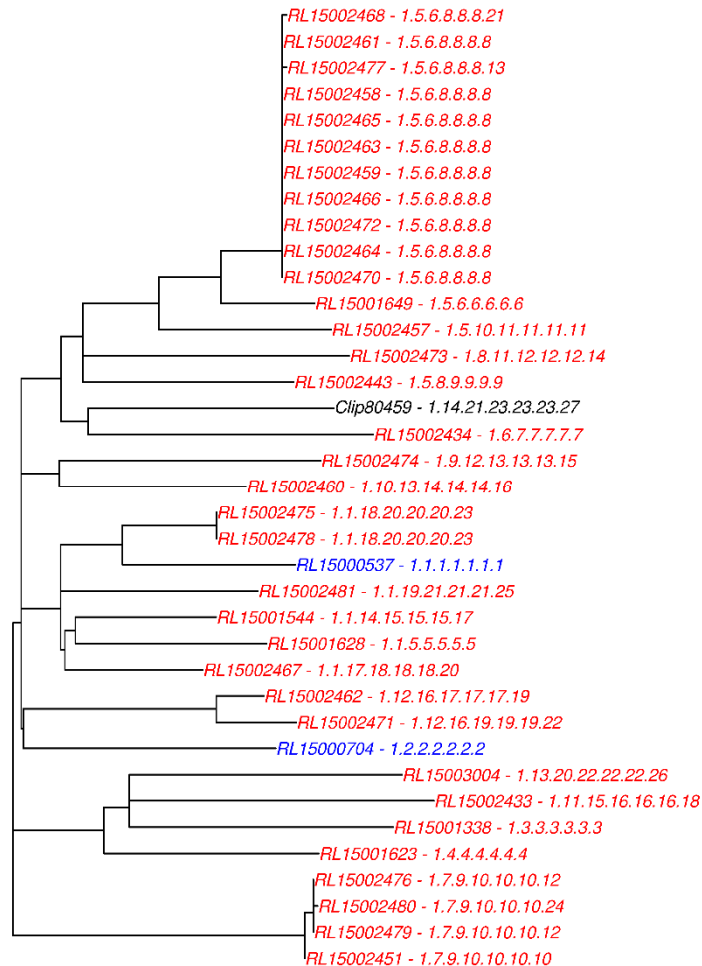


Figure 4.11.: Whole genome maximum likelihood phylogeny constructed using RAxML with the GTRGAMMA model of Clonal Complex 4. Clinical cases are coloured red and non-clinical isolates coloured blue. Taxa are labelled by strain identity and SNP address. High-resolution figure is available in full size at <https://github.com/lguillier/LISEQ-codes/blob/master/Chapter4/Chap4-5Trees.7z>

4.3. Conclusion

The phylogenetic analysis on four large CCs (CC8, CC9, CC101 and C121) reveals that within a CC, clinical strains are not associated to a specific clade of the tree. The CC4 phylogeny confirms recent results that this CC is associated to clinical strains and is probably one of the most virulent (Maury et al., 2016).

Whole genome sequencing has allowed us to define the population of *Listeria monocytogenes* from this study to an unprecedented resolution. It has provided the framework to answer questions on genetic

diversity in the different sources assayed in this strain collection as well as to explore possible epidemiological links between isolates.

5. Retrospective analysis of outbreaks

Many of the traditional typing methods have inherent problems; the discriminatory power of phenotypic methods is too low (serotyping) or suffer from biologic variability (phage typing) (Graves et al., 2007). Molecular methods such as PFGE have a much better discriminatory power, but are laborious and require a lot of work for creating comparable results between different laboratories. The most used *L. monocytogenes* PFGE protocol is created in the PulseNet organisation (Graves and Swaminathan, 2001), and the protocol specifies up to three different enzymes in order to get sufficient discriminatory power for secure outbreak detection meaning that the method is both expensive, time consuming and laborious (Tourdjman et al., 2014).

The outbreaks in this LISEQ project were defined by each submitter based on epidemiology and their local choice of molecular typing (e.g., PFGE and fAFLP).

WGS gives us the possibility to analyse the genome at close to the endpoint resolution of DNA typing, and the new challenge is no longer to chase the maximum resolution, but to find out what the actual variation within a listeriosis outbreak is. Besides the true biological variation, there will be variations in the results dependent on both lab work and analysis approach.

As described in Section 4, there are two main approaches to analyse the relationship between genomic sequences (SNP and gene by gene). Both approaches are in active use in reference laboratories in Europe for outbreak detection and/or investigations. Therefore, we chose to use and compare the two analysis approaches in order to properly address the specific objective 3 - “to perform a retrospective analysis of outbreak strains (i.e. using a subset of epidemiologically linked human and food isolates) to investigate the suitability of WGS as a tool in outbreak investigations”.

All of the outbreaks were national outbreaks without a previously known international component.

The results from this section were fed back to other parts of the project where epidemiological relationships are reported.

5.1. Methods

In eight of the nine analysed outbreaks, isolates from both human cases and linked food isolates were provided. The sequences from each outbreak were analysed together with all other isolates of the same clonal complex in the study regardless of epidemiological relationship to the outbreak. The diversity of isolates within outbreaks were explored using SNP and gene-by-gene based methods, thus validating the thresholds for cluster definitions.

The SNP analysis was made as described previously in Section 4, using ParSNP with a complete or high-coverage reference genome from the same clonal complex. The distance matrices shown are created with pairwise comparisons with pairwise deletion of ambiguous positions. The SNP analyses were visualized with maximum parsimony trees since this is a good, simple algorithm for relatively closely related isolates, resulting in a tree with branches of proportional lengths to the number of SNP. The maximum parsimony trees were created based on the non-ambiguous positions for all isolates in the clonal complex.

For the cgMLST the 1748 loci Pasteur scheme (Moura et al., 2016) was used and the analysis results were kindly provided by Alexandra Moura and Sylvain Brisse using an initial de novo assembly using CLC assembly cell from Qiagen Bioinformatics. cgMLST results were visualized using minimum spanning trees since this is a fast algorithm allowing short-term divergence and micro-evolution in populations to be reconstructed based upon sampled data.

Project isolates with the same CC, but with unknown epidemiological link to the outbreaks were included in the analysis to disclose possible additional cases and/or sources in same country as well as other EU countries.

5.2. Results

The outbreaks are presented individually in Sections 5.2.1 to 5.2.9. Each section contains an epidemiological description of the outbreak, with SNP- and cgMLST-based analysis consisting of a tree of the outbreak isolates + other isolates with the same CC. The accompanying matrices describe the pairwise number of differences between the cluster isolates only. The SNP trees are maximum parsimony trees and the cgMLST trees are minimum spanning trees.

The SNP and cgMLST analyses of the outbreak isolates (Table 5.1) show good concordance between the methods and generally the same level of SNP and allele differences were found. The exception is cluster #1 where the maximum number of allele differences is much higher than the number of different SNPs as described further in Section 5.2.1.

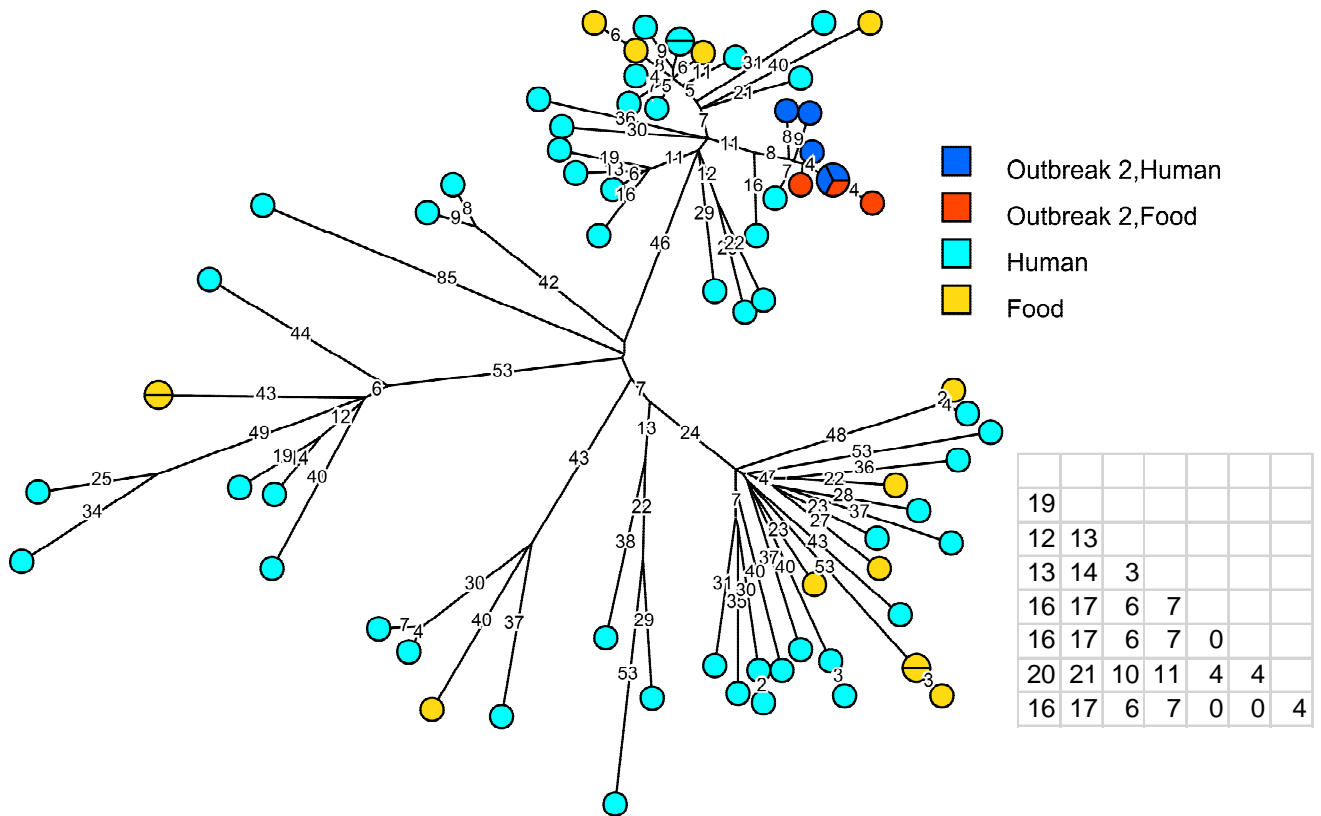
Table 5.1.: Descriptive data with CC, time period and pairwise genetic distances within each outbreak

Outbreak number		#1	#2	#3	#4	#5	#6	#7	#8	#9
	Clonal complex	CC155	CC1	CC7	CC59	CC415	CC398	CC87	ST14	CC4
	Years	2012-13	2007-13	2013-14	2009-11	2013-14	2013	2013-14	2012	2012
All isolates within CC	n human/food	13/41	55/17	24/20	6/12	9/3	8/1	17/16	6/16	34/2
SNP	Median	111	174	258.5	90	3	38	38	214	126
cgMLST	Median	56	71	59	46	4	19,5	20	23	54
SNP	Max	174	259	1368	243	93	65	74	337	183
cgMLST	Max	118	119	131	119	50	36	41	47	88
Outbreak isolates	n human/food	5/8	5/3	4/4	2/2	8/1	4/1	13/6	4/9	24/0
SNP	Median	0	10,5	2	9	3	0	5	2	0
cgMLST	Median	1	7,5	3	7	2	0	3	2	2
SNP	Mode	0	16	1	9	3	0	5	0	0
cgMLST	Mode	0	5	3	10	2	0	3	1	2
SNP	Max	2	21	4	12	4	1	8	8	2
cgMLST	Max	51	16	6	10	4	1	7	8	4

Note: In the rows describing SNP analysis the numbers refer to the number of SNP and in the rows describing cgMLST, the numbers indicate the number of allele differences. The table is grouped into two main parts, each headed by the tan coloured rows. The top part includes the outbreak isolates plus all other isolates in the LISEQ project belonging to the same CC. The bottom part concerns the outbreak isolates only. Each column corresponds to one outbreak and for each outbreak (or outbreak plus other isolates in the same CC), median and maximal pairwise distances are shown for both categories. For the outbreak only category the mode distance is also included. In outbreak 8 belonging to CC14 the numbers shown are for the ST instead of the CC as the defining unit since this CC is polyphyletic. In this table, the food category also includes some environmental isolates.

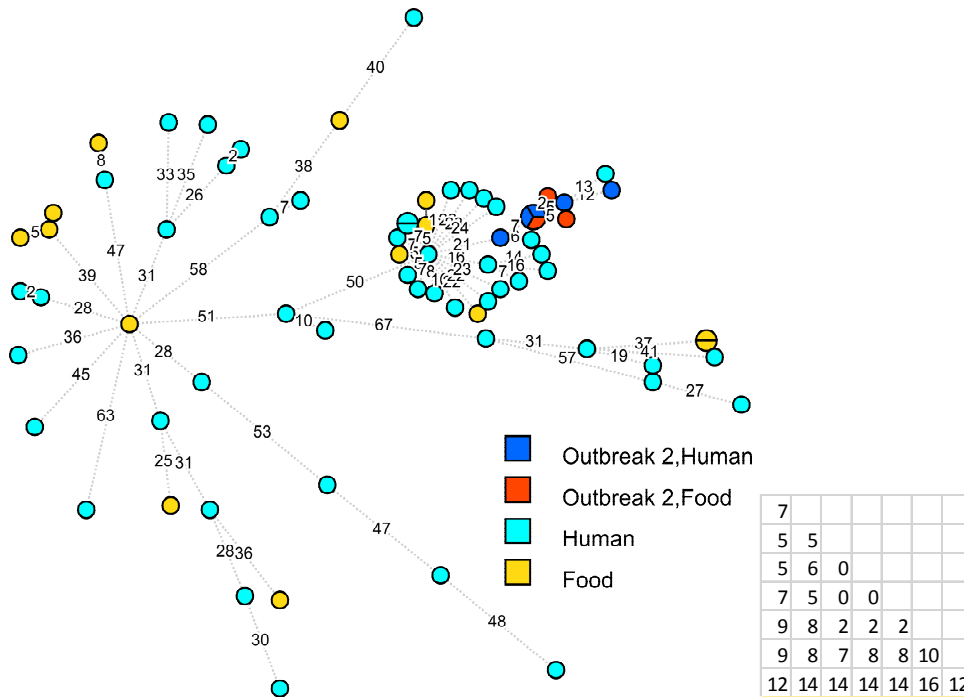
5.2.1. Outbreak 1 – CC155

An outbreak of 4 cases in the North West of country B within a 3-week period with the same rare molecular type (1/2a, XI.23). Three of the cases had a history of consuming pressed beef (a sliced meat product consisting of meat and other ingredients moulded and set in gelatine) purchased from different retailers but made by a single producer and a third had purchased raw meat from a butcher's shop also supplied with cooked meat products from the same producer. Isolates of the same



The tree is constructed using all isolates having the same CC as the outbreak isolates. To the right is shown a pairwise SNP distance matrix of the outbreak isolates. Number on branches indicates the number of SNP differences the branch corresponds to.

Figure 5.3.: SNP maximum parsimony tree of outbreak 2



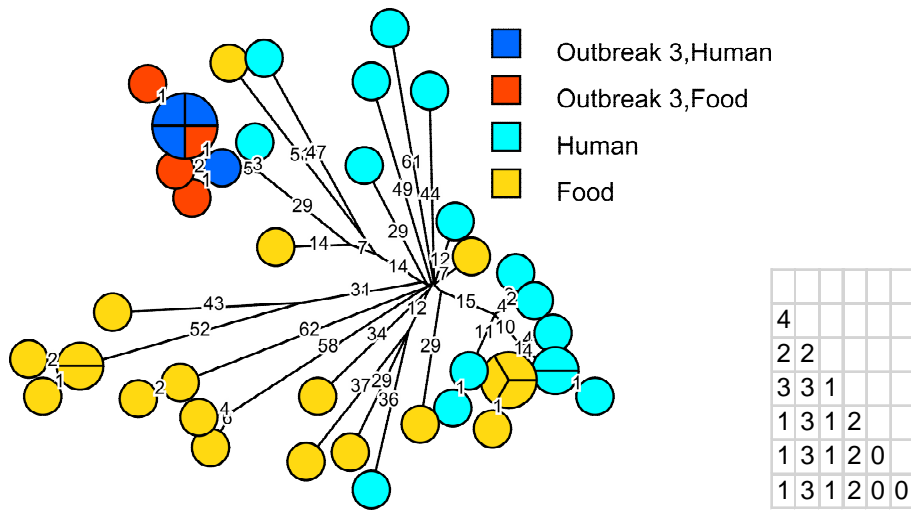
The tree is constructed using all isolates having the same CC as the outbreak isolates. To the right is shown a pairwise distance matrix of the outbreak isolates indicating the number of allele differences. Number on branches indicates the number of allele differences the branch corresponds to.

Figure 5.4.: cgMLST minimum spanning tree

5.2.3. Outbreak 3 – CC7

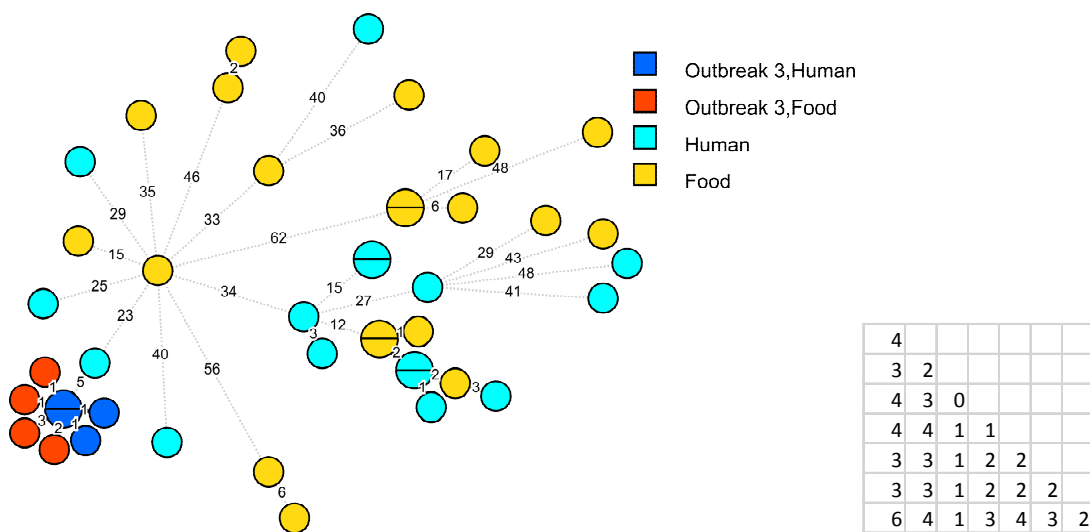
Four cases in a region of country B contracted listeriosis due to rare fAFLP type (1/2a, XIV.52b); one in May 2013 and 3 in Jan 2014. All cases were males between ages 54-67 years and all hospital inpatients. Review of PHE food and human fAFLP database detected isolates of same rare type in sandwiches at 3 hospitals in same geographical region as cases and from a local retailer. Only one case reported eating hospital sandwiches. Cases were at different hospitals and only one case was at hospital where the same type was found in sandwiches; this case did not report eating hospital sandwiches. Subsequently during the investigation by country B, WGS showed that isolates from 4 cases, and from sandwiches at 2 hospitals and a local retailer all grouped within same 5 SNP cluster and also identified a 5 case.

The WGS analysis shows that these isolates are very closely related, which matches the profile of a point source outbreak within a short time frame. With the exception of one isolate, the cluster is well separated from the other 37 CC7-isolates in the study. One other isolates in the collection clustered relatively close (8 SNP / 5 alleles) to the outbreak isolates. This isolate is a human sporadic case from the same country as the outbreak, but being isolated in 2010 it is not a temporal match.



The tree is constructed using all isolates having the same CC as the outbreak isolates. To the right is shown a pairwise SNP distance matrix of the outbreak isolates. Number on branches indicates the number of SNP differences the branch corresponds to.

Figure 5.5.: SNP maximum parsimony tree of outbreak 3

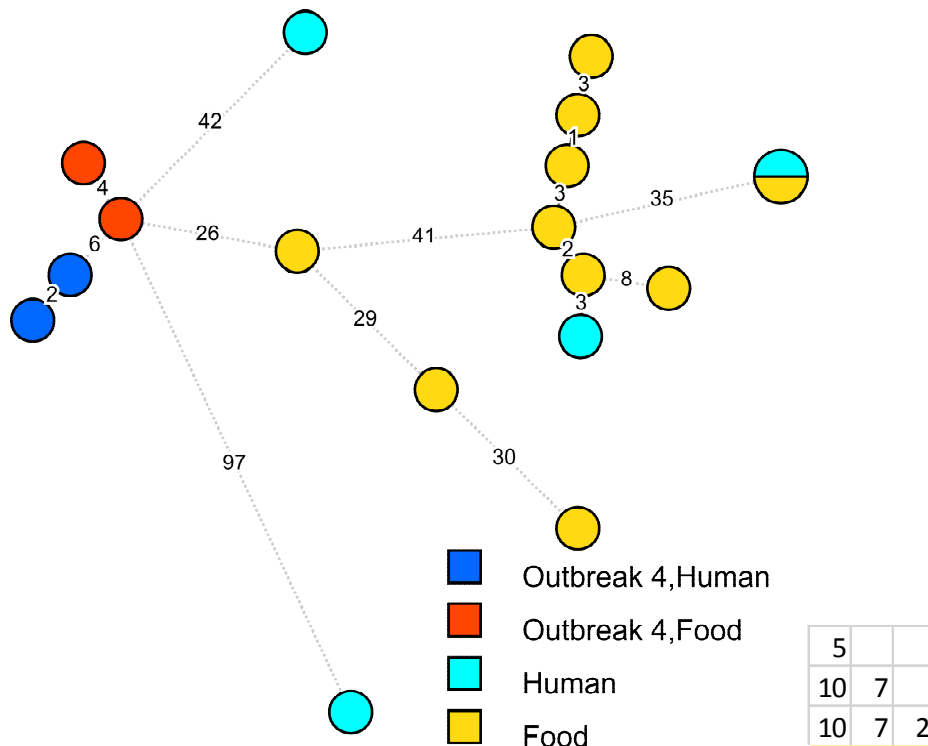


The tree is constructed using all isolates having the same CC as the outbreak isolates. To the right is shown a pairwise distance matrix of the outbreak isolates indicating the number of allele differences. Number on branches indicates the number of allele differences the branch corresponds to.

Figure 5.6.: cgMLST minimum spanning tree

5.2.4. Outbreak 4 – CC59

Ox tongue incident in country B that only involved a single confirmed case. The incident was published as a case of foodborne listeriosis linked to a contaminated food-production process (Lamden et al., 2013). The additional case used in this outbreak analysis is a tentatively linked case reported over 2 years prior to the published case.

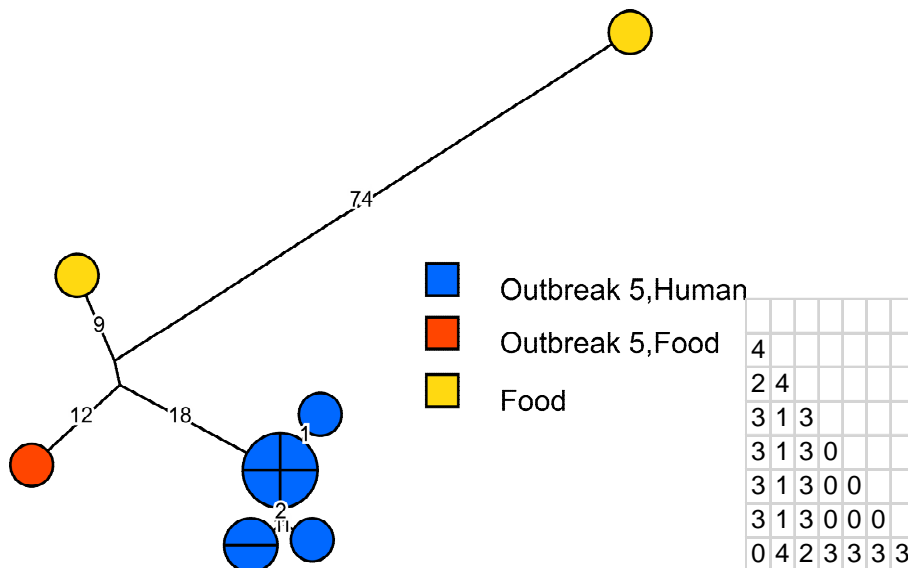


The tree is constructed using all isolates having the same CC as the outbreak isolates. To the right is shown a pairwise distance matrix of the outbreak isolates indicating the number of allele differences. Number on branches indicates the number of allele differences the branch corresponds to.

Figure 5.8.: cgMLST minimum spanning tree

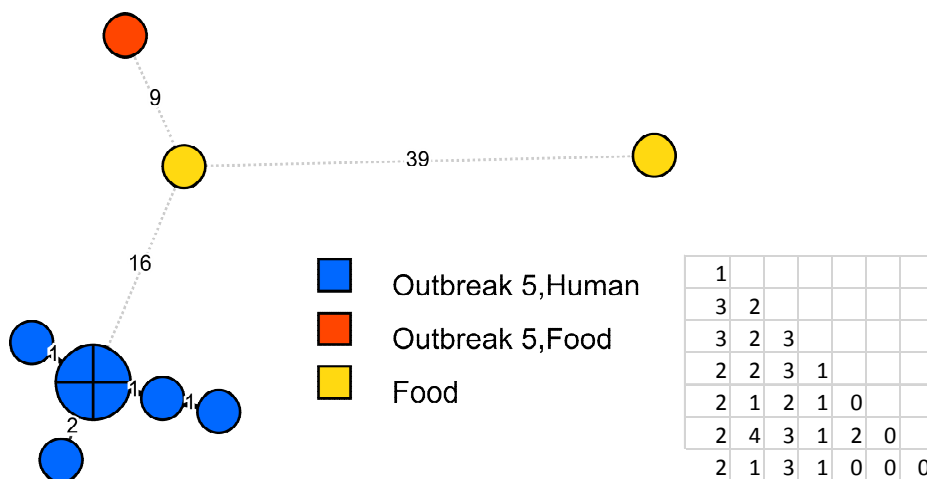
5.2.5. Outbreak 5 – CC415

The outbreak MLST type ST394 belongs to CC415, for which there yet is no complete genome for reference mapping. The SNP analysis therefore used a *de novo* assembled reference from ST394. All isolates are originating from country B. Eight isolates from humans associated with the outbreak were sequenced. One human isolate that originally was thought to be part of outbreak 5 (as listed in appendix 6) was shown to be of a different CC and the isolate was therefore not included in the outbreak analysis. One isolate from raw milk identified as having the same molecular typing profile by fAFLP but for which there was no epidemiological evidence of being linked to the outbreak were also sequenced. The initial SNP analysis showed that the raw milk isolate was quite distinct from the human isolates with over 1,000 SNPs difference, while the distance between the isolates from the eight cases were at most 4 SNPs. Closer inspection of the SNPs distribution showed that in the raw milk isolate, almost all SNPs were located within a phage in the *L. monocytogenes* genome. All human isolates had the same phage in this region, which was different but related to the phage in the raw milk isolate. On initial inspection, the difference in the number of SNPs between the raw milk and human isolates indicated that the milk isolate was not associated with the outbreak. Re-analysing the sequences for SNPs after removing these phage regions showed that the human isolates still clustered together, but the raw milk was now 30 SNP from the cluster. This illustrates one of the potential pitfalls with a whole genome analysis approach where several SNP/alleles can be acquired all in one event, and thus may not necessarily mean that strains are genetically unrelated. This is an important point that should be taken into consideration when using SNP analysis for investigating the genetic relatedness of strains i.e. where SNPs have occurred and not just the total number. In this instance, however, on re-analysis, the raw milk isolate was still 30 SNPs different to the human isolates and unlikely to be the cause of the outbreak.



The tree is constructed using all isolates having the same CC as the outbreak isolates. To the right is shown a pairwise SNP distance matrix of the outbreak isolates. Number on branches indicates the number of SNP differences the branch corresponds to. The food outbreak isolate is the one from raw milk, which was eventually discarded in the country’s investigation phase as not being part of the outbreak (see text above).

Figure 5.9.: SNP maximum parsimony tree of outbreak 5



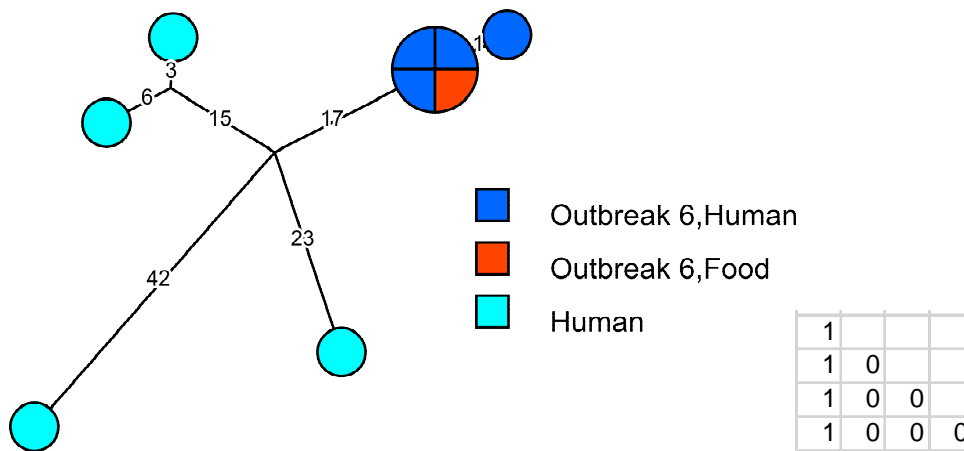
The tree is constructed using all isolates having the same CC as the outbreak isolates. To the right is shown a pairwise distance matrix of the outbreak isolates indicating the number of allele differences. Number on branches indicates the number of allele differences the branch corresponds to.

Figure 5.10.: cgMLST minimum spanning tree

5.2.6. Outbreak 6 – CC398

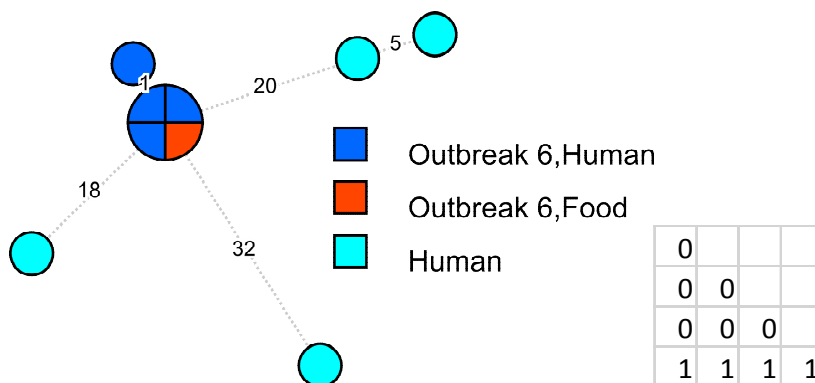
This outbreak was identified as a point source cluster with the rare MLST ST-802/CC-398 in country T. The source was epidemiologically identified as fermented fish; something that were confirmed by isolation of the same rare ST from food. We do not yet have a common reference strain for mapping analysis within CC398. We therefore used a *de novo* assembled genome from the outbreak itself as a reference. All five isolates in the outbreak were already sequenced by the national Institute of Public Health in country T, and were only analysed in this project. The SNP analysis of the four human and

one food isolates showed that there were only a single SNP found between all the isolates and hence a max distance of 1 SNP and the mode 0 SNP. This corroborates the single point outbreak epidemiology.



The tree is constructed using all isolates having the same CC as the outbreak isolates. To the right is shown a pairwise SNP distance matrix of the outbreak isolates. Number on branches indicates the number of SNP differences the branch corresponds to.

Figure 5.11.: SNP maximum parsimony tree of outbreak 6



The tree is constructed using all isolates having the same CC as the outbreak isolates. To the right is shown a pairwise distance matrix of the outbreak isolates indicating the number of allele differences. Number on branches indicates the number of allele differences the branch corresponds to.

Figure 5.12.: cgMLST minimum spanning tree

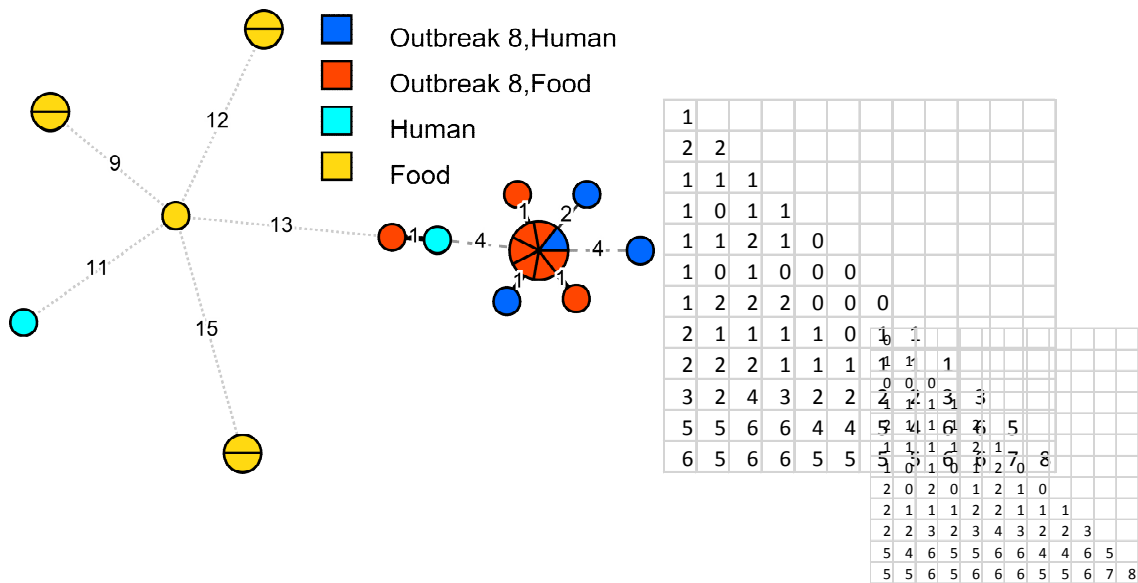
5.2.7. Outbreak 7 – CC87

This is analysis of two ST87 reported outbreaks between January 2013 – February 2014 from a specific region of country X (Perez-Trallero et al., 2014). ST87 belongs to the relatively rare human serotype 1/2b. The outbreaks were unusual in that there were a high proportion of pregnancy related cases (7 out of 12 human cases in the one analysed here). The two described ST87 outbreaks (that were temporally separated) were separated by a single band shift in PFGE *AscI* analysis (*SmaI* was identical).

There are two duplicates in this study; one with isolates from both mother and daughter, and one with both blood and CSF isolates from the same patient. These were 1 and 0 SNP apart respectively in the SNP analysis. In the cgMLST analysis the pairs were both identical.

WGS analysis shows that the two described outbreaks cannot be separated neither by SNP, nor cgMLST analysis. The pairwise SNP distances showed an unusual pattern with a bell shaped distribution centred around 4-7 SNP (Figure 5.13). With the information that ST87 was common in that geographical region throughout this year, with two described clusters and some sporadic cases, it could be hypothesized that the original source was somewhere with a *L. monocytogenes* population that have had the time to diverge for a relatively long time. The epidemiological investigation could not pinpoint the sources more closely than that, the first outbreak seemed to be related to ham, while in the second one the outbreak type were found in foie gras from the refrigerator of one case.

Our analysis found that two more human isolates could be assigned to the outbreak, both from country X in 2011 - 2 years before the outbreak.



The tree is constructed using all isolates having the same CC as the outbreak isolates. To the right is shown a pairwise distance matrix of the outbreak isolates indicating the number of allele differences. Number on branches indicates the number of allele differences the branch corresponds to.

Figure 5.17.: cgMLST minimum spanning tree

The outbreak isolates cluster closely together and have the food isolates mixed in. This solidifies that the correct source was identified.

5.2.9. Outbreak 9 – CC4

An outbreak in country C with an epidemiologically identified source in brie cheese (Tourdjman et al., 2014). Unfortunately, the bacterium was not isolated from the cheese samples, so the sequenced isolates are all human. The 25 isolates submitted were a mix of actual outbreak isolates and background isolates from the same country and period. In the original PFGE typing all the submitted isolates were identical using both restriction enzymes *AscI* and *ApaI*. The third PFGE enzyme *SmaI* could properly separate the 11 outbreak isolates from the background.

The WGS analysis confirmed the *SmaI* separation, but in a much more convincing way. Long branches separate the outbreak isolates from all the isolates where only *AscI* and *ApaI* were identical.

5.3. Conclusions

The cgMLST and SNP analyses showed concordant results with similar numbers (SNP and allele differences) when the outbreak isolates were analysed (Table 5.1). In the following discussion the numbers in the following conclusion will refer to SNP differences, but aside from the 2-3 deviant cgMLST results in outbreak 1 (Section 5.2.1) the same conclusions apply for cgMLST. Most of the outbreaks are tightly clustered; six out of nine show a typical point-source-like pattern with a median pairwise distance of ≤ 5 SNP and a maximum pairwise distance ≤ 10 SNP.

The other three outbreaks each have their own specific profiles. Cluster 7 shows an unusual pattern where the isolates have distances between 4 and 7 SNP. In an ideal world, it would be desirable to use a fixed cut-off threshold for cluster definition, but in the real world, this is not possible. Using a cut-off of 5 SNP in cluster 7 would fail to define the outbreak properly, but a cut-off of 10 would include all of the confirmed isolates. It can be noted that country X saw an increase of this type for a prolonged period. The type was found in diverse geographical locations and there were also small variations found in the PFGE patterns. The presence of the outbreak isolate in an extended spatial and a temporal space is congruent with a larger variation in the population. This is also confirmed by the variations in WGS.

Outbreak 2 and 4 both show somewhat higher pairwise SNP distances: median of 10.5 and 9, with maximums of 21 and 12, respectively. It should be noted that both of these clusters are occurring during relatively long time spans. If using a single linkage model, the longest branch needed to link all the cluster isolates would be 12 and 9 SNP, respectively.

Outbreak 5 showed the impact that the auxiliary genome can have, unless the proper reference genomes are used. It should be noted that in the analysis of all these outbreaks, the focus was on the core SNP and core cgMLST. The auxiliary genome can also be very valuable (Wang et al., 2015), but currently the core analysis is the most stable and also the one amenable to tracking the isolates in time.

The story told by the WGS analyses is reflected in the epidemiological descriptions (see individual outbreak descriptions). Higher diversity is often linked to temporally extended outbreaks, even though the cause of variation seen in outbreak 7 only can be speculated about.

No food isolates in the study collection, apart from those already described as part of the outbreaks, were similar to any of the nine outbreak types. In four of the outbreaks, one or two human isolates submitted as sporadic cases clustered together with the outbreak isolates. In all four instances, these isolates originated from the same country as the outbreak. The isolates we have sequenced in this project obviously is a subset of all the existing *L. monocytogenes* types found in food and humans in Europe and in the world. Although the LISEQ data set was designed to give a good coverage of the European situation in the years 2010-2011, there were not identified any international components in these epidemiologically confirmed outbreaks. This does not mean that international outbreaks do not occur, just that that the nine specific outbreak strains were not found among the study isolates from other countries.

The cgMLST analysis showed for the most part very concordant results with the SNP analysis. In 8 outbreaks the median and maximum sizes of branches within a whole CC was shorter (Table 5.1) compared to SNP branches. This is to be expected since several SNP within the same loci only results in a single allele difference. Between the outbreak isolates where the genetic distance is a lot shorter the differences are a lot smaller and there is no clear trend that either cgMLST or SNP results in a higher resolution. The exception here is outbreak 1, which is described in Section 5.2.1. We have used current state of the art bioinformatic methods, but since the two approaches use different computational methods for defining the differences in number of SNP or number of alleles, the results cannot be expected to be identical. In theory the number of pairwise SNP differences should always be higher (you can fit several SNP into one allele but not vice versa) but since the methods have computational limitations in this young research field there are still misassemblies (causing false allele calls), erroneous base calls (resulting in false SNP calls) etc. There is still a need for the scientific

community to perform targeted studies on the correlation between the methods, but both methods work well for defining clusters in outbreaks.

The WGS analysis managed to clearly separate the outbreak isolates from the background, so WGS is very well suited for clearly defining outbreaks. The results also indicate that every outbreak should be considered in its own context and that one should not use a single universal cut off value for separating an outbreak from background isolates. Outbreaks extended in time such as number 2 and 4 had up to 21 SNP or 16 cgMLST loci between individual isolates. More clonal outbreaks, such as number 6 and 9 only had a maximum of 2 SNP/4 cgMLST alleles between any isolate, even though outbreak 9 was the largest outbreak analysed (n=24). In the nine outbreaks analysed here, there were no issues separating the outbreak isolates from the background, even with a fixed cut of value if that value is set high enough. Nevertheless, with increasing number of isolates analysed, the problem of a fixed cut-off will be exacerbated.

6. Genetic diversity

Section 4 illustrates that *L. monocytogenes* contains a large number of variants. The extent of this variation, or diversity, may differ by source reservoir or from humans as has been exemplified for *Campylobacter* (Strachan et al., 2013). This genetic diversity can be characterised in a number of ways including by Simpson's diversity index (Simpson, 1949) and rarefaction (Heck et al., 1975). All of the different microbial typing methods used earlier in this report can be utilised. However, for pragmatic reasons only 7 locus MLST or 30 locus rMLST are practical because cgMLST and SNP based methods would produce too many "types" which would not really provide useful information about variation (practically every genome would yield a different type).

The primary aim of this section is to address Specific objective 2 part (i) - **to explore the genetic diversity of *L. monocytogenes*** within and between different sources and human origin. This section also investigates the genetic distance between each source by Nei's genetic distance (Nei, 1975). This methodology provides information on the genetic relatedness of isolates between sources and in particular whether sources have distinctive populations of *L. monocytogenes*.

6.1. Methods

6.1.1. Simpson's Diversity Index

Simpson's Diversity Index was used to obtain an estimate of the diversity of strains by source (Simpson, 1949),

$$Diversity\ Index = 1 - \sum_{All\ STs} (f_i)^2$$

where f_i is the relative frequency of ST i in a specific source. A value of 0 of the diversity index indicates that all strains are the same and a value of 1 indicates that they are all different (maximum diversity). Confidence intervals were calculated by the bootstrap method using the Pop Tools add-in for Microsoft Excel (www.poptools.org) and significant differences between pairs of sources were calculated using a randomisation test (Manly, 2007).

To generate bootstrapped confidence intervals the ST's were resampled with replacement. This was done 10,000 times using Pop Tools. From this the mean values of Simpson's diversity index were calculated as well as 95% confidence intervals.

To calculate significant differences between each pair of populations (e.g. human and bovine) these were randomized in Excel using PopTools and Simpson's diversity index calculated for each. Correction for differences in sample size between the sources was carried out using the following method. For each pair of sources to be compared, the one with the lowest number of isolates (I_{low}) was kept

constant whilst the other was resampled without replacement for I_{low} isolates. This process was then repeated 10,000 times using the Monte Carlo Excel add-in @RISK (Palisade Ivybridge, United Kingdom). The posterior distribution of Simpson's diversity indices was then compared with the non-randomised diversity index to obtain the level of significance (P value).

6.1.2. Rarefaction

The extent to which the isolates from sources had sampled the maximum number of genotypes was characterised using rarefaction. Rarefaction is a data re-sampling technique that indicates whether all of the genotypes have been sampled which results in the curve reaching a plateau or if the curve is still increasing there are still more genotypes in the population to be sampled (Heck et al., 1975). Statistical significance was determined by randomisation test as described above.

6.1.3. Nei's genetic distance

Standardized genetic distances (d_1) between pairs of sources were determined using the method of Nei (Nei, 1975) and applied to genetic locus and SNP data by the method of Manly (Manly, 2007). Briefly, for N_{loci} the distance is calculated as

$$d_{1ij} = \frac{1}{N_{loci}} \sum_{All\ loci} 0.5 \left(\sum_{All\ alleles} |p_i - q_j| \right),$$

Where p_i and q_j are the frequencies of the alleles (or SNPs) at each locus in source i and j , respectively.

A Nei's value of 0 indicates that the populations are identical whilst a value of 1 indicates that the two populations have no genotypes in common. Statistical significance was determined by randomisation test as described in 6.1.1.

6.1.4. Graphical visualisation and cluster analysis

A phylogenetic tree utilising the 50,297 SNPs generated by Parsnp (Treangen et al., 2014) was generated by MEGA (Tamura et al., 2013) utilising the nearest-neighbour joining technique. *L. innocua* was used to root the tree. This enabled visualisation of isolates around the phylogenetic tree.

To determine whether there was clustering of isolates from each source on the phylogenetic tree the following analysis was performed. Within each source pairwise SNP distances were calculated. The percentage of pairwise SNP distances less than a cut-off value (100 SNP's) was calculated. The percentage of pairwise SNP distances between isolates outside the source was then calculated using the same cut-off criterion as previously. If the percentage between source isolates was greater than the percentage between outside isolates, this was taken as evidence of clustering.

6.1.5. Analyses

Table 6.1 describes the list of analyses that were performed for each of the different methods and also the level of molecular analysis. The cgMLST for all the study isolates except for *Listeria innocua* were defined with the help of the Institute Pasteur as described in Section 4. The scheme applied was composed of 1,748 genes representing 125,029 alleles. From this 7 locus MLST and 30 locus rMLST (30 rMLST loci are utilised in the Institute Pasteur cgMLST scheme) data were obtained. SNPs detection was performed using the Parsnp software for all of the genomes in the database resulting in 39,529 core genome SNPs (cgSNP).

Table 6.1.: Diversity, rarefaction and genetic distance analyses carried out by level of molecular analysis

Level of molecular analysis	Simpson's Diversity	Rarefaction	Nei's genetic distance
7 locus MLST	√	√	√
30 locus rMLST	√	√	√
1,748 cgMLST	na	na	√
39,529 cgSNP	na	na	√

na – not applicable because the large number of loci results in practically every isolate being unique.

6.1.6. Selection of Genomes for analysis suitable for genetic diversity and source attribution analysis

Table 6.2 provides details of the number of human genomes and also the number that have cgMLST profiles but are not part of an outbreak. In a previous source attribution analysis (Little et al., 2010) human cases had been separated into two groups (one younger than 60 years and the other greater or equal to 60 years). Of the human data that were not part of an outbreak and had age data there were 50 in the <60 years age group with 121 being in the older age group. To determine whether there is a difference in age stratification (See Appendix 7) Nei's genetic distance was calculated from the 7 locus MLST data. No significant difference was found using a randomisation test ($P=0.141$). In addition Simpson's diversity index and rarefaction were carried out but no significant differences were found between the two groups. Hence, all of the human data were analysed as one dataset as the results show that there is no difference in diversity and genetic relatedness by age group.

Table 6.2 also shows the number of genomes that were allocated to particular sources. The designation of source did not depend on the part of the food chain from which the isolates originated. For example, genomes allocated to fish would include those from a fish sampled at a fish farm, all the way along the food chain, to those sampled at retail. The mixed category primarily comprises complex foods made of a number of ingredients such as sandwiches etc. A number of the sources were represented by a small number of genomes, which were insufficient for the analysis of diversity. It was decided to only consider those distinct sources with ≥ 25 genomes available for analysis. This cut-off was based by work done on *Campylobacter* (Smid et al., 2013), where they advised >25 isolates should be used per source.

Table 6.2.: Numbers of genomes categorised to source and the subset which were not part of an outbreak and for which cg MLST data were available

Human and source	Number of Genomes	Number of genomes with 7 locus MLST and not part of an outbreak
Human*	333	261
Mixed	30	27
Poultry*	32	25
Bovine*	80	61
Shellfish	3	0
Swine*	114	112
Fish*	325	324
Unspecified	101	101
Vegetable	5	5
Ovine*	117	89
Caprine	3	3
Total	1,143	1,011

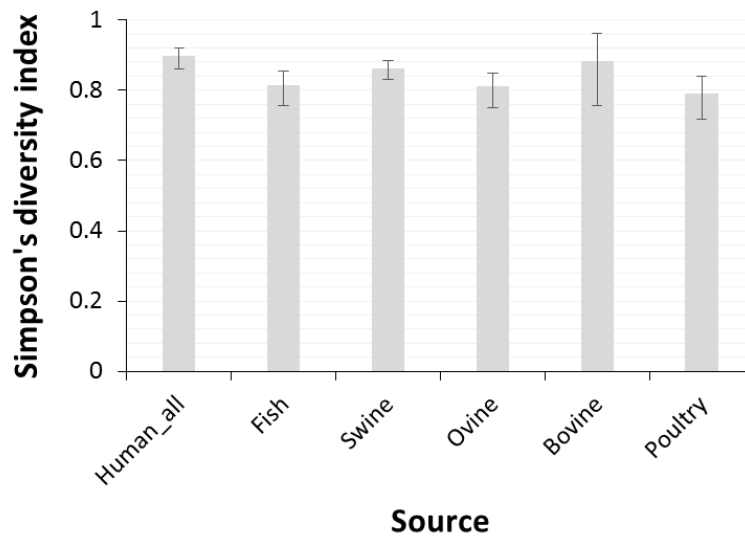
*denotes used in source attribution comprising 872 genomes. The other sources were not included as they comprise <25 genomes or were mixed/unspecified source.

6.2. Results and Discussion

6.2.1. 7 locus MLST

Simpson's diversity index was determined for isolates from humans and each of the sources (Figure 6.1. (a)). All exhibited high diversity ranging from 0.897 (Humans) to 0.811 (ovine). However, pairwise comparisons between each pair of sources yielded no significant differences ($P > 0.05$). Rarefaction was also carried out (Figure 6.1. (b)) and it can be seen that bovine and human isolates have a higher number of new STs per isolate sampled compared with fish and ovine. Swine is intermediate and poultry has too few isolates to determine a trend. It is also noticeable that the rarefaction curves have not plateaued and that if a larger sample had been obtained then additional novel genotypes would have been found.

(a)



(b)

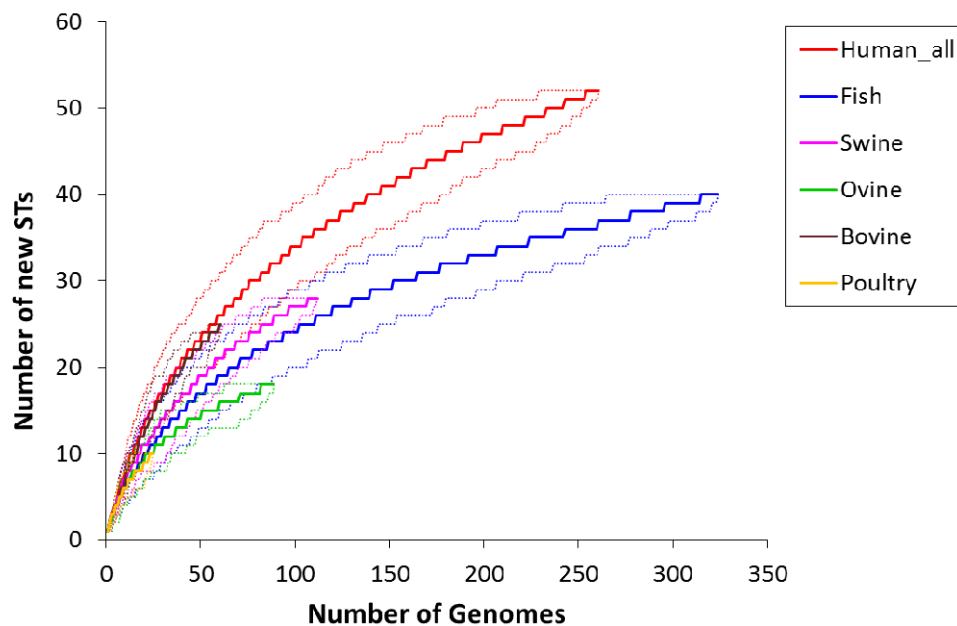
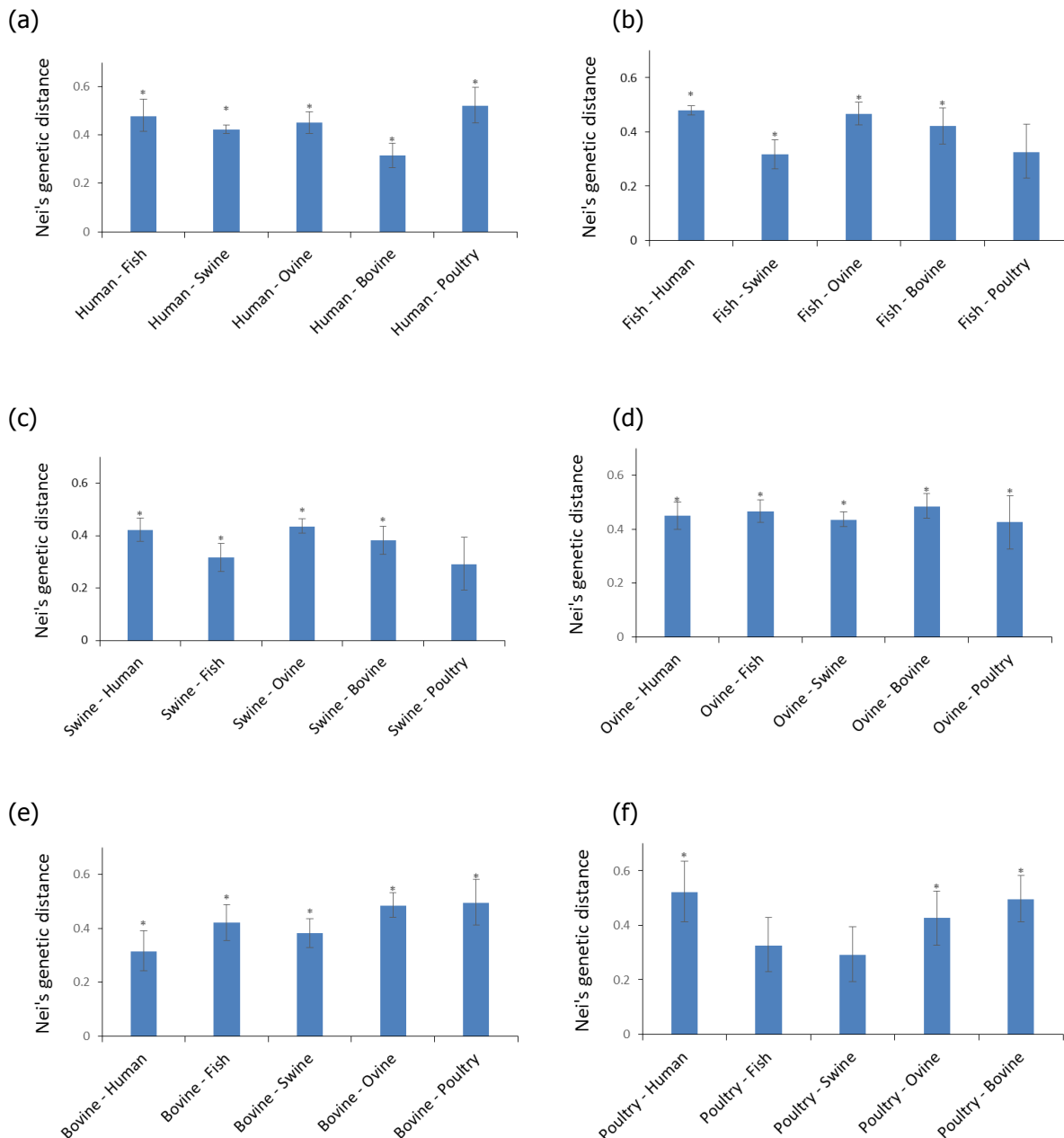


Figure 6.1.: (a) Simpson's diversity index with 95% bootstrapped confidence intervals, and (b) rarefaction of the 7 locus MLST human and source genome data (dashed lines show 95% bootstrapped confidence intervals)

Nei's genetic distance was determined between isolates from humans and the 5 sources (Figure 6.2). The genetic distance was significantly different between human and all of the other 5 sources ($P < 0.05$). Bovine had the closest genetic distance to human. All but two of the other pairwise comparisons (i.e. those that did not involve humans) also showed significant differences. The two which were statistically similar ($P > 0.05$) involved poultry, which was the source present in lowest numbers and it may be that this could be an artefact due to small sample size.



Confidence intervals (95%) were generated by the bootstrap method. Asterisks denote significant pairwise difference between each pair of sources ($P < 0.05$) using randomisation test described in 6.1.1.1.

Figure 6.2.: Pairwise Nei's genetic distance by 7 locus MLST between (a) humans, (b) fish, (c) swine, (d) ovine, (e) bovine, (f) and poultry and the remaining sources

6.2.2. 30 locus rMLST

Simpson's diversity index was again high for the 30 locus rMLST ranging from 0.91 in humans to 0.83 in ovine (Figure 6.3). No significant differences between any of the sources was observed ($P > 0.05$) as was the case for 7 locus MLST.

Rarefaction by 30 locus rMLST comprises more genotypes than for 7 locus MLST (e.g. for humans MLST provides 52 genotypes whilst rMLST has 73). Hence, the rMLST curves are steeper than for

MLST and again the curves do not plateau indicating that increasing sample size will generate additional novel genotypes. For rMLST isolates from humans have the highest number of new STs per isolate sampled as was the case for 7 locus MLST. However, bovine which had a similar curve to human in 7 locus MLST is now more similar to the other sources.

Nei's genetic distance shows that the *L. monocytogenes* population in humans is distinct compared to that in the other source reservoirs ($P < 0.05$) (Figure 6.4.(a)). However, again as in 7 locus MLST, the distance between humans and bovine is the smallest. All of the comparisons between sources (Figure 6.4. (b)-(e)) all show significant pairwise differences ($P < 0.05$).

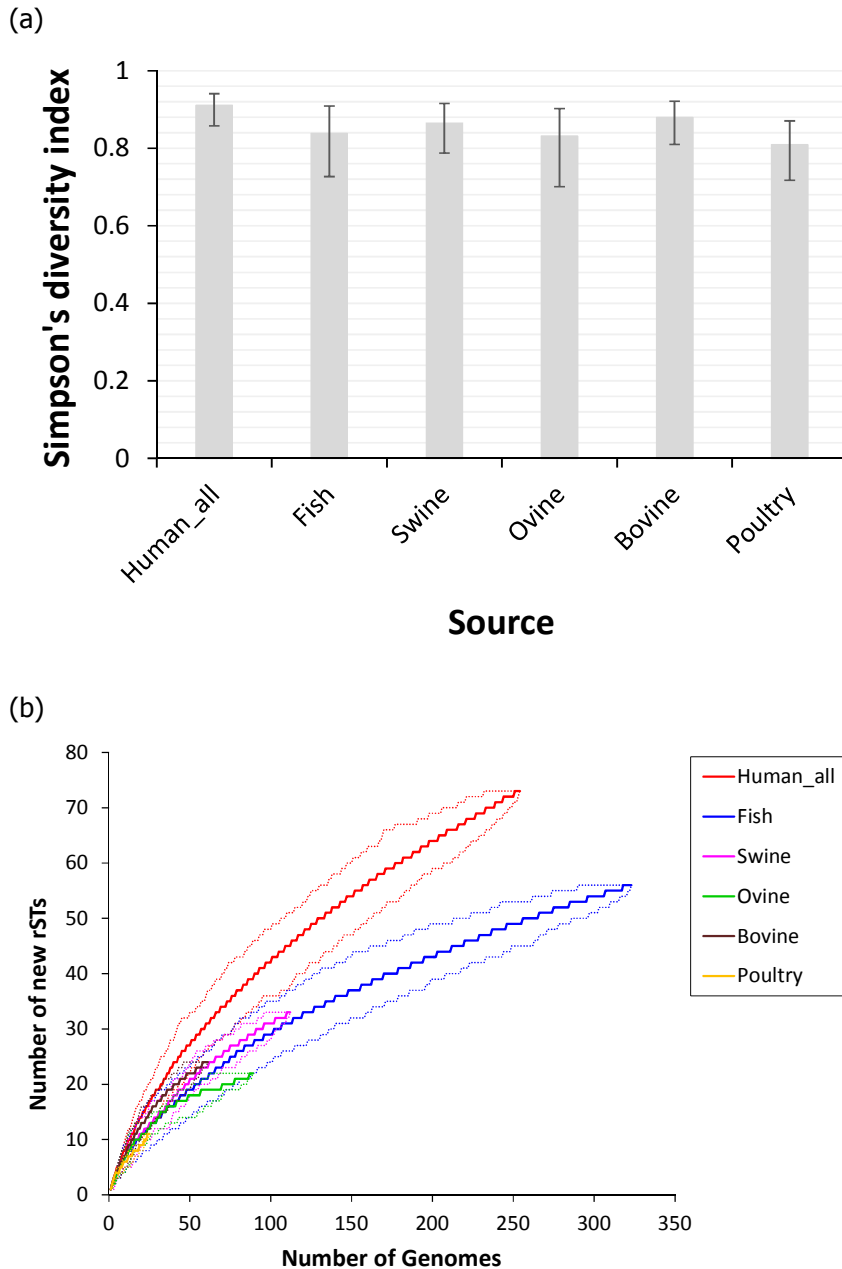
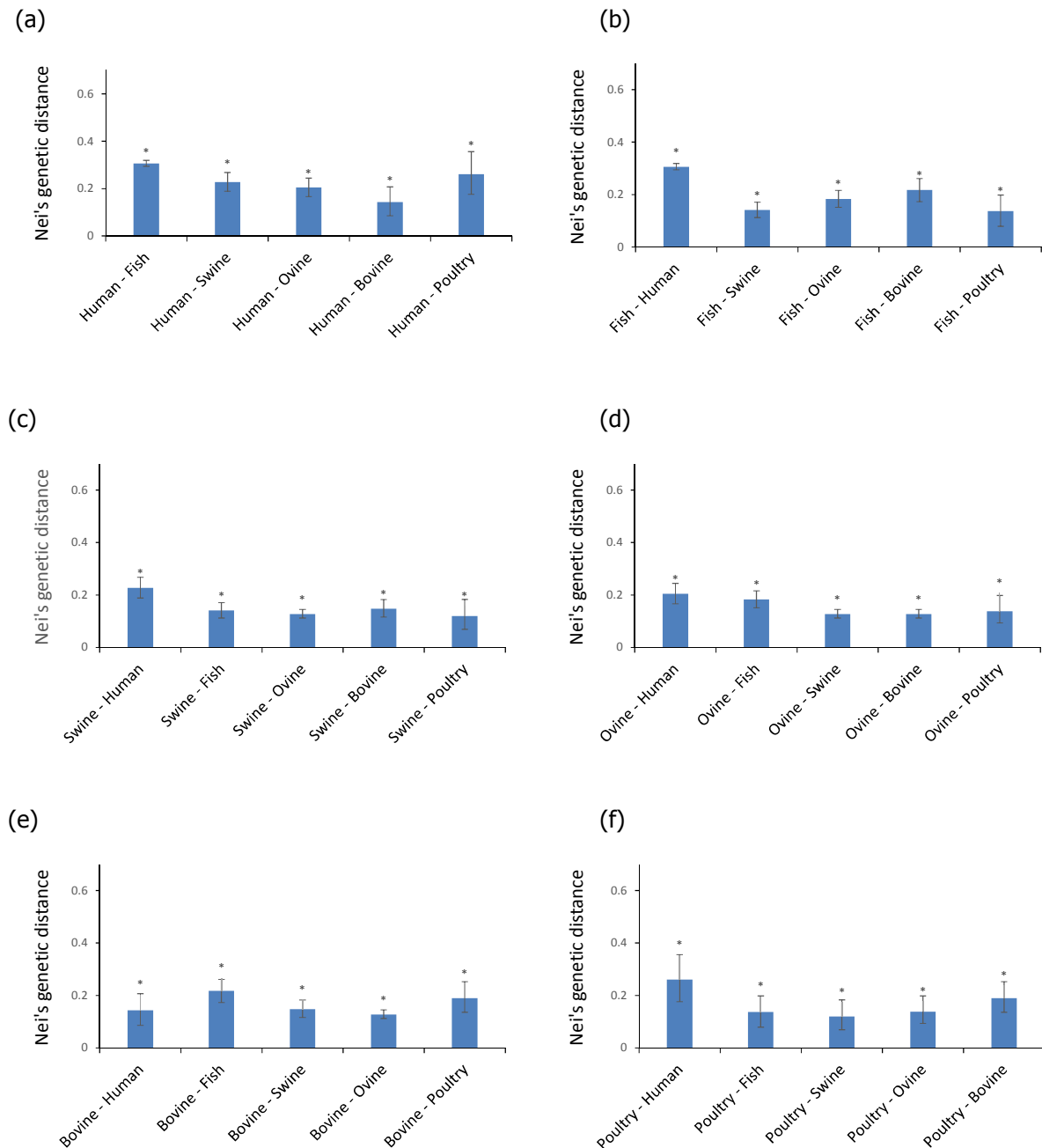


Figure 6.3.: (a) Simpson's diversity index with 95% bootstrapped confidence intervals and (b) rarefaction of the 30 locus rMLST human and source genome data using randomisation test described in 6.1.1.



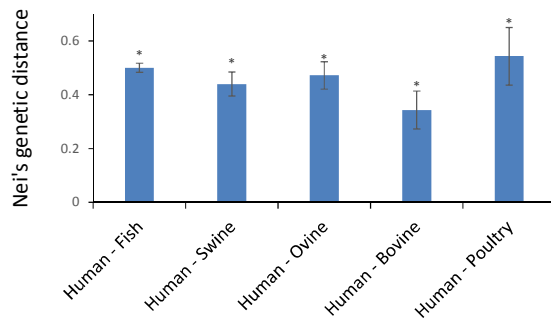
Confidence intervals (95%) were generated by the bootstrap method. Asterisks denote significant pairwise difference between each pair of sources ($P < 0.05$) using randomisation test described in 6.1.1.

Figure 6.4.: Pairwise Nei's genetic distance by 7 locus MLST between (a) humans, (b) fish, (c) swine, (d) ovine, (e) bovine, (f) and poultry and the remaining sources

6.2.3. 1,748 locus cgMLST

Figure 6.5 shows Nei's genetic distance, comparing humans with the source reservoirs, for 1,748 locus core genome MLST. There is a significant difference ($P < 0.05$) in the distance between humans and all of the sources. However, bovine has the smallest distance to human. However, bovine has the

smallest distance to human. It was not possible to perform comparisons between the source reservoirs within the timeframe of the project.

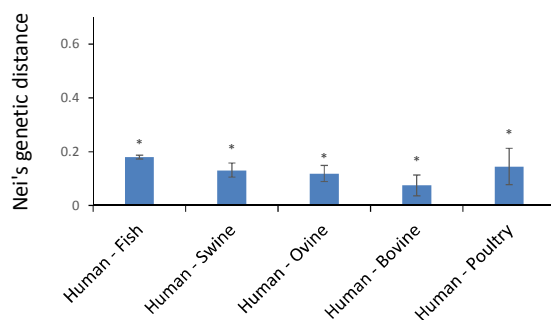


Asterisks denote significant pairwise difference between each pair of sources ($P < 0.05$) using randomisation test described in 6.1.1.

Figure 6.5.: Pairwise Nei's genetic distance by 1748 locus cgMLST between (a) humans and the remaining sources

6.2.4. 39,529 cgSNP

Nei's distance again shows significant differences between Humans and all of the animal reservoirs and that the smallest distance is between humans and bovine (Figure 6.6). Comparisons between sources were not carried out because of the long computational times required.

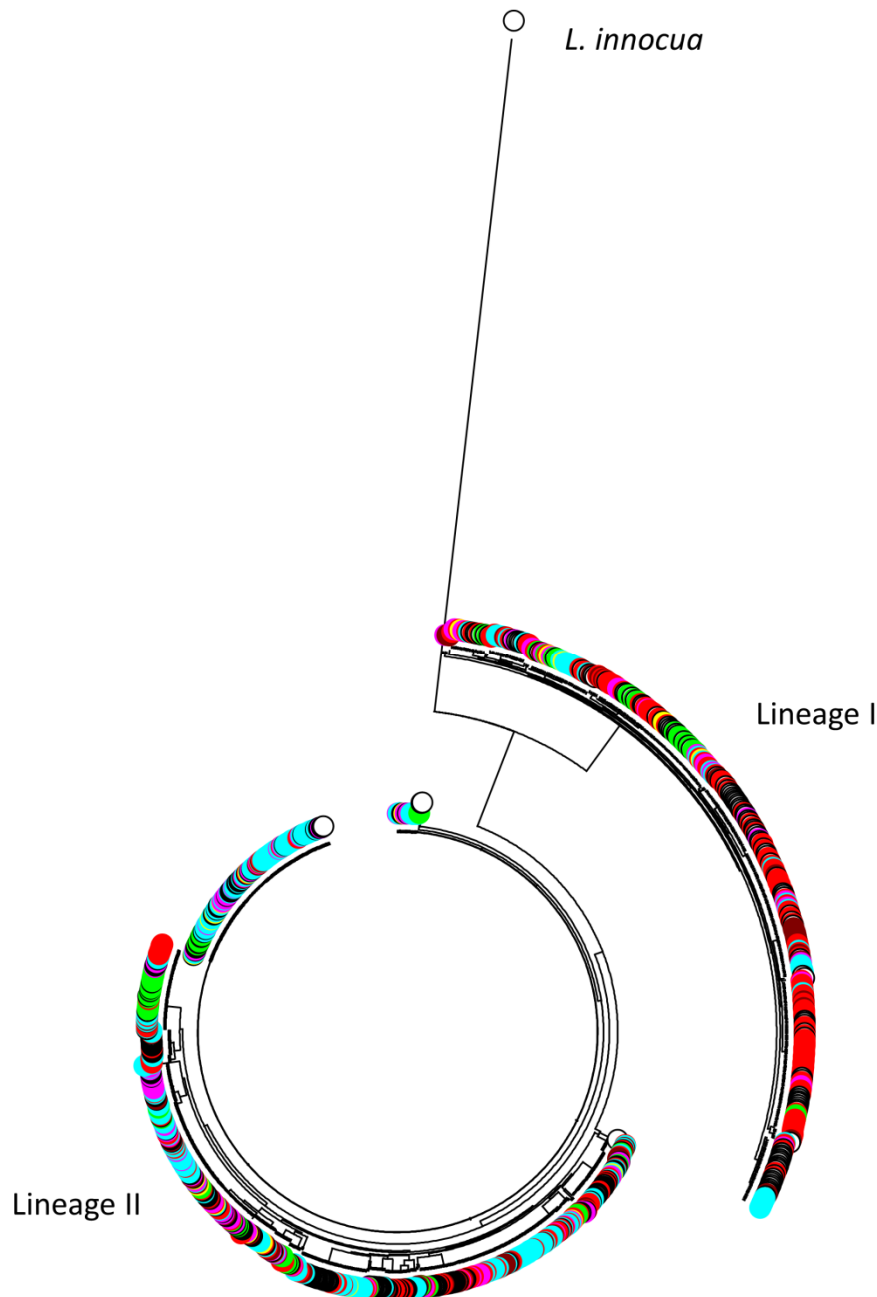


Confidence intervals (95%) were generated by the bootstrap method. Asterisk denote significant pairwise difference between each pair of sources ($P < 0.05$).

Figure 6.6.: Pairwise Nei's genetic distance by 39,529 locus cgSNP between (a) humans, (b) fish, (c) swine, (d) ovine, (e) bovine, (f) and poultry and the remaining sources

6.2.5. Graphical Visualisation

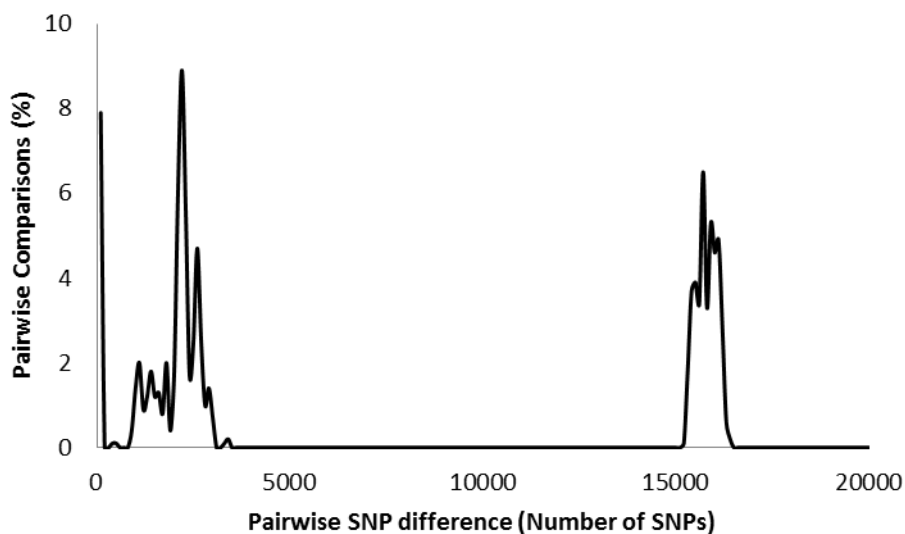
Figure 6.7 shows the phylogeny of the *L. monocytogenes* by source and appears to show that there is a non-even distribution of isolates by source. For example there appear to be more human isolates (red) in lineage I whilst there are more fish isolates (blue) in lineage II. Also, within the lineages by visual inspection there appears that there may be clustering of sources.



Scale: *L. innocua* to Lineage I is 31200 SNPs. The tree is drawn to scale using the Neighbor-Joining method (See 6.1.4).

Figure 6.7.: SNP based neighbour joining tree of *L. monocytogenes* rooted with *L. innocua* (fish – blue, swine – pink, ovine – green, bovine – brown, poultry – yellow, human – red and other - white).

Figure 6.8 shows the SNP differences between all of the 872 isolates. The peaks to the right (large SNP differences) show the differences between isolates from different lineages. There is a peak of 7.9% for pairwise comparisons <100 SNPs. This was used as the cut-off in the further analysis to establish clustering of isolates within the phylogeny.



Here the number of SNP differences between each pair of isolates was calculated and plotted on the graph.

Figure 6.8.: Pairwise SNP difference comparisons between all 872 isolates

Table 6.3 shows the clustering analysis that was performed. The mean and median values of pairwise SNP distance within each source are not in themselves that helpful because representatives of each source occur across different lineages. The SNP distance between lineages is large and the relative distribution of a source across lineages is the main driver for the size of the mean and median values. The 95 percentiles are very broad because isolates from humans and the different sources are spread across and within the different lineages. The final column in Table 6.3 shows the percentage of pairwise comparisons with a cut-off of 100 SNPs. Pairwise comparisons between all isolates (irrespective of source) indicates that 8.2% are within 100 SNPs of each other. Pairwise comparisons within human (8.6%) and bovine (7.3%) isolates are in the same range as "all" and hence do not appear to show evidence of clustering at this level. Whereas poultry (21.0%), fish (16.1%) and ovine (16.0%) appear to all show evidence of clustering.

Table 6.3.: Pairwise SNP differences within each source

Source	Mean SNP difference ($\pm 95\%$ CI)	Median SNP difference	% less than 100 SNPs
Human	8,655 (0 – 16,190)	15,290	8.6
Fish	4,781 (0 – 16,072)	2,155	16.1
Swine	7,407 (0 – 16,192)	2,333	11.7
Ovine	7,217 (0 – 16,104)	2,661	16.0
Bovine	8,344 (0 – 16,121)	2,733	7.3
Poultry	6,849 (0 – 15,997)	2,153	21.0
All	7,638 (0 – 16,140)	2,538	8.2

CI: confidence interval; SNP: single nucleotid polymorphism.

Note: The mean, median and 95% confidence intervals are presented. Also, the percentages of pairwise comparisons within each source that have <100 SNP difference are also given.

6.3. Conclusions

Simpson's index for humans and the 5 sources exhibited high diversity (>0.8) for both 7 locus MLST and 30 locus rMLST. Simpson's index of diversity between each of the sources was indistinguishable.

Rarefaction showed that for both 7 locus MLST and 30 locus rMLST that all of the genotypes had not been sampled. However, for the sources with the largest number of samples, it does show that the number of types from humans is considerably greater than that for fish. For both 7 and 30 locus MLST humans have the steepest slope, so have the largest number of genotypes per isolate sampled. Bovine is second, virtually indistinguishable from humans at 7 locus MLST but lower at 30 locus MLST.

Nei's genetic distance showed that there were significant differences between human and all sources at all levels of molecular analysis explored. Also, for all levels the distance between humans and bovine was the smallest. Computing times became long when dealing with analysis of cgMLST and cgSNS's. As a result calculation of Nei's genetic distance between sources and associated confidence intervals and randomisation tests were not carried out.

Visualising the SNP based phylogeny tree appears to show some areas of clustering by source, though there are many parts of the tree which are quite heterogeneous. When investigating this analytically and comparing pairwise SNP differences with a cut-off of 100 SNPs there was no evidence that there were no independent "host" clusters of bovine and human compared with chance whereas there was evidence to show that this occurred for a number of the other sources (poultry, fish and ovine). Whether this is a robust finding or an artefact of the sampling for this study can only be resolved when additional isolates become available to see if the pattern continues or otherwise. If it is generally found that there are parts of the phylogeny where there are clusters (<100 SNPs) comprising a particular source, then this indicates that they are closely related.

It should be noted that isolates for each source came from different points in the food chain. Those isolates obtained from sources closest to retail are likely to have had greater chance of being a result of cross contamination from another source. It was not possible to investigate this in the current study because there were insufficient data. However, this should be borne in mind when considering the robustness of the results. Future work should investigate the effect that isolates from different points (animal, factory, retail, human) along the food chain may (or may not) make to the analysis. It was not possible to do this here because there were insufficient isolates to perform this type of analysis.

7. Epidemiological relationship: Source attribution

The current section addresses specific objective 2(ii): "Assess the epidemiological relationship of *L. monocytogenes* from the different sources and of human origin considering the genomic information and the metadata available for each isolate." It achieves this using the method of source attribution.

The term "source attribution" has been defined (Pires et al., 2009) as: "...the partitioning of the human disease burden of one or more foodborne infections to specific source, where the term *source* includes animal reservoirs and vehicles (e.g. foods)."

Attribution can be carried out at different points along the food chain (Pires et al., 2009). This can include at production, distribution and consumption. In the current project because of the relatively small number of isolates, all of the isolates along the food chain that originate from a particular reservoir are combined (see Section 6). This enables the following sources of isolates and their respective genomes to be determined: bovine, ovine, swine, fish and poultry (see Section 6). Human clinical cases are then attributed to these sources by comparing the genotypic subtypes from the human and source isolates. The microbial subtyping approach involves characterization of isolates of a

specific pathogen by genotypic subtyping methods (e.g., MLST, cgMLST, cgSNPs etc). These data can then be used to perform source attribution utilising mathematical models (Mughini-Gras and van Pelt, 2014).

7.1. Methods

7.1.1. Source Attribution Methods

Availability of models

Appendix 8 provides the links to all of the attribution model programs used in this study.

Dutch Model

The Dutch model (Mughini-Gras and van Pelt, 2014) is a straight forward way to estimate the attribution of a particular genotype (e.g. ST) to a reservoir, when the frequency distribution of each type is known for each reservoir. If p_{ij} represents the frequency of type i (e.g. ST 19) in source j (e.g. poultry) then the attribution score of type i in source j is given by

$$Score_{ij} = \frac{p_{ij}}{\sum_j p_{ij}}$$

where the summation by j considers all the reservoirs where data exist (e.g. cattle, sheep, wild birds, chicken, turkey etc.).

When applied at ST level this model does not guarantee that all STs will be attributed to sources. This is because human types that are not found in the animal reservoir cannot be attributed. However, if genetic information exists at multiple loci, as in this study, then the Dutch Model can make use of the frequency of each individual allele at each individual locus, and estimate attribution even for STs that are not present in the animal reservoirs. In particular, at allele level the frequencies $p_{a_{ijk}}$ can be calculated for each allele a_{ijk} of all isolates from the animal reservoirs, where i is subtype, j source and k the loci number.

The attribution score of bacterial subtype i in source j is

$$Score_{ij} = \frac{p_{ij}}{\sum_j p_{ij}}$$

where $p_{a_{ijk}} = BetaInv(0.5, 0 + 1, N_{isolates} + 1)$ if its frequency is zero (*BetaInv* fn in Excel). This assumes that we have no prior knowledge of $p_{a_{ijk}}$ and so is maximally noncommittal or conservative.

Sample size correction and confidence intervals: Since the sample sizes of the sources are different a correction is incorporated. If the sample size of the smallest source is N_{min} then the Dutch model is run by sampling without replacement of N_{min} isolates from each source (e.g. $N_{min}=25$, which is the sample size of poultry reservoir or 61 (sample size of cattle) when poultry are discarded from analysis). This process is repeated for 10,000 iterations.

After each iteration the attribution scores of each human isolate to each source are re-calculated using the above equation. These scores are then averaged across the number of isolates (e.g. $n = 254$ for humans) and stored. The mean, standard error and 95% confidence intervals of the attribution scores over the 10,000 iterations are then calculated.

Applicability to level of molecular analysis: This model can be readily applied to ST, 7-locus MLST, rMLST and cgMLST. The method can also be applied to cgSNPs but there can become implementation problems at large numbers of SNPs. In the current project an implementation of 15,000 of the 39,529 SNPs was achieved.

Hald Model

This model was developed in Denmark for the attribution of human salmonellosis (Hald et al., 2004). This “Danish *Salmonella* source attribution” model uses a Bayesian framework with Markov Chain Monte Carlo simulation to attribute sporadic laboratory-confirmed human *Salmonella* infections caused by different *Salmonella* subtypes as a function of the prevalence of these subtypes in animal and food sources and the amount of each food source consumed. The model takes into account the uncertainty for all these factors and also includes travel as a possible risk factor.

This model was improved by (Mullner et al., 2009) to include the introduction of uncertainty in the estimates of source prevalence and an improved strategy for identifiability and is called the “Modified Hald Model”. This is the model that is used here and does not include information on amount of food consumed as is the case for the Dutch model.

In summary, the modified Hald model achieves source attribution by comparing the frequencies of human infections caused by different pathogenic subtypes (e.g. serotypes for *Salmonella* (Mullner et al., 2009)), with the subtype frequencies found in the different sources accounting for potential subtype- and source-dependent characteristics, that may influence their chance to cause human illness (Hald et al., 2004)).

The model utilises a Bayesian approach to estimate and quantify the uncertainty of the parameters.

Briefly,

$$o_i \sim \text{Poisson} \left(\sum_j \lambda_{ij} \right)$$

where o_i is the observed number of human infections caused by subtype i that is assumed to be generated by a Poisson probability distribution, whose mean parameter λ is given by the summation over sources of individual λ_{ij} , which are the Poisson parameters for each subtype i in source j and are given by

$$\lambda_{ij} \sim p_{ij} \times q_i \times a_j$$

where p_{ij} is the prevalence of subtype i in source j , q_i is the subtype-dependent factor, which putatively accounts for differences in survivability, virulence and pathogenicity for subtype i , and a_j is the source-dependent factor, which putatively accounts for the ability of source j to act as a vehicle of listeriosis.

The attribution score to each source j is calculated as follows

$$\text{Score}_j = \frac{\sum_{i=1}^I \lambda_{ij}}{\sum_{i=1}^I \lambda_{ij} \sum_{j=1}^N \sum_{i=1}^I \lambda_{ij}}$$

where I is the number of subtypes and N the number of sources.

According to Mullner et al. (Mullner et al., 2009) the following default priors were used for the above mentioned factors.

(a) Source dependent factor

$$a_j \sim \text{dexp}(0.002)$$

(b) Genotype dependent factor

$$\log(q_i) \sim \text{Normal}(0, \tau),$$

where τ is given by a fairly diffuse $\text{Gamma}(0, 0.01, 0.01)$ distribution.

(c) Prevalence

The priors for the prevalence (p_{ij}) were chosen to be independent beta distributions,

$$p_{ij} \sim \text{dbeta}(\alpha_{ij}, \beta_{ij}),$$

where the parameters α_{ij} and β_{ij} were determined from the posterior distributions of a separate Bayesian analysis of the prevalence data, for each source j and subtype i (Mullner et al., 2009; Mughini-Gras and van Pelt, 2014) (see prevalence sub-model below).

Posterior distributions of the attribution proportions Prop_j in each source j were obtained by a Markov Chain Monte Carlo simulation implemented in WinBUGS1.4 (<http://www.mrc-bsu.cam.ac.uk/software/bugs/>). Five independent Markov chains were run, each using 30,000 iterations (10,000 burn-ins). This was sufficient to provide convergence using the method developed by Gelman and Rubin (Gelman and Rubin, 1992).

Prevalence sub-model:

Briefly, the prevalence was modelled as

$$p_{1ij} \sim \pi_j \times r_{ij},$$

where $\pi_j \sim \text{dbeta}(1, 1)$ is the overall prevalence of subtypes in source j and r_{ij} is the relative frequency of genotype i in source j , which is given by

$$\left(r_{1j}, r_{2j}, \dots, 1 - \sum_{i=1}^{I-1} r_{ij} \right) \sim \text{ddirchlet}(X_{1j}, X_{2j}, \dots, X_{Ij}).$$

Here X_{ij} represents the number of isolates of genotype i in source j (Mughini-Gras & van Pelt, 2014).

The mean values and the standard deviations of the posterior distributions of p_{1ij} were used to calculate α_{ij} and β_{ij} (the parameters of the beta distribution used in the main model) as follows

$$\alpha_{ij} = \left(\frac{\langle p_{1ij} \rangle}{SD_{p_{1ij}}} \right),$$

$$\beta_{ij} = \left(\frac{1}{SD_{p_{1ij}}} \right).$$

Sample size correction and confidence intervals: Sample size correction was not implemented in this model. The summary statistics (mean, standard deviation, median and confidence intervals denoted as 2.5% & 97.5% percentiles) of the attribution proportions were obtained from the posterior distributions of Score_j .

Applicability to level of molecular analysis: This model is only implemented at ST level.

STRUCTURE

This is a Bayesian clustering model designed to infer population STRUCTURE and to attribute individuals to population groups (Pritchard et al., 2000). The program has been used successfully for 7 locus *Campylobacter* MLST genotyping data (Strachan et al., 2013). Each isolate is attributed on the basis of a training dataset consisting of isolates from known populations (i.e. set USEPOPINFO to 1). The algorithm calculates the frequency of each particular sequence type in each population. Based on these frequencies, the probability of an isolate (e.g. a human isolate) belonging to a population group (e.g. source includes fish, bovine, ovine, poultry, swine etc.) is calculated. This is repeated 10,000 times using the Markov Chain Monte Carlo process with 1,000 burn-in steps.

Sample size correction and confidence intervals: The model appears to have no sample size correction within it and it was not possible to implement this either within or outwith the model. The mean and confidence intervals of the scores were calculated as in the Dutch model.

Applicability to level of molecular analysis: This model was implemented at 7 locus MLST, rMLST and 1,748 -cgMLST. However, it was not possible to carry this out for cgSNPs because of the computation time required.

The Asymmetric Island (AI) Model

This source attribution model incorporates a Bayesian approach and uses the allelic profile of the sequence subtypes to reconstruct the genealogical history of the isolates (Wilson et al., 2008). The host populations are considered to exist on separate "islands" (e.g. the sheep island). Mutations and recombination occur on each island. Migrations from between each reservoir (island) into the human population are used to estimate the degree of attribution to each source. This model has previously been applied to *Campylobacter* 7 locus MLST data from England (Wilson et al., 2008), Scotland (Sheppard et al., 2009) and New Zealand (Mullner et al., 2009).

The Asymmetric Island model assigns each human case to the potential source populations on the basis of DNA sequence similarity. It does this by encoding the DNA sequence data for each locus as an allele. By comparing human isolates to a panel of reference sequences of known source (e.g. cattle, sheep, chickens, pigs, wild birds and turkey), each human case can be assigned a probability of originating in each source population (i.e. an attribution score). The source attribution probabilities are calculated using a statistical model of the way the DNA sequences evolve in the populations of bacteria. In the statistical model, there are parameters representing the processes of mutation, DNA exchange between bacteria (recombination or horizontal gene transfer) and zoonotic transmission between populations. These processes lead to differences in gene frequencies between the source populations, facilitating source attribution. This model also uses a MCMC process which was conducted for 100,000 iterations, with the output file written once every 50 iterations. A symmetric Dirichlet (1) prior is used on the proportion of human isolates attributed to sources, in which all sources are considered equally likely a priori (Wilson et al., 2008).

Sample size correction and confidence intervals: The model appears to have no sample size correction within it and it was not possible to implement this either within or outwith the model. The mean and confidence intervals of the scores were calculated as in the Dutch and STRUCTURE models.

Applicability to level of molecular analysis: This model was implemented at 7 locus MLST and rMLST. The program fails to work beyond 250 loci and so higher level analysis was not possible.

The Aberdeen Model

In this method, attribution is based on the similarity between human isolates to isolates from different sources (e.g. fish, bovine, ovine, etc.). An isolate is attributed to the reservoir which has the maximum number of similar loci or SNPs. This is simply calculated by summing the number of loci that are identical. Hence, each of the 254 human isolates used in the study are allocated to a source. For example if 30 are allocated to ovine then the attribution score to the ovine source is $30/254 = 0.12$.

Sample size correction and confidence intervals: sample size was carried out as in the Dutch model as was generation of mean, standard error and 95% confidence intervals for the attribution scores.

Applicability to level of molecular analysis: This model can be readily applied to ST, 7-locus MLST, rMLST and cgMLST and 39,529 SNPs.

7.1.2. Self-Attribution

Self-attribution is a key performance measure for the source attribution models (Sheppard et al., 2009). This is the average percentage accuracy that any given isolate from a source can be correctly attributed back to its own source reservoir (e.g. the likelihood that the attribution model will assign an ovine isolate back to ovine). This can be performed in a number of ways. Here, the attribution is carried out and then all of the source isolates are re-introduced blind to the models and their scores to each source are determined. Average, standard error and confidence intervals are calculated as done for standard source attribution described above. This was carried out for all of the attribution models.

7.1.3. Analyses

Table 7.1 specifies the analyses that were performed based on what was possible to implement with the models. Source attribution analysis was performed for the 5 main sources (fish, swine, ovine, bovine and poultry) as described in Section 6, Table 6.2. However, since poultry was only represented by 25 isolates and the rarefaction results in Section 6 indicate that this only represents a limited proportion of poultry genotypes (at both 7 locus MLST and 30 locus rMLST) additional attribution analysis was carried out with the remaining 4 sources.

Table 7.1.: Source attribution models performed according to level of molecular analysis

Number of sources	Level of molecular analysis (number of loci)	STRUCTURE	Dutch	Asymmetric Island	Hald	Aberdeen
5 sources	ST(1)	nd	nd	nd	√	nd
	MLST(7)	√	√	√	np	√
	rMLST(30)	√	√	√	np	√
	cgMLST(1748)	√	√	np	np	√
	cgSNP(15,000 Dutch, 39,529 Aberdeen)	np	√	np	np	√
4 sources (excluding poultry)	ST(1)	nd	nd	nd	nd	nd
	MLST(7)	√	√	√	np	√
	rMLST(30)	√	√	√	np	√
	cgMLST(1748)	√	√	np	np	√
	cgSNP(15,000 Dutch, 39,529 Aberdeen)	np	√	np	np	√

nd – not done; np – not possible due to software being inoperable above a certain number of loci.

7.2. Results and Discussion

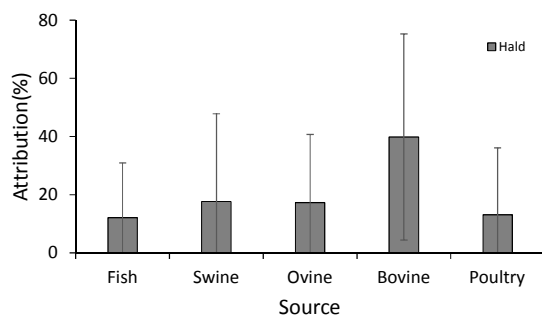
7.2.1. Source Attribution of 5 sources

Single locus ST level

The results of the source attribution model at the level of single locus ST are presented in Figure 7.1. (a). Bovine appears to be the main source that is attributed to human disease (38%). However, the confidence intervals in the model are very large indicating that it is difficult to determine which source is likely to be most important in terms of human infection. It may be that the relatively small sample size of the sources could play a role in this large uncertainty and/or the fact that single locus ST may not be a sufficient discriminating factor. Self-attribution (Sheppard et al., 2009) was used to determine the accuracy of the Hald model Figure 7.1. (b). On average the model was correct 50% of the time, but the problem of large confidence intervals persists, as for the attribution of human isolates.

Only the Hald model was conducted at single locus ST because the software code that has been developed only operates at a single locus. One disadvantage of this model is that if an ST occurs in humans but not in the source reservoirs then it cannot be included in the analysis. There is the potential to develop the Hald model for multiple loci. For example carrying out the analysis independently one locus at a time and then aggregating the results. This would however be computationally intensive and would require all the MCMC chains to converge for each individual locus. All of the other models could be performed at single locus MLST also but it was decided not to do this as it is already generally known that 7 locus MLST has improved performance.

(a)



(b)

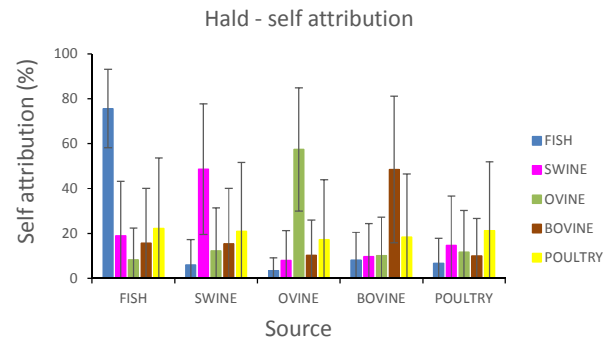


Figure 7.1.: (a) Source attribution of human cases and (b) Self-attribution, using single locus ST data and the Hald model (CIs = 95 percentiles)

7-Locus MLST

Self-attribution (Sheppard et al., 2009) was used to determine the accuracy of the three source attribution methods using the 7 locus MLST data (Figure 7.2.). On average Asymmetric Island model performed best, being correct 80% of the time, with STRUCTURE model next (45.9%) and Aberdeen and Dutch at 41%. This should be compared with what would be expected by chance which is 20% (i.e. 1 isolate partitioned to one of 5 sources). It is worth noting that the confidence intervals indicate that the Asymmetric Island has the smallest whilst the Aberdeen model has the largest.

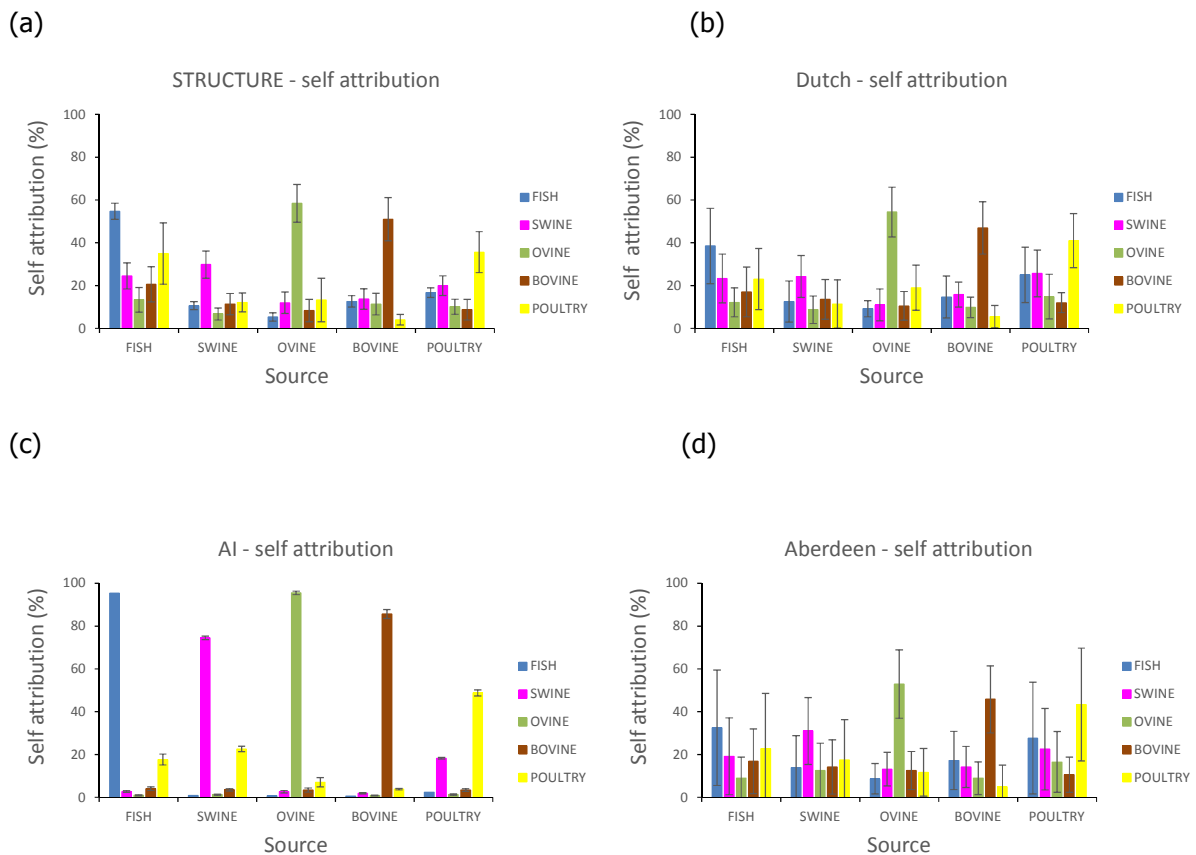


Figure 7.2.: Self attribution of 7 locus MLST data utilising (a) STRUCTURE, (b) Dutch, (c) Asymmetric Island and (d) Aberdeen models (error bars denote 95% confidence intervals)

Source attribution was then carried out using human data (Figure 7.3.). All four models indicated that the most likely source was bovine (38-64%) whilst the remaining of human isolates were shared across the other sources. The AI model confidence intervals are again the smallest and has the highest attribution to bovine (64%) compared to all of the other models.

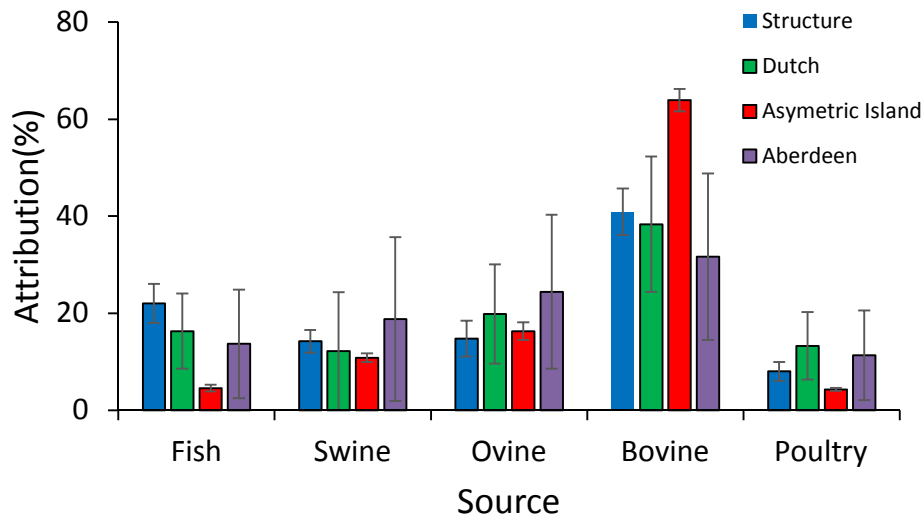


Figure 7.3.: Source attribution of human cases with 7 locus MLST data utilising STRUCTURE, Dutch, Asymmetric Island and Aberdeen models (error bars denote 95% confidence intervals)

30-locus rMLST

The self-attribution for 30-locus rMLST (Figure 7.4.) performs best for AI (80% whilst the other models give similar levels of performance (44% STRUCTURE, 43% Aberdeen and 36% Dutch). Again, the confidence intervals for the AI model are the smallest. The source attribution to human data (Figure 7.5) suggests that for all 4 models the most likely source was bovine (41-59%) whilst the remainder of human cases was shared across the other sources with poultry possibly being the lowest. The error bars are again very small for the AI model and it attributes the highest number of cases, of the four models, to bovine (59%). Since the self-attribution for AI is much higher than the other models it is likely that the higher attribution to bovine is credible.

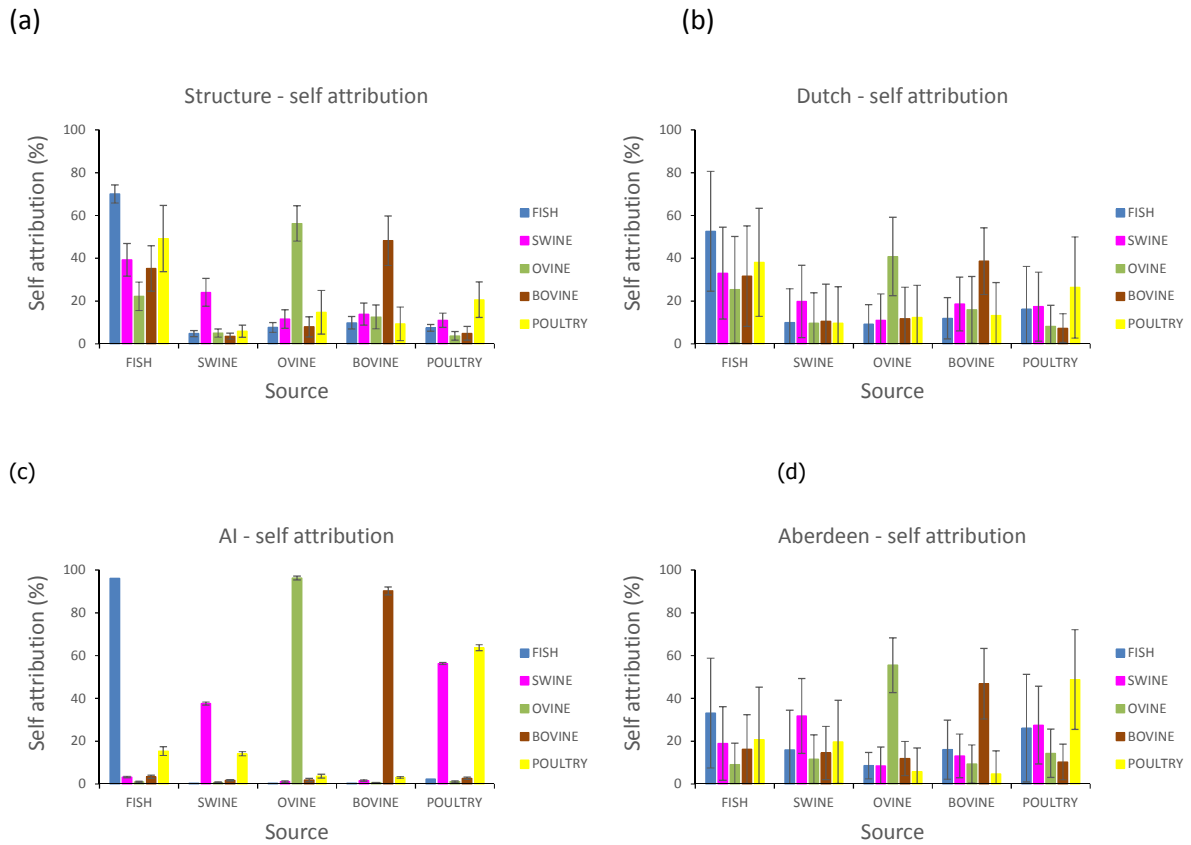


Figure 7.4.: Self attribution of 30 locus rMLST data utilising (a) STRUCTURE, (b) Dutch, (c) Asymmetric Island and (d) Aberdeen models (error bars denote 95% confidence intervals)

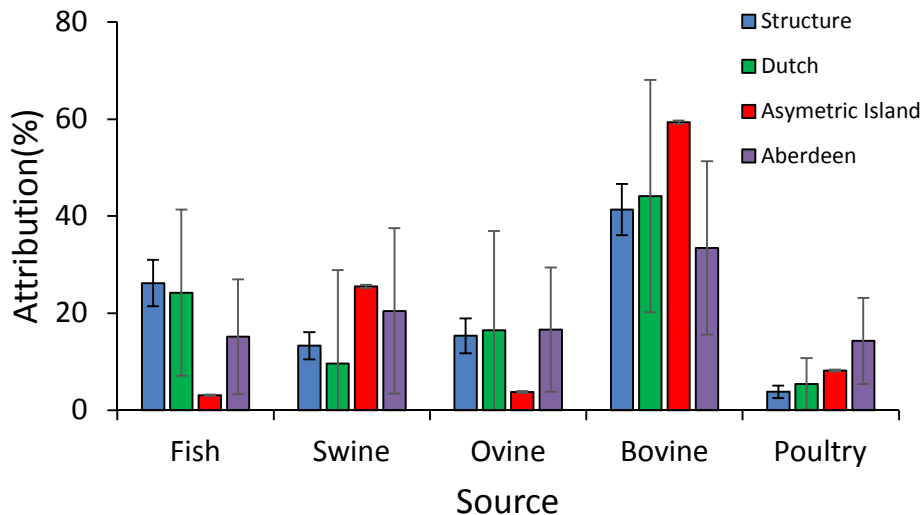


Figure 7.5.: Source attribution of human cases using 30 locus rMLST data and STRUCTURE, Dutch, Asymmetric Island and Aberdeen models (error bars denote 95% confidence intervals)

1748-locus cgMLST

The self-attribution results (Figure 7.6) indicate that the Aberdeen (61%) and STRUCTURE (60%) models perform best whilst the Dutch model performs at 44%. The error bars in the Dutch model appear to be generally larger than for the other models. Source attribution was carried out using human data (Figure 7.7). Although, all three models indicated that the most likely source was bovine (35-42%), this is less obvious than the findings from 7 locus MLST and 30 locus rMLST analyses.

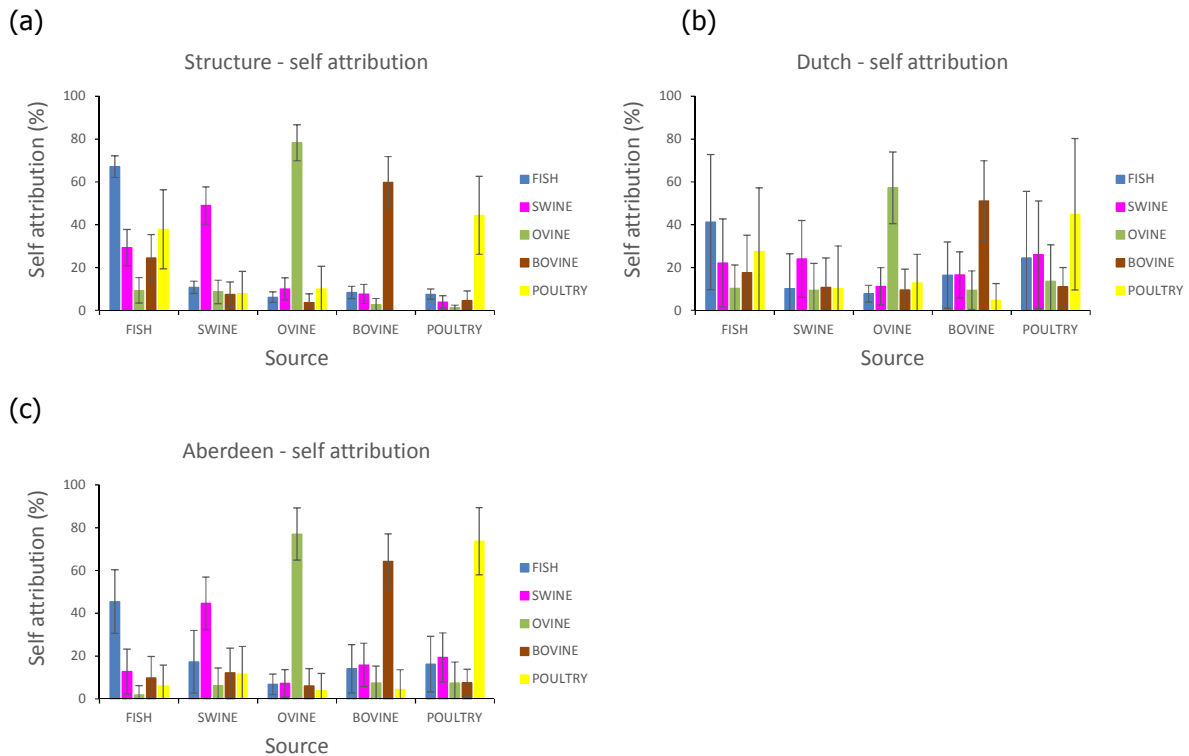


Figure 7.6.: Self attribution of 1748 locus cgMLST data utilising (a) STRUCTURE, (b) Dutch, and (c) Aberdeen models (error bars denote 95% confidence intervals)

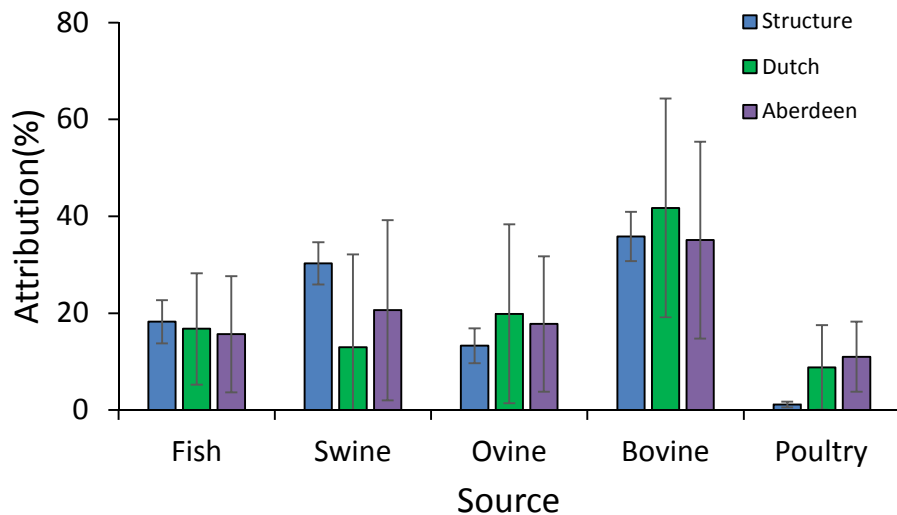


Figure 7.7.: Source attribution of human cases by 1748 locus cgMLST data utilising STRUCTURE, Dutch and Aberdeen models (error bars denote 95% confidence intervals)

39,529/15,000-locus cgSNPs

The Dutch model is less accurate than the Aberdeen model in predicting the origin of the isolates by self-attribution (30% compared with 60%) for the cgSNPs dataset. Note however that by chance the probability to identify the “right” origin of an isolate from 5 sources is 20%, then the Dutch model does considerably better than that. It is also worthy to note that the Dutch model tends to have a bias towards fish (Figure 7.8. (a)). The confidence intervals are also very large in the Dutch model showing that there is a high degree of variation between each of the iterations of the computer model. It should be noted that for computational reasons the Dutch model operated on only 15,000 SNPs and the Aberdeen model on the full 39,529 SNPs. However, the Aberdeen model self-attribution was repeated for 15,000 SNPs and the self-attribution was found to be 62%.

Source attribution was carried out using human data (Figure 7.9). Although in both models the most likely source was bovine (32 - 43%), the high uncertainty of the attribution results (see the size of the confidence intervals in Figure 7.9) suggest that this result is not significant. This could be due to the fact that many of the SNPs are not host associated and this is adding noise in the source attribution calculations.

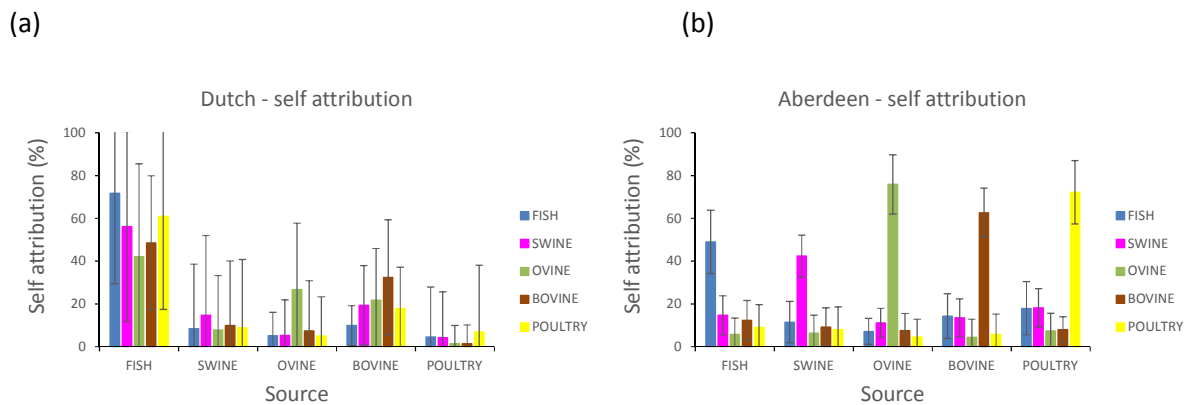


Figure 7.8.: Self attribution of cgSNPs data utilising (a) Dutch (15,000 SNPs), and (b) Aberdeen (39,529 SNPs) models (Error bars denote 95% confidence intervals)

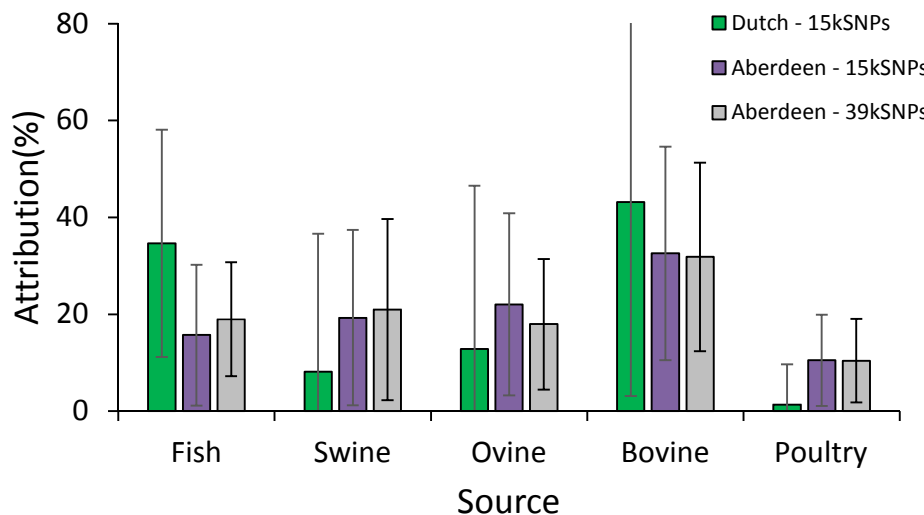


Figure 7.9.: Source attribution of 39,529 cgSNPs data utilising Dutch and Aberdeen models (Note Dutch used 15,000 SNPs after removing isolates with 1 SNP difference between isolate in reference) (error bars denote 95% confidence intervals)

7.2.2. Source Attribution of 4 Sources (Excluding Poultry)

Figure 7.10 shows the source attribution results after poultry have been removed from the analysis. As mentioned above this was done because poultry is represented by only 25 isolates. Further, it is worth noting that in the previous sections human attribution to poultry was generally amongst the lowest of the 5 source reservoirs. The results at the 4 different levels of sub-typing all show that bovine tends to have the highest rates of source attribution (7 locus MLST(35% to 61%), 30 locus rMLST(37% to 57%), cgMLST(33% to 51%), cgSNPs(34% to 55%)). The other 3 reservoir sources exhibit a range of attribution levels depending on the level of sub-typing and the source attribution model.

For 7-locus MLST Self attribution improved for all of the models after poultry was dropped as a source (Figure 7.11). For example by 8% for STRUCTURE, 11% for Dutch, 14% for AI and 13% for Aberdeen. AI has the highest self-attribution score of 94%.

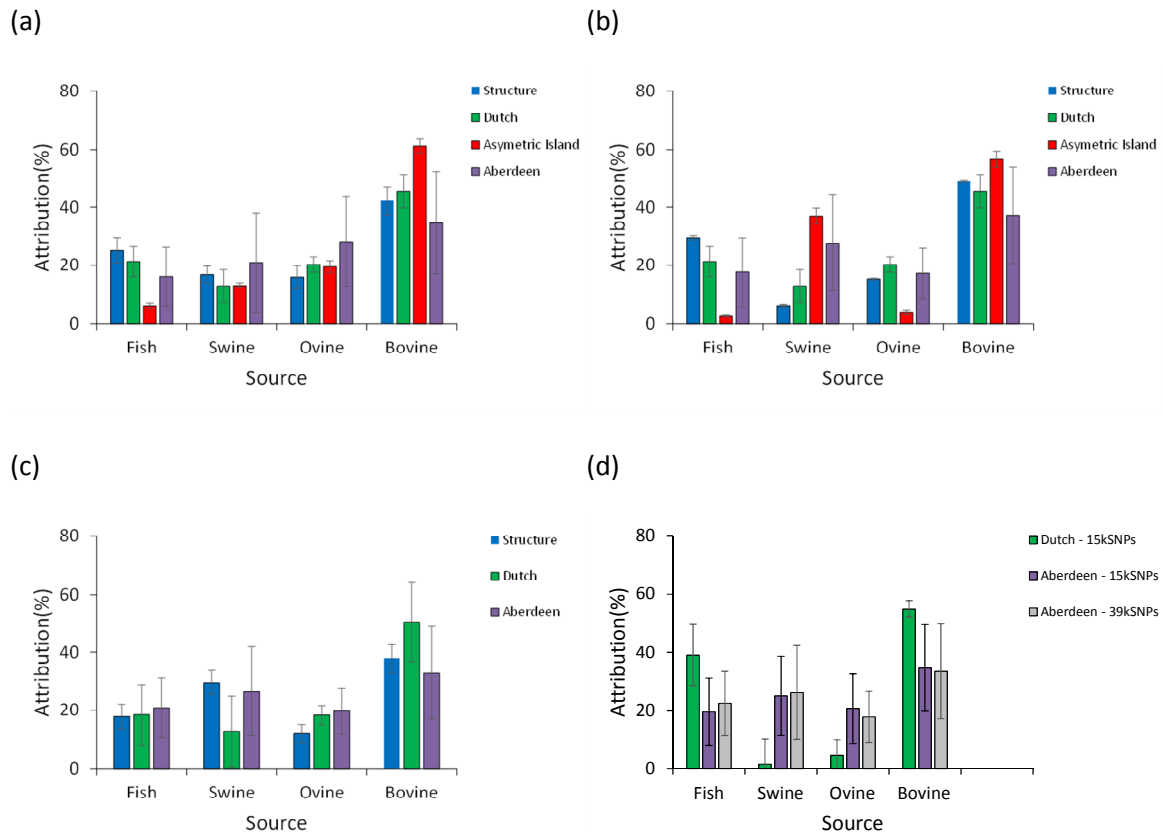


Figure 7.10.: Source attribution excluding poultry at (a) 7 locus MLST, (b) 30 locus rMLST, (c) 1,748 locus cgMLST and (d) 39,529 cgSNPs (error bars denote 95% confidence intervals)

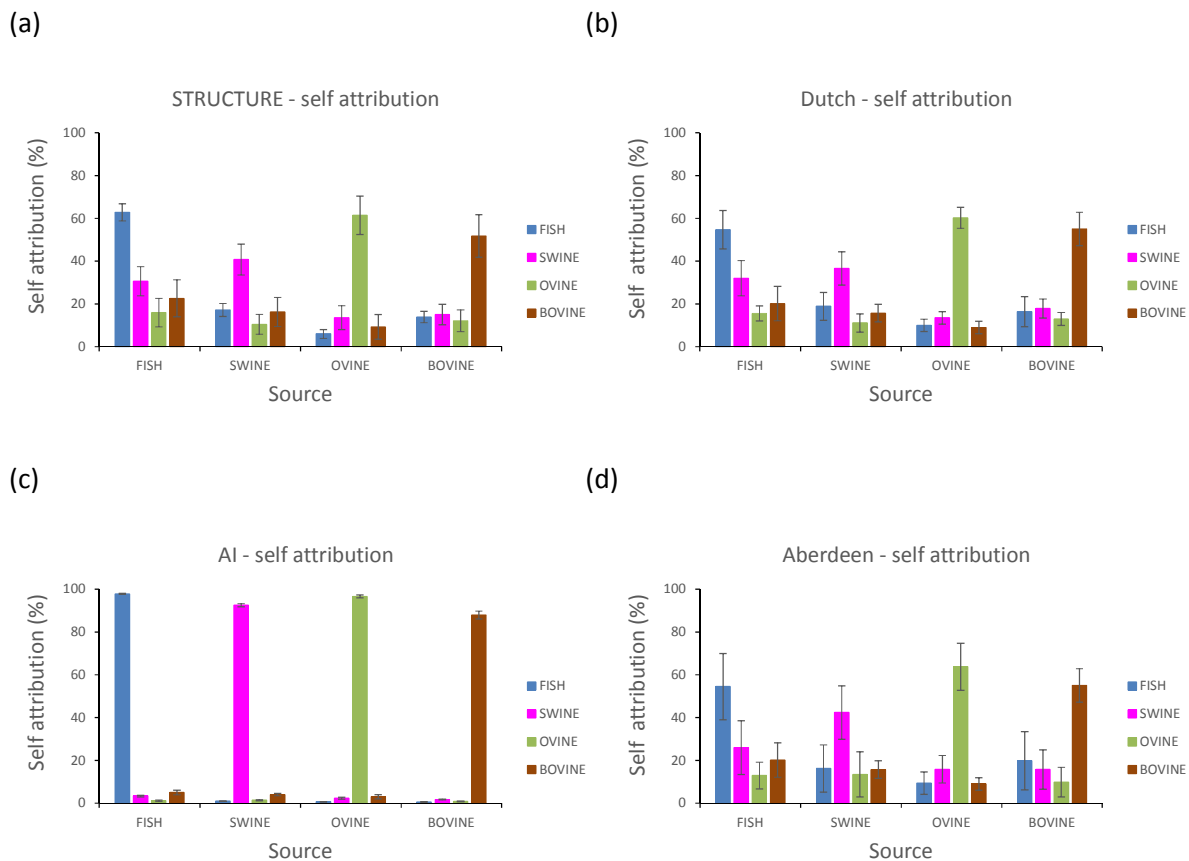


Figure 7.11.: Self-attribution of 7 locus MLST data, excluding poultry, utilising (a) STRUCTURE, (b) Dutch, (c) Asymmetric Island and (d) Aberdeen models (error bars denote 95% confidence intervals)

For 30-locus rMLST Self attribution improved for all of the models after poultry was dropped as a source (Figure 7.12.). For example by 8% for STRUCTURE, 6% for Dutch, 15% for AI and 13% for Aberdeen. AI has the highest self-attribution score of 95%.

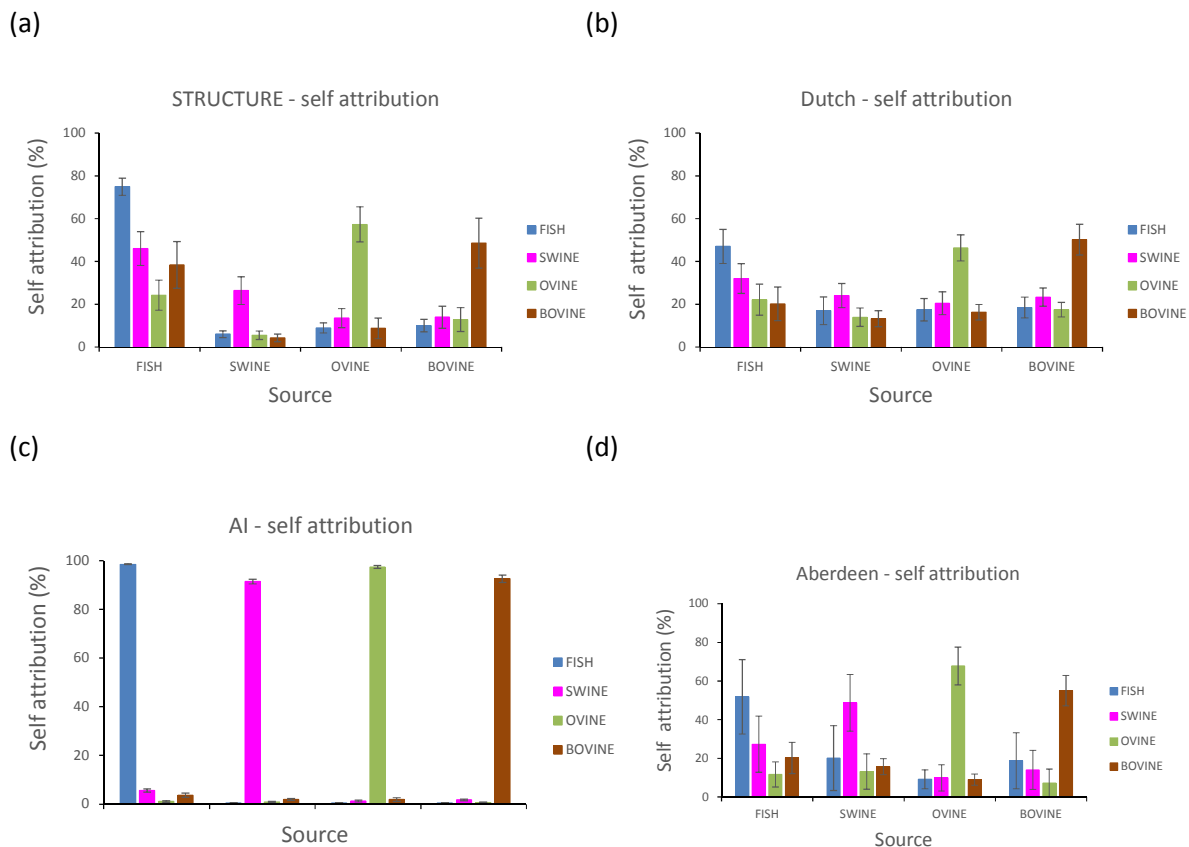


Figure 7.12.: Self-attribution of 30 locus rMLST data, excluding poultry, utilising (a) STRUCTURE, (b) Dutch, (c) Asymmetric Island and (d) Aberdeen models (error bars denote 95% confidence intervals)

For 1,748-locus cgMLST self-attribution improved for all of the models after poultry was dropped as a source (Figure 7.13), for example by 5% for STRUCTURE, 12% for Dutch and 10% for Aberdeen. The Aberdeen model has the highest self-attribution score of 71% followed by STRUCTURE (65%) and Dutch (56%).

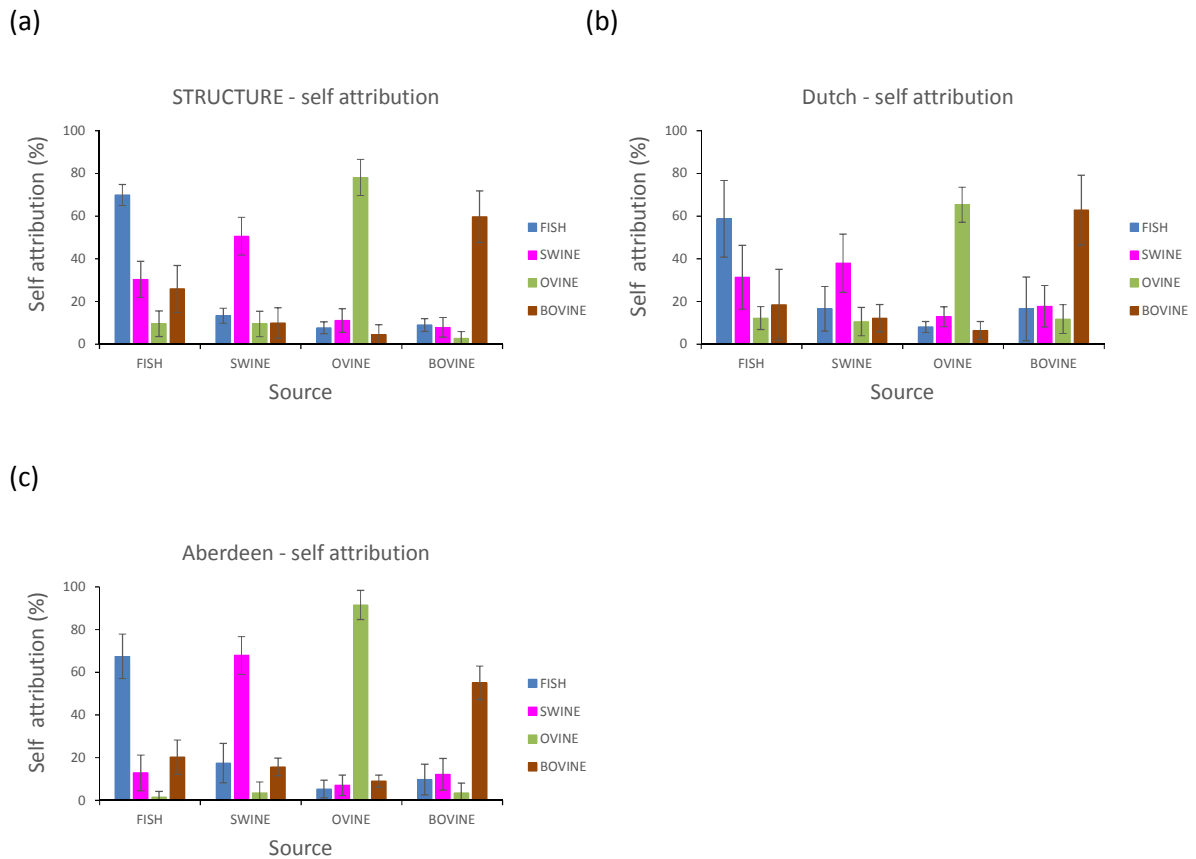


Figure 7.13.: Self attribution of 1748 locus cgMLST data, excluding poultry, utilising (a) STRUCTURE, (b) Dutch, and (c) Aberdeen models

For cgSNPs self-attribution improved for the two models with the Dutch at 38% and the Aberdeen model at 66% (Figure 7.14). The Dutch model appears to have a bias towards fish.

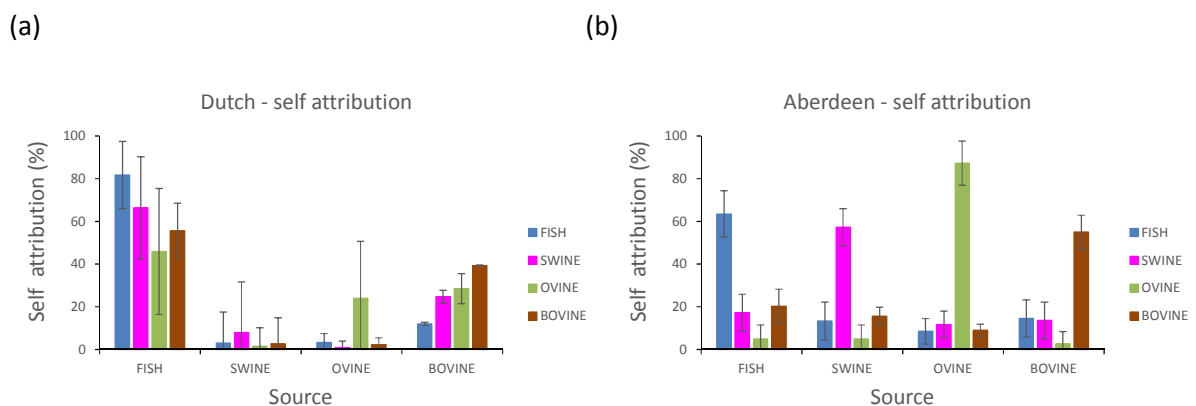


Figure 7.14.: Self attribution of cgSNPs data, excluding poultry, utilising (a) Dutch (15,000 cgSNPs), and (b) Aberdeen (39,529 cgSNPs) models (error bars denote 95% confidence intervals)

7.2.3. Discussion

All of the models at all of the different levels of molecular analysis tended to place bovine as the main source of human listeriosis (32% to 64%). The Dutch model was able to perform at all levels of molecular analysis but was limited to 15,000 SNPs because of its current implementation in VisualBasic. This limitation has the potential of being resolved by the software being further improved or written using another programming platform. The Dutch model tended to have low self-attribution compared with most of the other models and increasing the level of loci did not really improve its performance. The Hald model was of limited utility since it could only be applied at the level of ST and was unable to incorporate sequence types that were found in humans but not in the animal sources. STRUCTURE was operational up to the level of cgMLST but cannot be currently implemented for >15,000 loci as required in the cgSNP analysis. Although the AI model was operational up to 30 rMLST only, it had the highest self-attribution and tightest confidence intervals and also gave the highest source attribution to bovine. The Asymmetric Island model incorporates recombination and mutation. The model appears to be fairly complicated and the current explanations of its operation are difficult to comprehend. The newly developed Aberdeen model, which operated at all levels of molecular analysis, was relatively easy to implement and was not computationally intensive. Its self-attribution performance was similar to a number of the other models.

Those models that were able to operate at the whole genome level (cgMLST and cgSNP) did not appear to show improved performance from fewer loci. It is likely that a large number of the loci are not host related and this may add noise to the analysis. It may be best to pre-select loci for host specificity prior to source attribution. Methods need to be developed to achieve this in an unbiased way.

The number of genomes available for some of the sources was relatively small. In the source attribution analysis performed for 5 sources, poultry had only 25 genomes. The results indicate that this number is probably too small and this is also seen by the relatively wide confidence intervals. Most other published studies for source attribution tend to have at least 50, if not 100 representatives for each source. In an analysis for campylobacter it was reported (Smid et al., 2013) that it was preferable to have at least 100 isolates per source and the data presented showed that using 25 isolates gave a large uncertainty in the self-attribution scores as is being found in the current study. The source attribution results for 4 sources appear to be better with higher self-attribution scores and models producing tighter confidence intervals. Ideally selection of isolates for source attribution should include contemporaneous sampling of isolates from sources and humans from a fixed geographic area. In the current study, the geographic area was very broad (much of Europe) and a fairly broad sampling timeframe with source isolates not being uniformly distributed across Member States. Hence, the analysis should be treated with caution as there is the potential of bias.

None of the models applied utilised human consumption data. Only the Hald model has been made operational to do this should these data become available. However the Hald model only works at the level of ST (i.e. for a single locus). The other models provide a probability that an isolate comes from a given source. So potentially all human clinical isolates from a country (or the EU) can be assigned probabilistically to a source and this can be summed up to determine the likely number of cases associated with each source. Then knowing, the exposure (i.e. amount of meals consumed associated with each source) this can then be used to determine the average risk per meal. However, this is simplistic as there will be a lot of different types of meals (some posing greater potential risk than others) and also there will be variation in susceptibility of the host (immune-compromised compared with healthy). This is an area for future research and can potentially be linked to work utilising quantitative risk assessment.

7.3. Conclusions

Source attribution was applied utilising 5 models (Hald (ST only), Dutch (up to 15,000 cgSNPs), STRUCTURE (up to 1748 cgMLST), Asymmetric Island (up to 30 rMLST) and Aberdeen (up to 39,529 cgSNPs) for 5 sources (fish, swine, ovine, bovine and poultry) and 4 sources (removing

poultry). All of the models showed bovine as the main source of human disease (32% - 64% for 5 sources and 33%-61% for 4 sources) but for a number of the models there were broad confidence intervals. It was not possible to rank the relative order of importance for the other sources due to the variation in model outputs (i.e. overlapping confidence intervals). The confidence intervals were reduced when the poultry source was removed because of its small sample size. For all sources, isolates from different parts of the food chain had to be combined to produce a sufficient dataset on which source attribution could be performed. It is possible that the genetic distribution of isolates associated with a particular source may change along the food chain and that this could affect the source attribution results. This is an area worthy of future investigation.

The Asymmetric Island model, which was operational at only 7 locus MLST and 30 locus rMLST, had the highest self-attribution (>80%), had the smallest confidence intervals, and had the largest attribution to bovine (57% - 64%). The AI model therefore appears to provide the most robust results for this dataset. However, it should be noted that the dataset used here may be biased (e.g. due to non-uniform sampling across Europe) and would likely be more robust using larger, well-structured datasets. The AI model is not yet operational at the cgMLST and cgSNP level. Implementation and computational requirements of the models became more difficult the greater the number of loci that were being processed. Currently there appeared to be no great advantage in carrying out attribution at the highest levels of molecular resolution (i.e. cgMLST and cgSNP). New approaches are required to select the data from across the genome to be used in the source attribution model. This is because a number of the loci/SNPs are not informative about the source and appear to add noise to the attribution results. There is also the potential for future research to link the source attribution results to the risk from consuming a meal and quantitative risk assessment.

8. Epidemiological relationship – linking of genetically related isolates

Several recent studies demonstrated the added value of WGS for outbreak investigations by confirming and/or discriminating food and human isolates (Gillesberg Lassen et al., 2016; Jackson et al., 2016; Kvistholm Jensen et al., 2016). In a context where international surveillance is increasing (Paquet et al., 2005; Swaminathan et al., 2006), we intended to check the interest of WGS along with epidemiological information of the food and clinical isolate to assess, retrospectively, relationships between circulating strains of *L. monocytogenes* in EU within 2010-2012 period.

Within each clonal complex of *L. monocytogenes* the comparison of the isolate's full genome to an appropriate reference genome helped to identify individual nucleotide differences (single nucleotide polymorphisms or SNPs). SNP differences were used to identify clusters of clinical strains and food isolates. Clusters of interest were further investigated by focusing on metadata associated to each strain (geographical information, timeline and isolation context).

8.1. Methods

8.1.1. Definition of genetically clustered strains

The methodology used to determine clustering of strains was based on SNP. cgMLST has proved to be also efficient in cluster definition (Section 5). But SNPs bring currently the highest available discriminatory information for determining genetic links between strains.

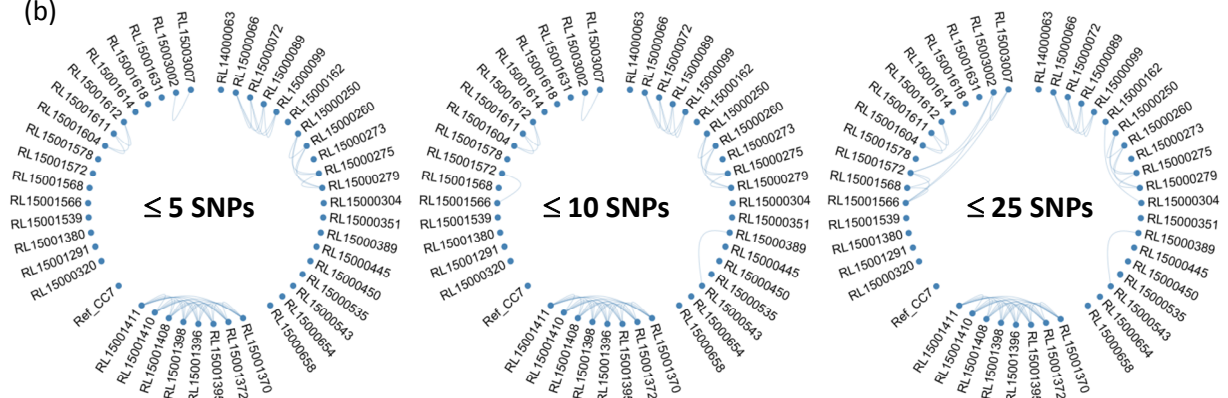
Within each CC, SNP pairwise distances were used to assess the genetic link between *L. monocytogenes* strains isolated in food and strains linked to sporadic cases. Figure 8.1. -a shows, the pairwise distance matrix for strains belonging to CC7. Some strains have less than 5 SNPs of difference (Figure 8.1. -b). With larger SNPs difference (e.g. 10 SNPs), more links can be established. The limit for defining genetically related strains was set to 25 SNPs according to detailed SNPs pairwise distance observed during the retrospective analysis of known outbreaks (Section 5). Figure 8.2 shows congruent clustering of strains according to maximum likelihood phylogeny established on

SNPs or networks established on SNP distance. For such defined genetically related strains, information on time of isolation and geography were used to retrospectively explore the links between isolates.

(a)

	Ref_CC7	RL15000089	RL15000099	RL15000072	RL15000066	RL15001370	RL15000250	RL15000351	RL15000260	RL15000275	RL15000279
Ref_CC7		268	274	162	208	429	190	585	198	194	...
RL15000089	268		0	1	2	266	242	251	243	241	
RL15000099	274	0		1	2	264	249	261	250	248	
RL15000072	162	1	1		3	172	174	178	175	174	
RL15000066	208	2	2	3		218	203	212	204	203	
RL15001370	429	266	264	172	218		248	410	256	253	
RL15000250	190	242	249	174	203	248		187	1	0	
RL15000351	585	251	261	178	212	410	187		188	186	
RL15000260	198	243	250	175	204	256	1	188		1	
RL15000275	194	241	248	174	203	253	0	186	1		
RL15000279	...										

(b)

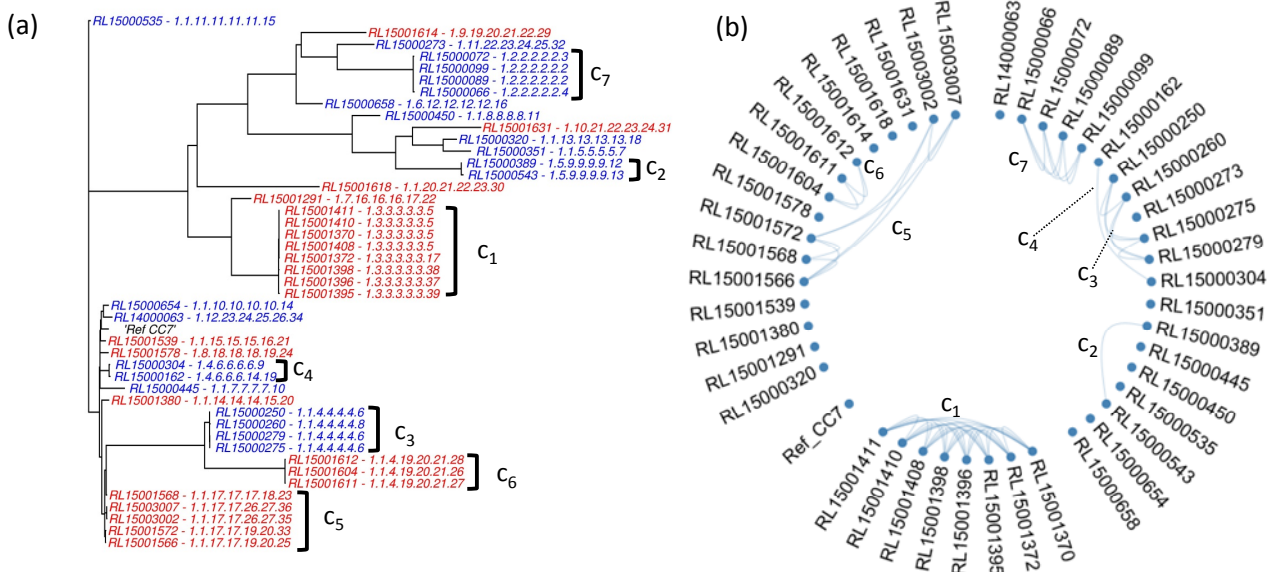


Ref_CC7 strain is public genome used for SNP calling. Upper left: sporadic isolates, lower bottom: isolates of known CC outbreak, upper right: food isolates. Links are established between isolates if the SNP distance is lower or equal to the threshold value indicated in the centre of circle.

Figure 8.1.: (a) Pairwise SNP distance matrix for CC7 stains (only first row column is shown). (b) Relation between the 44 food, sporadic and outbreak strains of CC7

8.1.2. R packages and software

Distribution of SNP distance graph was generated with R software 3.2.4 and ggplot2 package. Links of isolates within a CC, were investigated and summarized with a circular network figure produced with functions of edgebundleR (Bostock et al., 2016) and igraph (Csardi and Nepusz, 2006) packages. Maximum likelihood phylogenies were produced with RAxML, using the model GAMMALG, with 5,000 bootstrap replicates (Stamatakis, 2014).



Upper left: sporadic isolates, lower left: isolates of outbreak, right: food isolates. Links established for isolates according to distances ≤ 25 SNPs. The cluster numbers match with those listed in Table 8.1.

Figure 8.2.: (a) ML phylogeny of 45 strains of CC7. 7 clusters established on pairwise SNP distance are presented. (b) Relation between the 45 food, sporadic and outbreak strains of CC7

8.2. Results

The links between isolates were established for 21 CCs (CC1, CC2, CC3, CC4, CC5, CC6, CC7, CC8, CC9, CC11, CC14, CC31, CC37, CC59, CC87, CC101, CC121, CC155, CC204, CC220, CC415). For these 21 CCs, 151 "clusters" were identified according to SNP pairwise distances. Amongst them, 27 clusters were expected as they exclusively include strains isolated in the same context (like strains belonging to the same outbreak or strain isolated in the same factory). These clusters are not presented. Table 8.1 shows the 124 'unexpected' clusters that were identified based according to SNP pairwise distance.

Table 8.1.: List of *a priori* non expected clusters, established according to the pairwise SNP distance between all strains of the project

CC	Cluster	Outbreak	Sporadic	Food baseline survey	Food other	Food category*
CC1	cluster 1	B(8)	B(1)			
CC1	cluster 10		C(2)			
CC1	cluster 3				V(3)	cheese
CC1	cluster 4		X(7)		X(2)	cheese
CC1	cluster 5		A(1)	U(1)		smoked and gravad fish
CC1	cluster 6		T(2)			
CC1	cluster 7		Q(2)			
CC1	cluster 8		B(2)			
CC1	cluster 9		W(1),C(1)			
CC101	cluster 10		Q(2)			
CC101	cluster 11		Q(5)			
CC101	cluster 7		C(2)	C(1),E(1)	C(2)	smoked and gravad fish, dairy
CC101	cluster 8		F(1)	F(1)		smoked and gravad fish
CC101	cluster 9		A(1)	A(1)		smoked and gravad fish
CC11	cluster 1				A(2)	RTE meat
CC11	cluster 2		W(1)		Z(1)	RTE meat
CC121	cluster 1		C(1)	C(3), Q(1)		smoked and gravad fish
CC121	cluster 10			P(1)	B(3)	smoked and gravad fish
CC121	cluster 12			D(2)		smoked and gravad fish

CC	Cluster	Outbreak	Sporadic	Food baseline survey	Food other	Food category*
CC121	cluster 13			L(1),T(2),H(1)		smoked and gravad fish
CC121	cluster 14			L(2)		RTE meat
CC121	cluster 15			L(3)		smoked and gravad fish
CC121	cluster 16			L(2),X(1)		smoked and gravad fish
CC121	cluster 17			W(2)		smoked and gravad fish
CC121	cluster 18			Q(1)	B(2)	smoked and gravad fish
CC121	cluster 19			K(2)		smoked and gravad fish
CC121	cluster 2		T(2)	U(2),L(1),T(2),H(1)		smoked and gravad fish
CC121	cluster 20			J(5)		RTE meat(3), smoked and gravad fish(2)
CC121	cluster 23				C(5)	RTE meat
CC121	cluster 24				A(2)	RTE meat
CC121	cluster 25				A(1),J(1)	RTE meat
CC121	cluster 26				X(2)	RTE meat
CC121	cluster 27				B(2)	FPE, RTE meat
CC121	cluster 3			C(2)		smoked and gravad fish
CC121	cluster 4			C(4),S(1), E(1), A(1),P(1),J(1),L(2),W(1),Q(1)		smoked and gravad fish
CC121	cluster 5			F(3)		smoked and gravad fish
CC121	cluster 6			N(2)		smoked and gravad fish
CC121	cluster 7			B(2)		smoked and gravad fish
CC121	cluster 8			U(1),L(7)		smoked and gravad fish
CC121	cluster 9			A(1),W(1)		smoked and gravad fish
CC14	cluster 1	X(13)	X(1)			
CC14	cluster 2		T(3)			
CC14	cluster 3			U(2)		smoked and gravad fish
CC14	cluster 4			U(2)		smoked and gravad fish
CC14	cluster 5				V(2)	cheese
CC14	cluster 6				V(4)	cheese
CC155	cluster 2		X(2)	LA(1),N(1)		smoked and gravad fish
CC155	cluster 3		W(1),A(1)	F(1),U(2),W(1),K(2),J(1)		smoked and gravad fish
CC155	cluster 4		A(1)	D(1)	F(2)	smoked and gravad fish, cheese
CC155	cluster 5			N(2),U(6),H(2)		smoked and gravad fish
CC155	cluster 6			U(2)		smoked and gravad fish
CC155	cluster 7				X(5)	RTE meat
CC2	cluster 3		F(2)			
CC2	cluster 4		A(1)	LT(1)		
CC2	cluster 5				B(2)	FPE (1), Vegetable(1)
CC2	cluster 6			X(3)		RTE meat(3)
CC2	cluster 7				C(2)	RTE meat(2)
CC2	cluster 9				U(1),W(1)	smoked and gravad fish(1), RTE meat(1)
CC2	cluster 1		A(4)			
CC2	cluster 2		W(2)			
CC204	cluster 1			C(4)		smoked and gravad fish
CC204	cluster 2			G(1)	C(1),G(1),B(1)	RTE meat(3), FPE(1)
CC204	cluster 3			H(2)		cheese
CC204	cluster 4				B(2)	RTE meat(2)
CC3	cluster 3			F(2)		smoked and gravad fish
CC3	cluster 4			U(6)		smoked and gravad fish (5), RTE meat(1)
CC3	cluster 5		F(2)	F(1)		smoked and gravad fish
CC31	cluster 1				X(2)	RTE meat
CC31	cluster 2			B(1)	G(1),B(1)	smoked and gravad fish, RTE meat, FPE
CC31	cluster 3			A(1)	A(1)	RTE meat
CC31	cluster 4			D(1),Q(2)		cheese
CC31	cluster 5				C(5)	RTE meat
CC37	cluster 2				B(4),B(1),C(1)	FPE, cheese, RTE meat
CC4	cluster 1	C(16)	C(1)			
CC5	cluster 2			C(1)	C(2)	RTE meat, smoked and gravad fish
CC5	cluster 4				V(2)	RTE meat
CC59	cluster 1		A(1)	Q(5)	C(1)	smoked and gravad fish, dairy

CC	Cluster	Outbreak	Sporadic	Food baseline survey	Food other	Food category*
CC59	cluster 2		B(1)		B(1)	cheese
CC6	cluster 1				C(6)	cheese
CC6	cluster 10		Z(1)	Z(1)		smoked and gravad fish
CC6	cluster 2			Z(1)	C(1)	RTE meat, smoked and gravad fish
CC6	cluster 3				C(2)	cheese
CC6	cluster 4			X(2)		smoked and gravad fish
CC6	cluster 5				B(2)	FPE, Vegetables
CC6	cluster 6			N(2),U(1),A(1),W(1)		smoked and gravad fish
CC6	cluster 7		A(2)		A(1)	RTE meat
CC6	cluster 8		A(1),C(1), D(1)	Q(1)		smoked and gravad fish
CC6	cluster 9		Z(1)		B(1)	FPE
CC7	cluster 3			Q(4)		smoked and gravad fish
CC7	cluster 4			C(2)		smoked and gravad fish
CC7	cluster 5		W(2),T(3)			
CC7	cluster 6		X(3)			
CC7	cluster 7			U(4)		smoked and gravad fish, RTE meat
CC8	cluster 10			U(1),W(2),W(1),A(1)		smoked and gravad fish
CC8	cluster 11			K(2)		smoked and gravad fish
CC8	cluster 12			L(2)		smoked and gravad fish
CC8	cluster 13			U(1),Q(1)		smoked and gravad fish
CC8	cluster 3		Z(1)	Z(1)		RTE meat
CC8	cluster 4		Z(1),X(1)		C(7)	cheese, RTE meat
CC8	cluster 5		T(1)	K(6)		smoked and gravad fish
CC8	cluster 6		W(1)	U(7),W(2),L(1),Q(2),W(1)		smoked and gravad fish
CC8	cluster 7			C(1),X(1)	B(1)	smoked and gravad fish (2), FPE(1)
CC8	cluster 8			B(3)		cheese, FPE
CC8	cluster 9			S(1),W(1),J(1)		smoked and gravad fish
CC8	cluster 1		T(1),A(1)			
CC8	cluster 2		W(2)			
CC87	cluster 1	X(13+6)	X(2)			
CC87	cluster 2			Q(6)		smoked and gravad fish
CC87	cluster 3		X(1)		V(1)	RTE meat
CC9	cluster 10			L(2)		smoked and gravad fish
CC9	cluster 11				Z(3)	RTE meat
CC9	cluster 12				A(2)	RTE meat
CC9	cluster 13				A(3)	RTE meat
CC9	cluster 14				A(2)	RTE meat
CC9	cluster 15			X(2)		smoked and gravad fish
CC9	cluster 16			X(2)		smoked and gravad fish
CC9	cluster 3		Z(1)	Z(3)		RTE meat
CC9	cluster 4		A(1)	U(3),D(5)		smoked and gravad fish
CC9	cluster 7			C(2),Z(2),Q(1)		smoked and gravad fish
CC9	cluster 8			N(1),U(7),A(1),V(1)		smoked and gravad fish
CC9	cluster 9			Z(2)		smoked and gravad fish
CC9	cluster 1		A(3)			
CC9	cluster 2		X(1)		C(1)	RTE meat
CC98	cluster 1		W(1),Q(1)			

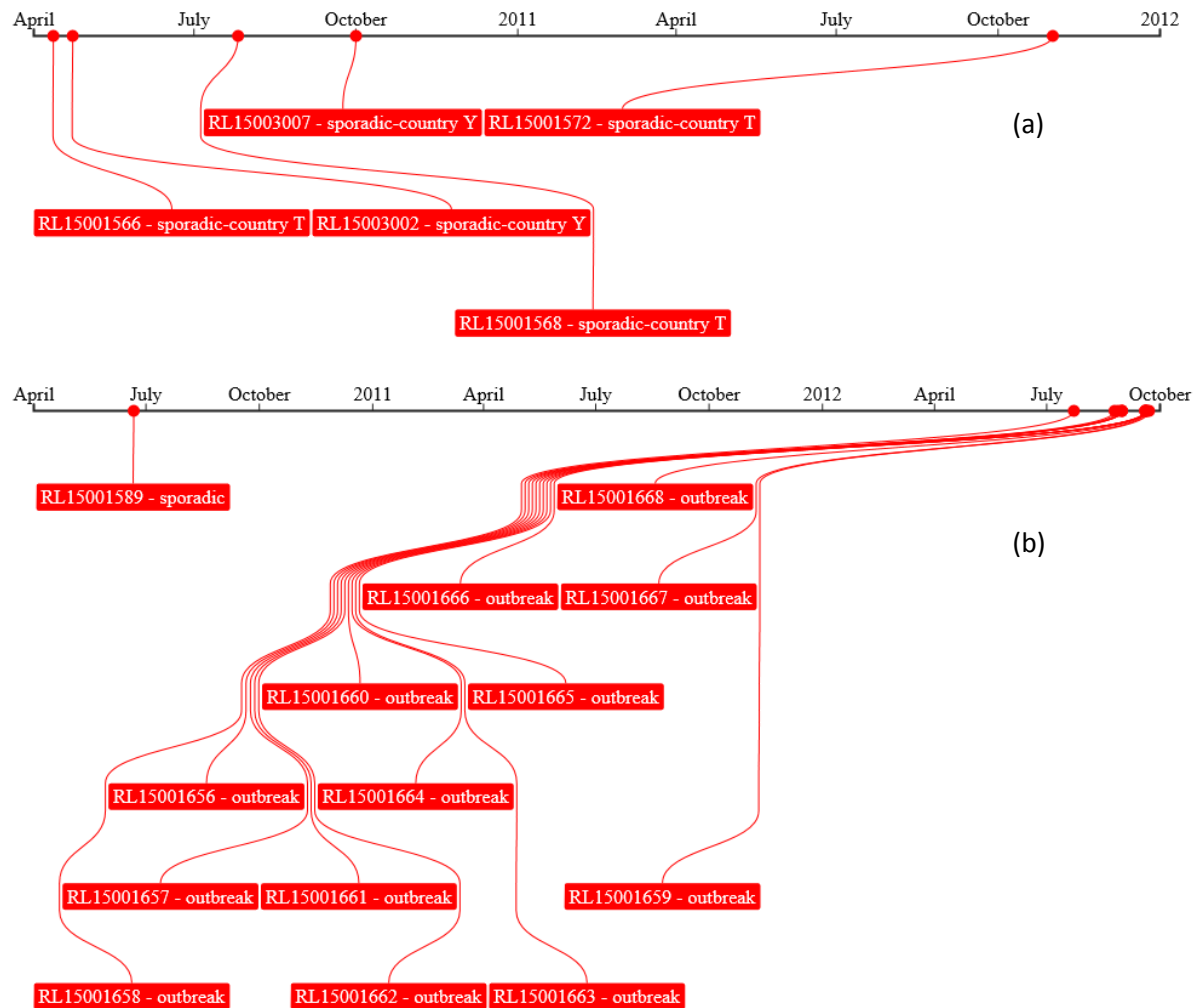
* FPE: food processing environment

Note: Letter identified code of countries, number in parenthesis indicates number of strains.

8.2.1. Epidemiological analysis of genetically clustered strains: link between human sporadic strains and potential relation with food strains

Forty-eight clusters out of the 124 included one or several sporadic human strains (representing a total 91 sporadic human cases). For 17 out of these clusters, only human sporadic strains were related (see e.g. cluster 5 of CC7 Figure 8.3. -a). Additionally, it was revealed that sporadic human

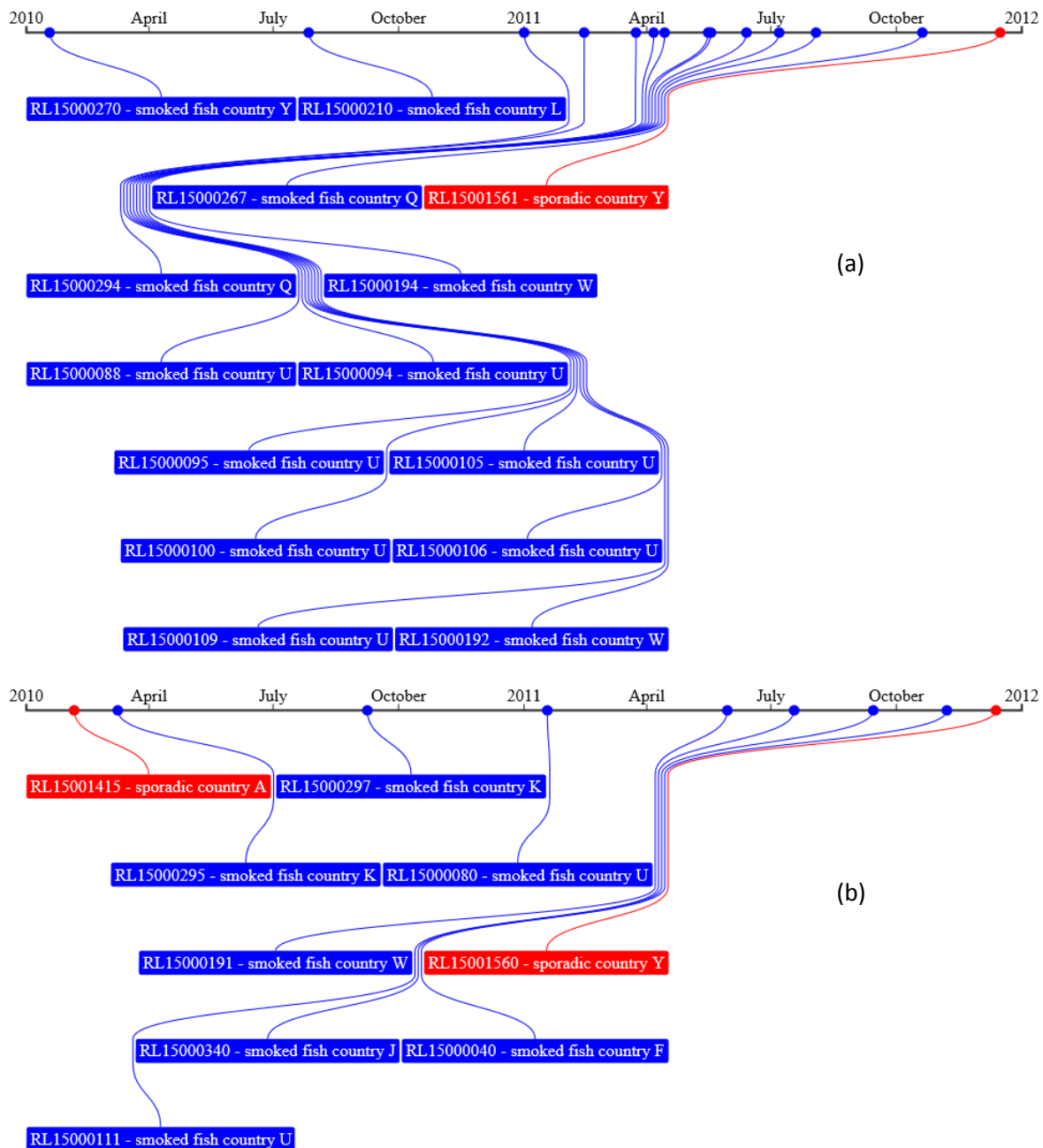
cases were related to four of the outbreaks studied in Section 5 (see e.g. Figure 8.3. -b). It is worth to notice that these sporadic strains were observed in the same country where the outbreak occurred.



Red colour is associated to strains isolated in sporadic cases or in an outbreak context.

Figure 8.3.: (a) Timeline and countries of a cluster (cluster 6 of CC7) associating sporadic strains. (b) Timeline of a sporadic strain genetically closely related to CC14 outbreak (sporadic and outbreak strains are from the same country, that is X)

For the 27 other clusters, at least one strain isolated from food (either from baseline survey, or other strains from national active or passive surveillance) was involved, potentially relating sporadic human cases to contemporary food isolates that circulate in EU. Although the three categories of RTE food products, that is smoked and gravad fish (Figure 8.4), cheese and RTE meat (Figure 8.5), were involved, most (16) of the clusters were related to smoked fish.

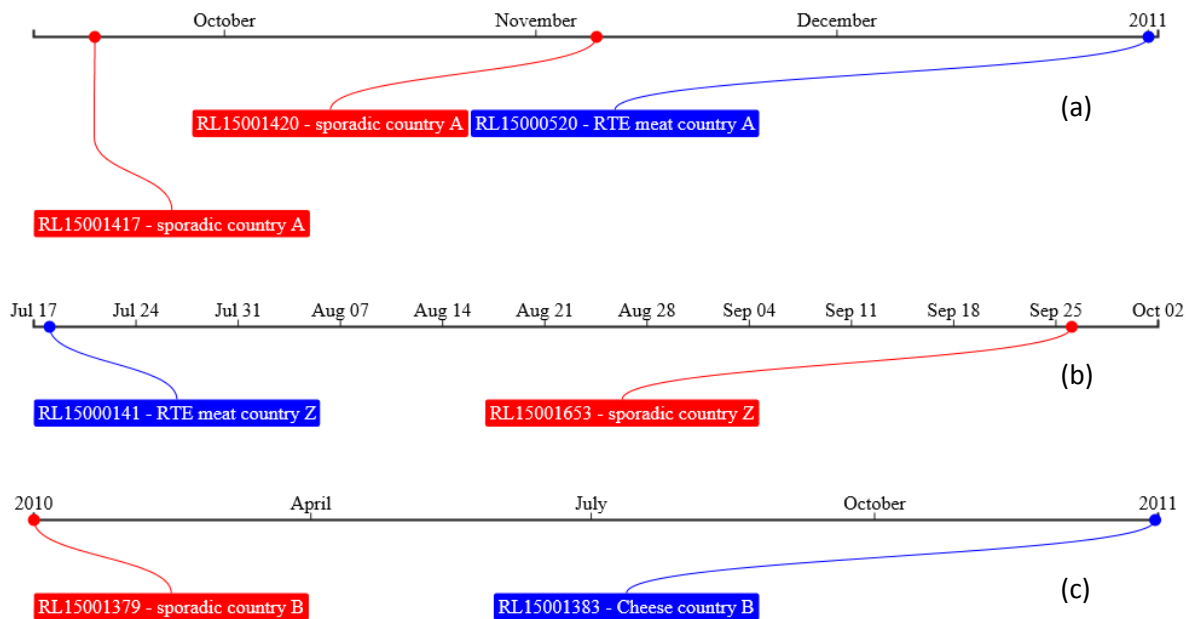


Red colour is associated to strains isolated in sporadic cases or in an outbreak context. Blue colour is associated to food isolates.

Figure 8.4.: Two clusters observed including sporadic human cases and strains isolated in smoked salmon category from baseline survey (a) for CC8 cluster 6, (b) for CC155 cluster 3

It confirms that this smoked and gravad fish food category is of concern for the risk of listeriosis (Pouillot et al., 2009; Tocmo et al., 2014). Yet, it cannot be concluded that sporadic cases are most likely linked to this type of product as a majority of strains we matched against belonged to the smoked and gravad fish category. Moreover, the majority of strains for cheese and RTE meat categories came from passive national surveillance of a more limited number of countries. The strains of these categories just matched less to real exposition of consumers than do the baseline survey

strains for smoked salmon. Difference in exposure could be better approached with quantitative microbial risk assessment as this approach takes into account food exposure.

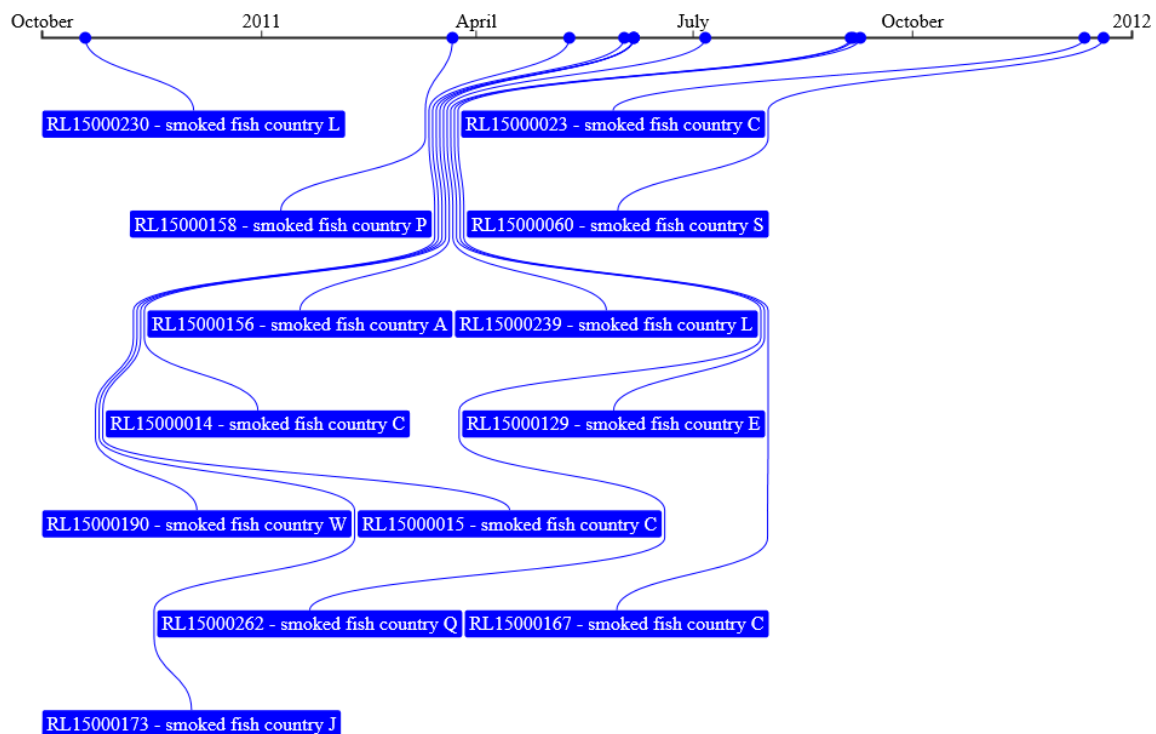


Strain isolated from RTE meat strain of CC8 cluster 2 was isolated during the baseline survey in 2011. Red colour is associated to strains isolated in sporadic cases or in an outbreak context. Blue colour is associated to food isolates.

Figure 8.5.: Timeline and countries implicating sporadic and RTE meat strains for (a) CC6 cluster 7 and (b) CC8 cluster 3 and for (c) cheese CC59 cluster 2

8.2.2. Geographical and temporal widespread of genetically clustered strains

Seventy-six clusters were established for food strains, i.e. not including human strains. The analysis of these clusters revealed that strains circulated in several countries as 21 clusters involved from two up to nine countries for cluster 4 of CC121 (Figure 8.6). This European circulation of strains is particularly obvious for smoked and gravad fish category. As for the attribution of sporadic infections, it cannot be inferred that trans-national circulation of strains is less present in RTE meat and cheese due to lower number of strains available. Food exchange between EU Member States as well as consumption habits through EU would help to determine if exposition for these two categories is country specific or not.



Blue colour is associated to food isolates.

Figure 8.6.: Timeline and countries for CC121 cluster 7 implicating the largest number of countries (9) of all clusters of strains

The established links for sporadic strains as well as food clusters revealed that some clonal isolates circulate for years in RTE products and confirmed the results of retrospective outbreak investigation (see Section 5).

8.2.3. Consistency of clusters established

Consistency of genetically established clusters can be assessed with epidemiological information associated to each strain. The countries where the strains were isolated and the type of food are the two main elements for consistency assessment. Time of isolation is another criterion, but the relatively short period (2 years) is not very informative.

Dealing with food categories, among the 76 clusters established between food strains and the 27 clusters where human and food strains were linked, only 13 links between more than one food category were established (see Table 8.1, e.g. CC8 cluster_4). Different hypotheses can be advanced to explain the contemporary presence of the same clone in different categories of RTE food. The first is linked to cross-contamination at retail level. It has been recently shown that cross-contamination at retail is of major importance for *L. monocytogenes* (Pouillot et al., 2015; Gallagher et al., 2016). Yet cross-contamination probably concerns products of the same food category, e.g. cheeses (Heiman et al., 2015). Another reason could be the use of common ingredients or equipment in the food chain of the different food categories. Yet the data available for the strains of the present study is not precise enough and there is no scientific literature that may help supporting or checking this hypothesis. Finally, the presence of strains from different categories could be explained by false positively associated strains. Indeed the threshold used to distinguish strain was set to 25 SNPs based on retrospective analysis of outbreaks (Section 5). SNP pairwise distributions for some CCs show that this threshold probably induces a loss of specificity. The first two of the multimodal distributions of pairwise SNPs distribution (obtained of a set of diverse strains) help to distinguish closely related

strains. In our analysis, for some CCs, it appeared that the threshold of 25 is the second mode (data not shown).

8.3. Conclusion

For any outbreak investigation, making a linkage between clinical isolates and possible food sources requires distinguishing the suspected pathogen from the circulating background population whatever methodologies are employed. The retrospective analysis conducted here shows that numerous consistent genetic linkages, between a priori unlinked strains, can be established with WGS. Data to support the establishment of the actual epidemiological linkages between the genetically related strains was not available in this retrospective study.

With less discriminatory method (PFGE, MLST), outbreak detection is mainly based on cluster of time-linked strains that shared the same profile (Yde et al., 2012). Systematic comparison (e.g. without considering time-linked) with these microbiological methods is not possible as it would result in too numerous potential epidemiological links and investigations in the field. The discriminatory power of WGS completely changes the paradigm of outbreak investigation. Direct comparison (based on SNP or any discriminatory method like it, e.g. cgMLST) of genomes, even in low number and/or timely separated by several months, would result in specific and sensitive potential links.

We used a maximum SNP distance (that was confirmed by phylogenies) for establishing the link between strains. Although we used a single rule whatever the CCs for retrospective investigation, setting a single diversity threshold might not be the most specific and sensitive approach. According to the diversity of subtypes in each CCs and the timeframe, this level could probably be adapted. Furthermore, the threshold used is only valid for the workflow used to generate SNP pairwise distances. Other variant calling workflow would result in different distributions (Sahl et al., 2016), and thus different thresholds.

9. Putative markers

L. monocytogenes is widely found in the environment. Its ability to persist in a diverse range of niches is supported by its ability to respond to the different stresses it encounters (Gandhi and Chikindas, 2007). These stress responses confer on it the ability to persist in environment (Gandhi and Chikindas, 2007), as well as ensuring successful transition from food into the gastrointestinal tract of hosts (Toledo-Arana et al., 2009). Genomics data from a large collection of isolates provides the means to identify marker genes associated with pathogen stress survival and/or virulence (Franz et al., 2014).

While the majority of *L. monocytogenes* isolates are generally susceptible to a large number of anti-microbials, a small portion (Wieczorek et al., 2012; Barbosa et al., 2013; Khen et al., 2015) demonstrate resistance to certain clinically used anti-microbials recommended for treatment of listeriosis infection in pregnancy (Donovan, 2015). Even such low level of resistance is of concern as it may represent an emerging pattern of developing resistance (Khen et al., 2015). Along with virulence factors, antibiotic resistance genes that have been previously described (Charpentier and Courvalin, 1999; Lungu et al., 2011) were sought in the LISEQ collection.

Over the last 15 years, numerous virulence factors have been identified (Vázquez-Boland et al., 2001; Toledo-Arana et al., 2009; Maury et al., 2016). We aimed at comparing the presence/absence of these virulence factors in the genomes of the LISEQ clinical and food isolates. We sought to find any population level differences at the lineage level which may suggest adaptation or association of particular factors to survival in the environment or to the clinical manifestation of listeriosis.

L. monocytogenes can remain on equipment or surfaces (Mettler and Carpentier, 1999) for several months or years (Jessen and Lammert, 2003; Carpentier and Cerf, 2011). WGS has recently been shown to be an invaluable tool to detect persistent strains in processing plants (Fagerlund et al., 2016; Morganti et al., 2016). What is of concern is that the presence of persistent cells on food-

contact surfaces can be a source of recontamination (Lundén et al., 2002; Reij et al., 2004). One hypothesis to explain persistence is the ability of bacteria to adapt to and survive environmental stresses such as nutrient deprivation, hot or cold temperatures, sanitisers and preservatives, desiccation, low pH, and high salt concentrations (Thévenot et al., 2006; Carpentier and Cerf, 2011; Melo et al., 2015). Some studies suggested that persistent bacteria are genetically distinct from transient strains (Autio et al., 2003; Wulff et al., 2006; Holch et al., 2013). For persistence, the first objective was to test the ability of WGS to detect potential persistent strains among all strains collected in a cheese plant from country Q. The second objective was to compare the presence/absence of specific genes involved in persistence in strains isolated in food processing environment (potentially persistent strains) to those in strains isolated in raw product (potentially non persistent, or transient strains).

Finally, molecular genotyping techniques may also assist in the identification of potential host-associated genetic markers. This association has already been tested for several foodborne pathogens (Sheppard et al., 2013; Hayward et al., 2016), we aim here to carry out this search for markers for *L. monocytogenes*.

9.1. Methods

9.1.1. Antibiotic resistance genes

Resistance to tetracycline, penicillin, benzalkonium chloride, quaternary ammonium sanitizers and antiseptic were assayed in the genomes of the isolates in this study. Tetracycline resistance was inferred from the presence of *tetM* and *tetS*, penicillin resistance inferred by the presence of *penA*, benzalkonium chloride by the detection of the *bcrABC* locus and the *Tn6188* insertion. Resistance to quaternary ammonium sanitizers and antiseptic was inferred by the presence of the efflux pump *emrE* (Charpentier and Courvalin, 1999) and *qacA* (Lungu et al., 2011). For detection of genes presence, “paired-end” reads of each strain were mapped against the reference gene sequences using Bowtie2 v.2.2.5. (Langmead and Salzberg, 2012). The resulting alignment .sam file were then converted into .bam files and sorted by using SAMtools (Li et al., 2009). Genes were defined as detected if they covered greater than 80% of the query sequence with greater than 80% nucleotide identity. Genes with coverage less than 100% were also classified as truncated.

9.1.2. Published virulence factors

A comprehensive set of 115 genes identified as putative or confirmed virulence factors were used from the two studies (Camejo et al., 2011; Maury et al., 2016). The gene sequences were extracted from *L. monocytogenes* EGD-e (accession NC_003210.1) apart from the LIPI3 clusters of gene is extracted from *L. monocytogenes* F2365. Genes were detected as described in 9.1.1.

9.1.3. Genes implicated in persistence

Genetic loci involved in persistence were selected through a bibliographic research of significant genes related to three main bacterial functions: cold growth, biofilm and resistance (Felix et al., 2015). The list is given in Table 9.1. Genes were detected as described in 9.1.1.

Table 9.1.: Loci targeted through bibliographic research for persistent marker study

Main function	Genes	Gene Id	Gene functions	References
Biofilm	actA	Lmo0204	Aggregation factor	(Travier et al., 2013)
Biofilm	-	lmo0673	Flagellar operon	(Renier et al., 2011)
Biofilm	bapL	lmo0435	Peptidoglycane associated protein	(Renier et al., 2011)
Biofilm	recO	lmo1460	DNA gap repair protein	(Tremoulet et al., 2002)
Biofilm	-	lmo2504	cell wall-binding protein	(Lourenço et al., 2013)

Main function	Genes	Gene Id	Gene functions	References
Biofilm	<i>luxS</i>	lmo1288	Quorum sensing A12 biosynthesis protein	(Bonsaglia et al., 2014)
Cold adaptation	<i>cspB</i>	lmo2016	RNA chaperon protein	(Schmid et al., 2009)
Cold adaptation	<i>cspD</i>	lmo1879	RNA chaperon protein	(Schmid et al., 2009)
Dessication resistant	<i>fljP</i>	Lmo0676	flagellum biosynthesis	(Hingston et al., 2015)
Dessication resistant	<i>flhB</i>	lmo0679	flagellum biosynthesis	(Hingston et al., 2015)
Dessication resistant	<i>flgD</i>	Lmo0696	flagellum biosynthesis	(Hingston et al., 2015)
Dessication resistant	<i>flgL</i>	lmo0706	flagellum biosynthesis	(Hingston et al., 2015)
Dessication resistant	<i>motB</i>	Lmo0686	Motor control	(Hingston et al., 2015)
Dessication resistant	<i>fliM</i>	lmo0699	Motor control	(Hingston et al., 2015)
Dessication resistant	<i>fliY</i>	NC_019556.1	Motor control	(Hingston et al., 2015)

No gene symbol: -

9.1.4. Markers of host association

Analyses of genetic markers between isolates from each source to isolates from humans were carried out to identify genetic markers, which differentiated significantly between these hosts. Four genotyping methods were used to identify genetic characters: 7-locus MLST, 30 locus rMLST, 1,748 locus cgMLST and 39,529 locus cgSNP. To reduce the size of the cgSNP dataset loci with only 1 SNP difference were removed from analysis, which left only 19,902 loci.

For each of the genotyping datasets, for each allele at each locus, odds ratios were determined for the difference in allele abundance in one host compared to human isolates. Statistical significance ($P < 0.05$) of these odds ratios was determined by Fisher's exact test with Bonferroni correction incorporated for multiple comparisons.

9.2. Results

9.2.1. Antimicrobial resistance

Table 9.2 shows the percentage of strains in the study harbouring the assayed resistance genes. The resistance profile for each strain is included in the supplementary file (Annex A). Less than 1% of isolates showed likely resistance to tetracycline via *tetM* with no detection of *tetS*. Benzalkonium chloride resistance was conferred in 18.5% of isolates by Tn6188 insertion and approximately 5% of isolates by the *bcrABC* loci. Less than 1% of isolates harboured the efflux proteins *emrE* and *qacA* whilst the efflux protein *qacC* was found in 18.3% of isolates and generally found in conjunction with Tn6188. No isolates showed likely resistance to penicillin through the presence of *penA*.

Table 9.2.: Percent of isolates in the study harbouring the assayed resistance genes

Gene	% Detection
<i>tetM</i>	0.6
<i>tetS</i>	0
<i>bcrA</i>	4.9
<i>bcrB</i>	4.9
<i>bcrC</i>	4.7
<i>emrE</i>	0.3
<i>qacA</i>	0.5

Gene	% Detection
<i>qacC</i>	18.3
<i>Tn6188qac</i>	18.5
<i>penA</i>	0

9.2.2. Published virulence factors

The supplementary file (Annex A) shows the presence and absence of 115 putative virulence markers across the strain collection. Of the 115 markers 2 were absent across all isolates, conversely 92 markers were present in greater than 95% of isolates. Figure 9.1 shows for each virulence marker the proportion that was present in lineage I and lineage II isolates.

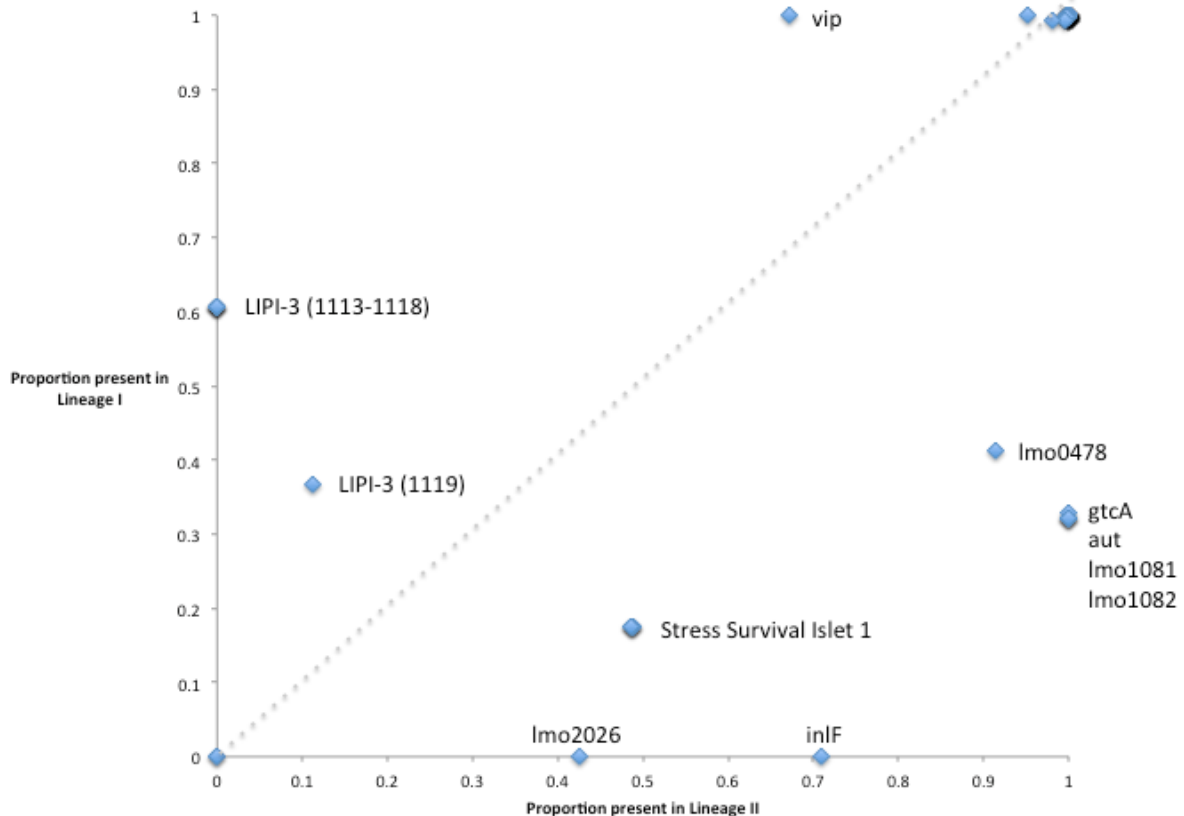


Figure 9.1.: Scatter plot showing the proportion each of the 115 putative virulence markers found in lineage I or lineage II

In total, 21 putative virulence markers had significant variability in their detection across the strain collection. As described by Maury *et al* (Maury *et al.*, 2016) the *Listeria* pathogenicity island 3 (LIPI-3) was found in 60% of isolates from lineage I (ubiquitous in CC1, CC3, CC4 and CC6) but completely absent in lineage II isolates. LIPI-3 loci 1119 showed a different presence and absence profile to the other LIPI-3 alleles with it being found in a minority of lineage II isolates (12/187 CC121, 11/54 CC155, 14/98 CC8 and 11/110 CC9) in absence of the other LIPI-3 loci. Conversely, in lineage I isolates some isolates do not possess loci 1119 and have an otherwise intact LIPI-3.

The known virulence surface protein Vip (Cabanes *et al.*, 2005) was found across all isolates in lineage I but only 70% of lineage II isolates (absent in CC204, CC21, CC31, CC37 and only present in 1/43 isolates in CC7 and 3/98 isolates in CC8). Several putative virulence factors were found in a greater proportion in lineage II isolates compared to lineage I isolates. These included the internalins *Imo2026*

(absent in lineage I and ubiquitous in CC155, CC18, CC20, CC204, CC21, CC37, CC415, CC7 and CC9 in lineage II) and *inlF* (absent in lineage I and only absent in CC121 and CC14 of lineage II) previously shown to be detected variably in different serotypes (Chen et al., 2009).

The five gene locus termed the stress survival islet (SSI-1) (Ryan et al., 2010) which has previously been associated with growth of *L. monocytogenes* under sub-optimal conditions, contributing to survival of certain strains in food environments, was over-represented in lineage II isolates. However, when we consider the number of clonal complexes this association is less clear. SSI-1 is present in CC3 and CC5 of lineage I and conversely absent in lineage II CCs 101, 121, 14, 20, 21, 415 and 7.

Ubiquitous amongst lineage II isolates was the recently described *rmlADBC* L-rhamanose biosynthesis loci (Imo1081 and Imo182) (Carvalho et al., 2015) involved in producing wall teichoic acids providing protection against the activity of antimicrobial peptides as was *gtcA* (Promadej et al., 1999) also involved in ecoration of cell wall teichoic acid of *L. monocytogenes*. The autolysin *aut* was also found across all lineage II isolates but only in CC3, CC5, CC59 and CC87 of lineage I, which is perhaps surprising given its proposed role in entry of *L. monocytogenes* to non-phagocytic mammalian cells (Cabanés et al., 2004). Finally, the surface adhesion *lapB* required for entry into mammalian cells is present across all lineages but absent in all isolates of CC31.

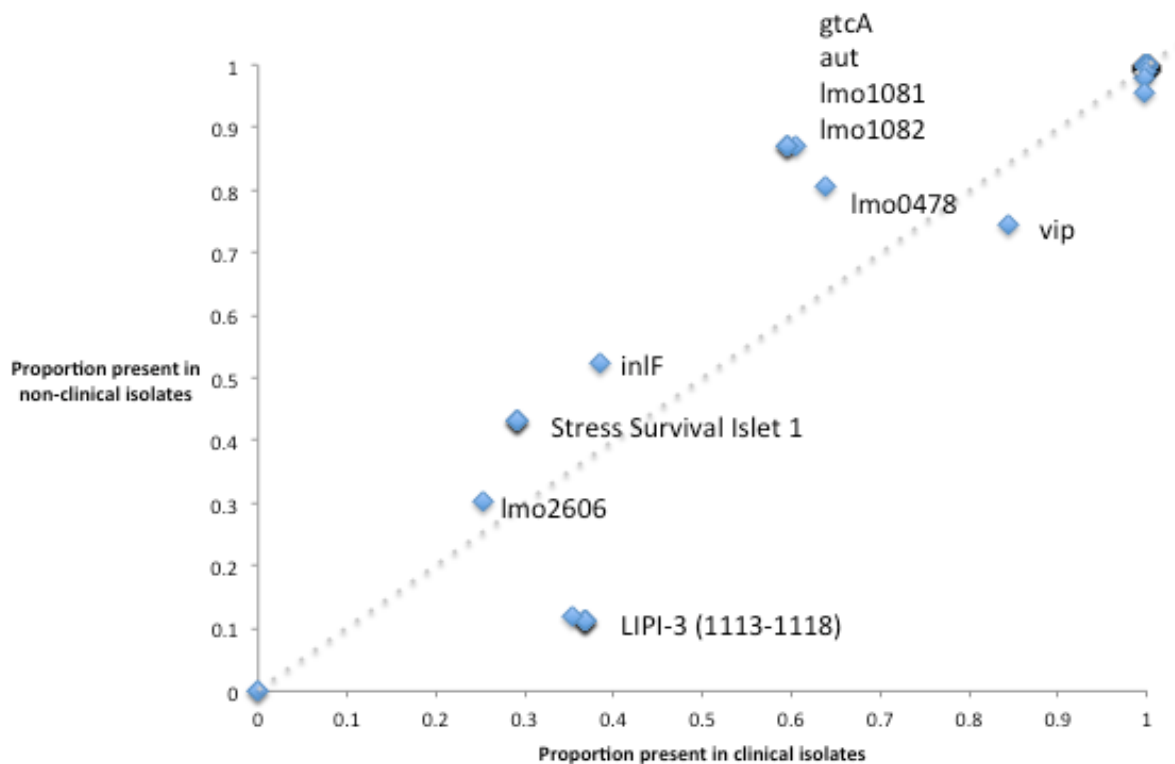


Figure 9.2.: Scatter plot showing the proportion each of the 115 putative virulence markers found in clinical or non-clinical isolates

Figure 9.2 shows for each virulence marker the proportion that was present in clinical and non-clinical isolates. When compared to the assortment by lineage there seems less effect than by whether the isolate was from a clinical sample or not.

Loss of function through partial gene deletion or miss-sense mutations is also known to be important in virulence attenuation. To explore this, genes with less than 100% coverage of the query sequence were designated as truncated (see supplementary file – Annex A). Several genes had loss of function truncation in lineage II but were found intact in lineage I, these include the already described *inlA*

deletion (Maury et al., 2016) as well as lmo0257, the terminal SSI loci lmo0478, the autolysin ami and the actin-assembly inducing protein precursor actA. Conversely several genes were truncated in lineage I but intact in lineage II isolates. These included the internalins inlH, inlJ, lmo1290, the stress protein clpB and the flagellar motor switch protein lmo0698.

9.2.3. Genes implicated in persistence

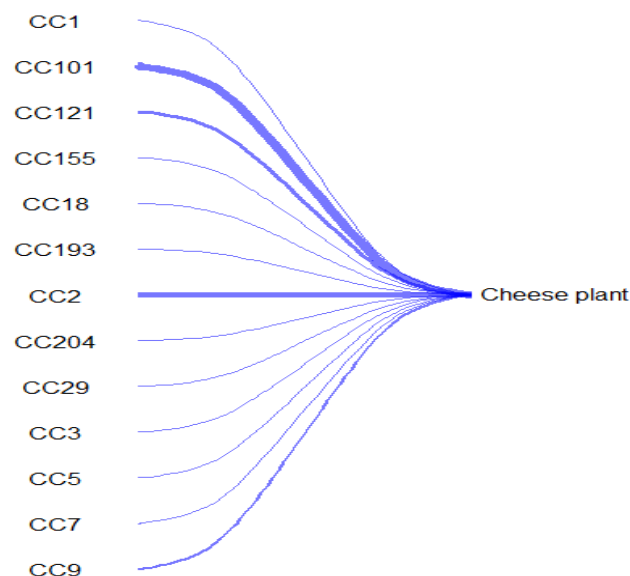
For persistence, the food processing facilities of three food sectors were investigated. For two cheese production environments, WGS was used to decipher the strain diversity and the origin of contamination. For the pork strains, the dataset was constituted of previously identified persistent and non-persistent strains. The presence of markers within these two subsets was investigated. For the salmon producers of country B, WGS was used to identify the origin of contamination and to identify persistent strains within other strains. The research of presence of putative markers was carried out on these strains.

Persistence in dairy plants

Two different dairy plants were investigated. The first one corresponds to a cheese plant in country B, the second one to a cheese plant in country Q.

In total there were 10 isolates that were genome sequenced from contamination of full fat semi soft unpasteurised cheese made from bovine milk (check this) originating from country B. Six of the isolates were from two cheese products (First product: RL15000630, RL15000631, RL15000, second product: RL15000637, RL15000638 & RL15000639) and 4 from the factory environment (RL15000635 (rack in the chill that cheese were stored on), RL15000634 (swab of brine trolley handle), RL15000633 (swab of top of brine trolley), RL15000632 (swab from brush used in factory)). All of the isolates were ST37 except RL15000639 which was ST121. This isolate was the first isolated (January 2013). The remaining isolates were obtained in the following 2 months. As all ST37 strains present less than 25 SNPs, it can be concluded that a single clone persisted in the factory environment and was at origin of the cheese contamination.

Within the 100 strains isolated in the cheese plant from country Q in 2012, 13 different CCs were found. The most prevalent CCs were CC101 and CC2. Figure 9.3 shows the repartition of the different CCs.



Line thickness is proportional to number of isolates (largest CC represents 32 isolates, smallest CCs are represented by 1 strain).

Figure 9.3.: Repartition of the thirteen CCs isolated in a cheese plant from country Q

Phylogeny established with SNP analysis data of 16 CC2 strains presented on Figure 9.4 helps to decipher which strains are related with other. At least 3 different clusters of strains of CC2 circulate in the processing plant. Yet none of these clusters of strains in the environment match with the strain isolated from cheese. Such an analysis would not have been possible with less discriminatory methods.

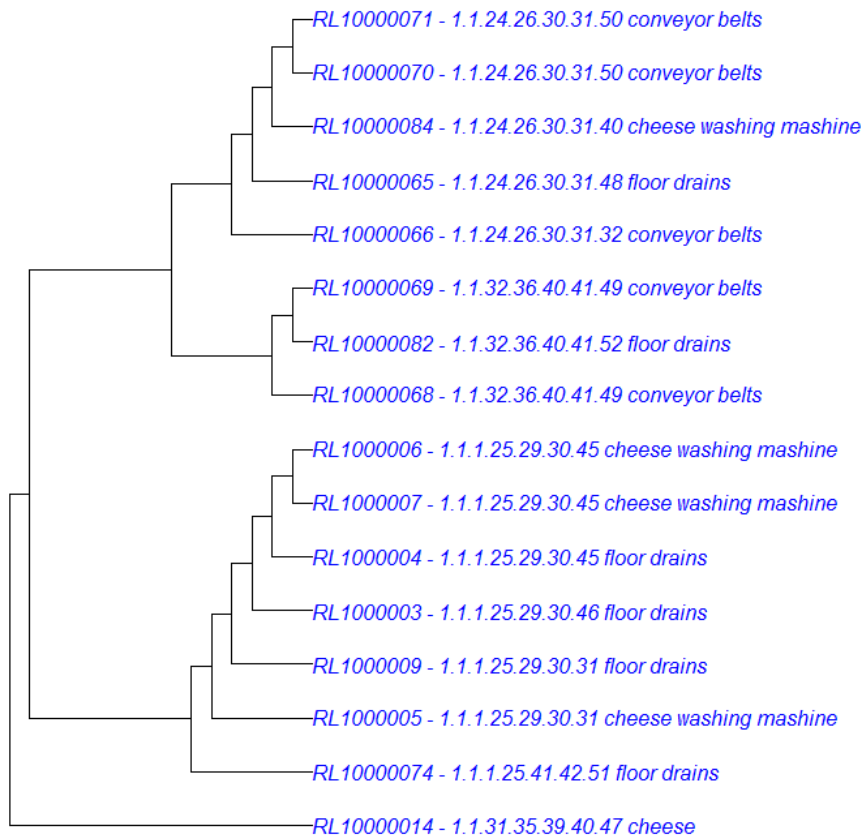


Figure 9.4.: ML phylogeny tree of 16 strains of CC2 isolated in the same cheese factory

Persistence in the pork processing environment

The pork processing strains from cutting plants were specifically selected to look for potential differential presence/absence of putative markers for persistence. The presence/absence analysis for the 15 gene loci identified to be of importance for persistence in food processing environment is given in Table 9.3.

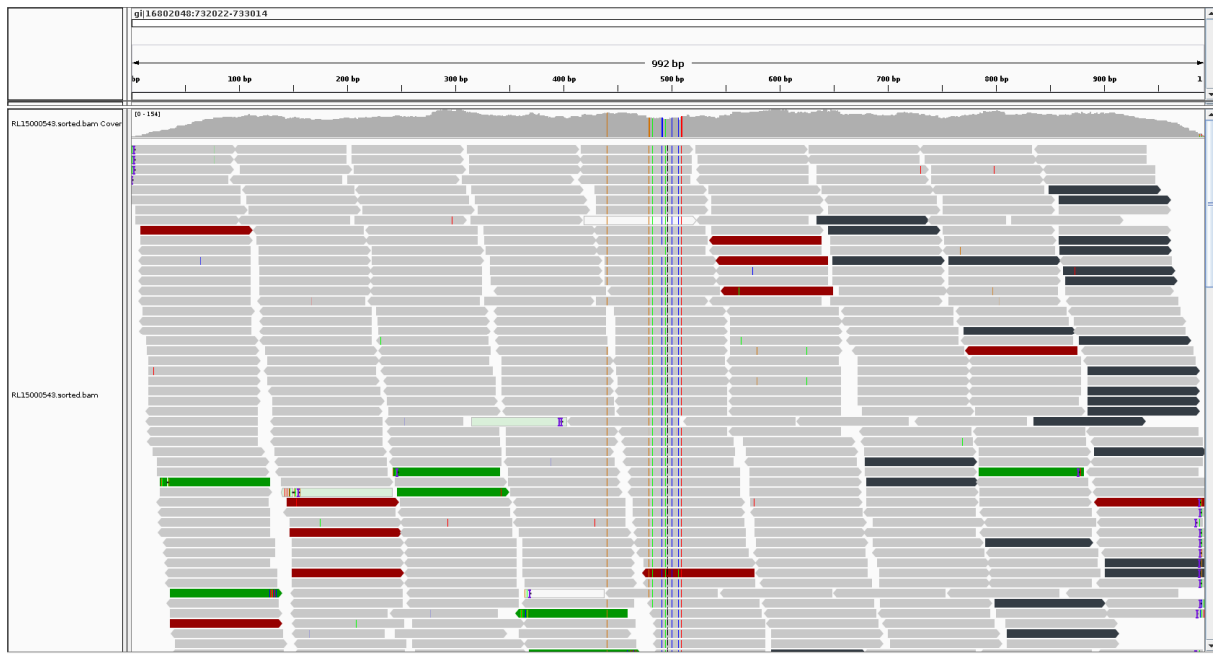
Table 9.3.: Presence/absence of putative markers for persistence in two groups of strains: persistent strains isolated in cutting plants, non-persistent strains isolated in raw material of the same cutting plants (pork processing, Country C)

Gro up	RL num ber	Presence of potential markers for persistence														NC_019 556.1 FluY	
		lmo0 204	lmo0 673	lmo0 435	lmo1 460	lmo2 504	lmo1 288	lmo2 016	lmo1 879	lmo0 676	lmo0 679	lmo0 696	lmo0 706	lmo0 686	lmo0 699		
Persistent	RL1500 0543	X	X		X	X	X	X	X	X	X	X	X	X	X	X	
	RL1500 0542	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	
	RL1500 0541	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	
	RL1500 0540	X	X		X	X	X	X	X	X	X	X	X	X	X	X	
	RL1500 0539	X	X		X	X	X	X	X	X	X	X	X	X	X	X	
	RL1500 0538	X	X		X	X	X	X	X	X	X	X	X	X	X	X	
	RL1500 0393	X	X		X	X	X	X	X	X	X	X	X	X	X	X	
	RL1500 0392	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	
	RL1500 0391	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	
	RL1500 0390	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	
	RL1500 0389	X	X		X	X	X	X	X	X	X	X	X	X	X	X	
	RL1500 0388	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	
	RL1500 0387	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	
	Non persistent	RL1500 0361	X	X		X	X	X	X	X	X	X	X	X	X	X	X
		RL1500 0362	X	X		X	X	X	X	X	X	X	X	X	X	X	X
RL1500 0363		X	X		X	X	X	X	X	X	X	X	X	X	X	X	
RL1500 0364		X	X		X	X	X	X	X	X	X	X	X	X	X	X	
RL1500 0365		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	
RL1500 0366		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	
RL1500 0367		X	X		X	X	X	X	X	X	X	X	X	X	X	X	
RL1500 0368		X	X		X	X	X	X	X	X	X	X	X	X	X	X	
RL1500 0370		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	
RL1500 0371		X	X		X	X	X	X	X	X	X	X	X	X	X	X	
RL1500 0372		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	
RL1500 0373		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	
RL1500 0374		X	X		X	X	X	X	X	X	X	X	X	X	X	X	
RL1500 0375		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	
RL1500 0376		X	X		X	X	X	X	X	X	X	X	X	X	X	X	
RL1500 0377	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		
RL1500 0378	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		

Note: Empty cells indicate absence of markers.

Whatever the group of strain, i.e. the group of persistent strains isolated from food processing environments or the group of strains isolated from recently imported raw materials into the plant, at least 14 out of 15 gene loci were present. The gene locus lmo0435, was not present in all isolates in either of the two groups. Yet the proportion of strains without this locus was the same both groups of strains.

Within the 14 gene loci present, no large deletion or insertion were found in the strains of the two groups (see the example Figure 9.5).




The upper part shows the coverage of reads along the gene. The lower part presents the mapping of each reads. For a deletion, a section of DNA is absent in the subject genome compared to the reference genome. In the case of an insertion, a section of DNA is present in the subject genome that is not represented in the reference genome. Position of SNP (regarding reference gene) in aligned reads can be seen with vertical lines. Insertions are indicated by a purple *I* () and deletions are indicated with a black dash (–). Alignments that are displayed with light gray borders and transparent or white fill, have a mapping quality equal to zero. Green reads present a poor mapping quality. Red reads include one sequencing mutation that corresponds to a sequencing error.

Figure 9.5.: Integrative Genomics Viewer (IGV) screen capture of pair-end reads mapping to *flhM* gene. Link to figure in high quality: <https://github.com/lguillier/LISEQ-codes/tree/master/Chapter9>

No allelic profile was found to explain the persistent phenotype (see two examples for *actA* and *fliY* in Figure 9.6).

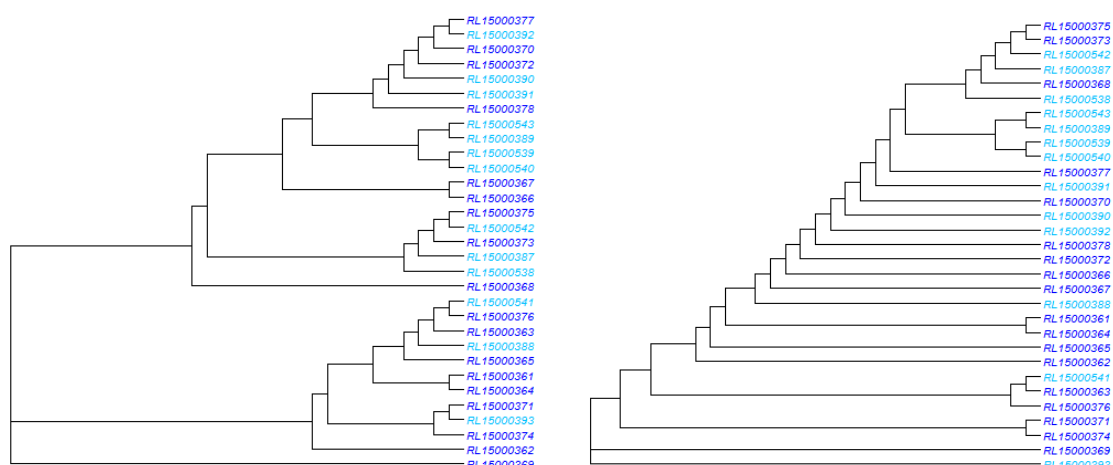
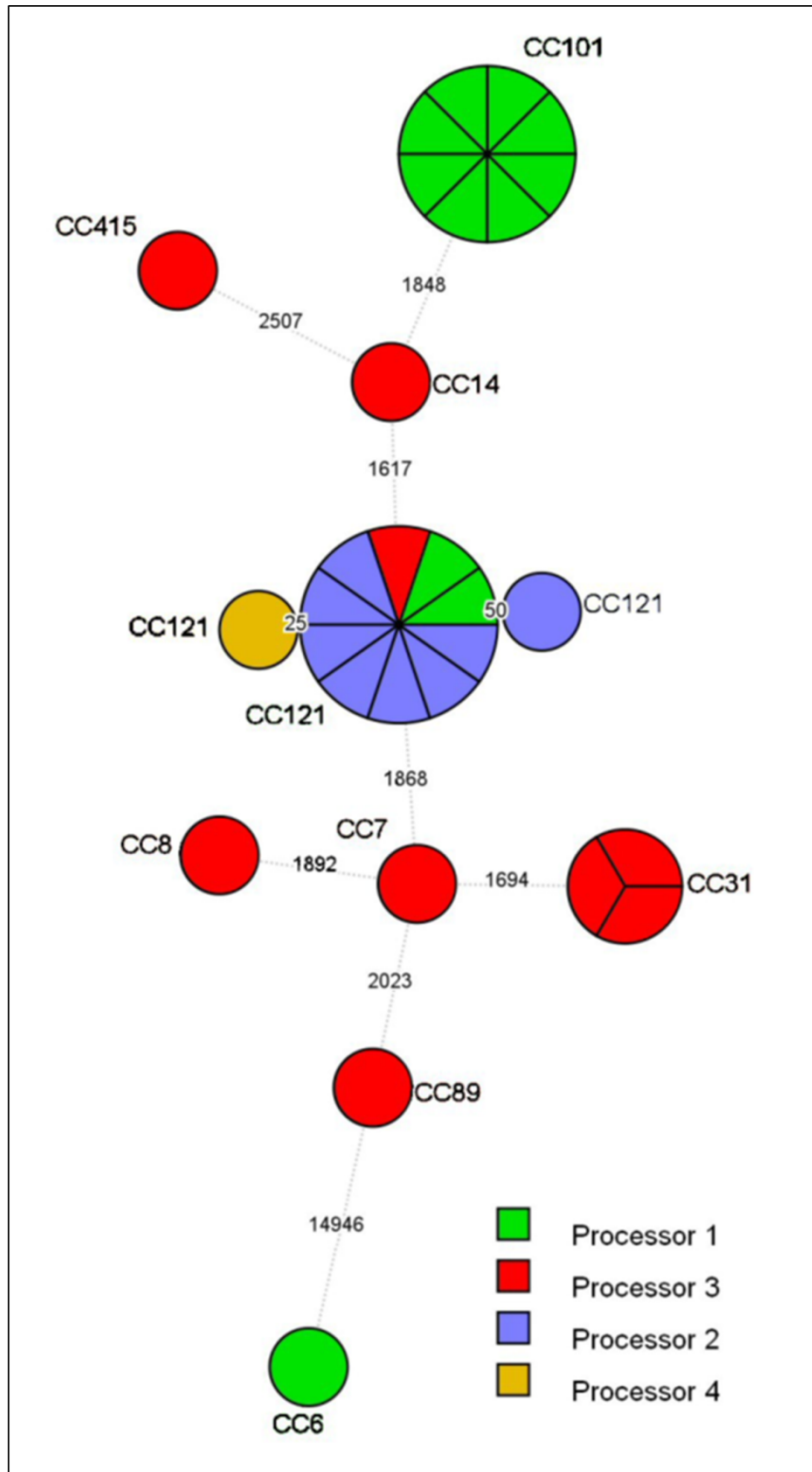


Figure 9.6.: ML phylogeny tree of persistent (light blue) and non-persistent strains (blue) based on SNPs for gene *actA* (left) and *fliY* (right)

Persistence in salmon processing plants

The commonality between 29 isolates collected from four salmon processors in country B, over 4 years, from along the processing chain were determined using WGS.

There were nine different ST from nine different Clonal Complexes, with three multi-isolate ST (CC121, CC101 and CC31) represented by 12, 8 and 3 isolates respectively with the remaining six ST being singletons (Figure 9.7).



Distances on the branches in the figure are measured in SNP differences. Clonal complex 121 is spread throughout all the processors, while processors 1 and 3 also have site-specific clones of Listeria present.

Figure 9.7.: Minimum spanning tree of strains isolated from country B Salmon Processors indicating distribution of isolates by CC and by Processor

Phylogeny (as well as SNP address or SNP pairwise distance) for CC101 strains isolated in Processor 1 shows that the same strains circulated in the plant during the period considered, that is 2011-2013. The same clonal group was present either in food processing environment (e.g. RL15000641 was a strain isolated from a swab) or in raw fish. But none of the strains were isolated in final products.

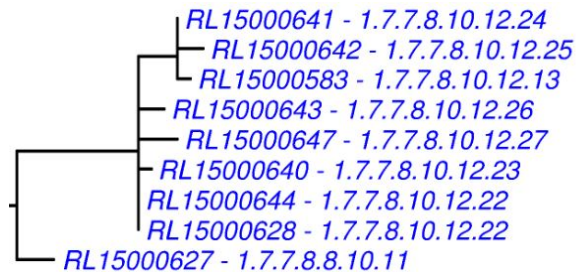
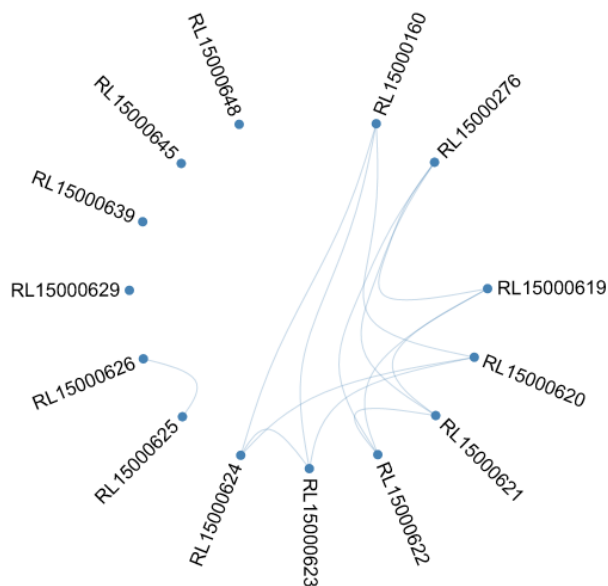


Figure 9.8.: Clade of the ML phylogeny tree of CC101 (extracted from the complete phylogeny presented in Section 4) regrouping all the strains isolated in smoked salmon Producer 1

For CC121, out of the 12 isolated strains in Producer 2, three different clonal groups were identified (Figure 9.9.). It is worth noting that two of them match with two strains from the baseline survey (RL15000160 and RL15000276, cf. Table. 8.1 CC121 clusters 10 and 18).



Upper right: two strains isolated in smoked salmon product sampled for baseline survey.

Figure 9.9.: Links between strains of CC121 established by pairwise SNP distance (below or equal 25)

The presence or absence of putative markers for persistence was tested in the strains of this CC. The strains isolated more than once from Producer 2 were considered as potentially persistent and the four strains that were isolated once were considered as non-persistent (Table 9.4). No large insertion/deletion or SNP help to distinguish both groups of strains (Figure 9.10).

Table 9.4.: Presence/absence of putative markers for persistence in two groups of strains: persistent strains isolated in salmon processing environment and/or finished product more than once, non-persistent strains isolated in food processing environment only once

Group	RL number	Presence of potential markers for persistence														
		Imo0204	Imo0673	Imo0435	Imo1460	Imo2504	Imo1288	Imo2016	Imo1879	Imo0676	Imo0679	Imo0696	Imo0706	Imo0686	Imo0699	NC_0191556.1 FljY
Persistent	RL15000620	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
	RL15000623	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
	RL15000624	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
	RL15000619	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
	RL15000621	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
	RL15000622	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
	RL15000625	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Non persistent	RL15000626	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
	RL15000629	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
	RL15000639	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
	RL15000645	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
	RL15000648	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X



The upper mapping corresponds to an a priori non-persistent strain (RL15000629), the lower to a persistent strains (RL15000621). For a deletion, a section of DNA is absent in the subject genome compared to the reference genome. In the case of an insertion, a section of DNA is present in the subject genome that is not represented in the reference genome. Position of SNP in aligned reads can be seen with vertical lines (regarding the reference gene sequence shown at the bottom of the graph). Insertions are indicated by a purple *I* (I) and deletions are indicated with a black dash (-). Alignments that are displayed with light gray borders and transparent or white fill, have a mapping quality equal to zero. Green reads present a poor mapping quality. Red reads include one sequencing mutation that corresponds to a sequencing error.

Figure 9.10.: Integrative Genomics Viewer (IGV) screen capture of pair-end reads mapping to fljM gene. Link to figure in high quality: <https://github.com/lguillier/LISEQ-codes/tree/master/Chapter9>

9.2.4. Markers of host association

The number of isolates in the different host sets and the number of loci screened for each genotyping dataset is detailed in Table 9.5. For each method and host pair Table 9.5 indicates the total number of alleles identified which were significantly different for that host pair and secondly the number of loci which harboured these alleles.

Overall the number of loci which could differentiate between human isolates and a source isolate were rare for human-bovine and uncommon for human-poultry, suggesting that the strains found in bovine and poultry sources are genetically similar to those in human cases. Other sources had greater numbers of distinguishing loci between the host and human isolates.

7-locus MLST and rMLST both use loci that are considered selectively neutral whilst cgMLST and cgSNP comprise markers which span the spectrum from neutral through to those loci under selection (both negative and positive selection). Genetic variation in 7-locus MLST, rMLST and cgMLST loci is classified using alleles which differ from each other by sequence polymorphisms which can be anything from a single nucleotide through to several nucleotide differences; thus these alleles are not truly independent of each other as they may harbour common polymorphisms. cgSNP, on the other hand, are defined as a unique site in the genome (i.e. a particular nucleotide position) and so will be independent of each other. Secondly cgSNP, as with other genotyping schemes, have allelic variants at each of the cgSNP loci in the form of four alternative bases.

An ideal genetic marker for molecular host attribution would be one which was found exclusively in one host source and not in other sources. Since attribution works with several hosts all such ideal host specific markers would be pooled together and used collectively. The benefit of this strategy is that those markers which do not significantly contribute to host specificity - they are 'neutral' and can only reduce the strength of the host specific signal – are excluded, and so more robust attribution scores should result.

Table 9.5.: Abundance of putative markers to differentiate a source host isolate from a human isolate

Genotyping method	Number of Loci	Human isolates (N=254) vs					Loci found in any host
		Bovine (N=61)	Fish (N=323)	Ovine (N=89)	Poultry (N=25)	Swine (N=112)	
MLST	7	0 / 0	15 / 7	10 / 7	0 / 0	8 / 7	7
rMLST	30	0	45 / 21	10 / 9	1 / 1	32 / 19	22
cgMLST	1,748	4 / 4	3,900 / 1745	1,684 / 1,399	100 / 100	2,506 / 1,567	1,748
cgSNP	19,902	0	16,801 / 8,393	2,203 / 1,129	112 / 58	14,379 / 7,227	9,164

For each method and host pair, the cell indicates the total number of alleles identified which were significantly different for that host pair and secondly the number of loci which harboured these alleles.

9.3. Conclusion

Antimicrobial resistance in *Listeria* sp. has been studied in various food, environmental and clinical settings (Bertrand et al., 2005; Morvan et al., 2010; Granier et al., 2011; Jamali et al., 2015). *Listeria monocytogenes* has generally been shown to be more susceptible to antimicrobial agents than other species in the genus. In this study we found remarkable low resistance to tetracycline (<0.1%) and penicillin (1%). Resistance to detergents and antiseptics via efflux activity was significant with mechanisms detected at a prevalence approaching 20%. Whilst it is encouraging that the isolates in this study show low levels of antimicrobial resistance it is important to remain vigilant for emerging resistance. Whole genome sequencing allows antimicrobial resistance monitoring to be done as a cost-

neutral activity if WGS is part of routine microbial surveillance and therefore allows this potential threat to be reviewed going forward.

Whole genome sequencing provides the opportunity for rapid interrogation for markers of virulence. In this study 115 putative markers of virulence were assayed for their presence or absence in this data set. Less than 20% of markers were present in less than 95% of the isolates suggesting that most putative markers described in the literature are fairly ubiquitous across at least lineage I and lineage II *Listeria monocytogenes*. Of those that vary the majority were over-represented in food and/or lineage II isolates with markers associated with stress survival or cell wall modification particular enriched. Conversely the recently discovered *Listeria* pathogenicity island 3 and the surface protein vip were enriched in clinical and/or lineage I isolates. Although most virulence markers were present in all strains we do not know if the genes are in-fact expressed. Several truncations were identified in virulence genes across the dataset with some having an increased propensity for truncation dependent on lineage.

The present study confirms recent studies that showed that WGS and SNP-based analysis is well-suited to investigated persistence and contamination routes of *L. monocytogenes* in food processing facilities and in the food chain (Fagerlund et al., 2016).

The presence/absence of genes thought to promote persistence was not found to be pertinent for predicting persistent phenotype. SNPs as well as insertion and deletion in these genes were not helpful either. The study of expression of gene marker for persistence (Mazza et al., 2015) or proteome analysis (Rychli et al., 2016) have recently appeared to be more promising for predicting persistence phenotypes. The analysis of the accessory genome is also an important element in persistence study as it has been recently shown that conservation of the accessory might be associated with persistence (Fagerlund et al., 2016).

This study did not consider the accessory genome, which by definition comprises genes, which are not present ubiquitously across the population. Such genes will make a significant contribution to the variation in biology seen between strains and therefore should be a rich source for the discovery of polymorphisms associated with host association, and indeed many other features. This pilot study suggests that cgSNP (see 9.2.4), and by extension SNP in the accessory genome, are likely to be the most fruitful source of host associated polymorphisms, which may be of use in refining molecular attribution models.

10. Conclusions

The overall objective of this study was to compare *L. monocytogenes* isolates from the EU-wide BLS on ready-to-eat foods conducted in 2010-11, with isolates from compartments along the food chain and from human cases using WGS analysis. This was achieved by meeting the three described specific objectives.

The first specific objective was met by assembling a fully representative isolate collection that consisted of a total of 1,143 *L. monocytogenes* isolates from across the EU, including 810 isolates from along food chain and 333 human clinical isolates. The food chain isolates comprised 353 from the EU-wide baseline survey (BLS) on the prevalence of *L. monocytogenes* in certain RTE foods, 423 from national surveys, control programmes or research projects and 34 food isolates from outbreak investigations. The clinical isolates were provided voluntarily by European national public health laboratories and comprised 262 isolates from sporadic cases and 71 from outbreaks. Isolates were selected within a time frame of 2010-2012 as far as possible although this was extended as necessary to ensure the strain collection was as representative as possible within the scope of the study. The majority of isolates were whole genome sequenced at Public Health England's sequencing facilities using state of the art equipment and methodologies under an accredited quality management system. For a minority of isolates WGS data was already available with sequencing having also been undertaken at PHE and were included in the analysis subject to WGS data meeting the same quality

metrics used for WGS data generated as part of this study. A database was constructed with the available metadata for the isolates with links to their respective genome sequences.

In order to fulfil the second specific objective it was necessary to investigate the phylogeny of the *L. monocytogenes* isolates and produce data sets, in order to provide a framework for further analyses on the genetic diversity and potential epidemiological associations. This was carried out using a range of bioinformatic procedures including several gene-by-gene based approaches such as 7-gene MLST, rMLST, cgMLST as well as SNP-based methods including cg SNP analysis. This study has facilitated the WGS analysis of a unique and large data set of *L. monocytogenes* isolates and has enabled the population to be defined to an unprecedented level of resolution from lineage to nucleotide. The phylogeny showed a clear delineation between *L. monocytogenes* lineages and between clonal complexes within lineages. All isolates in the study were in either lineage I or II. There was a huge amount of diversity among the genomes sequenced and they cover the diversity in lineages I and II as described previously by Ragon et al. (2008). There was an uneven distribution of isolates both between the two lineages and also amongst the clonal complexes within each lineage but this is to be expected due to the number of isolates selected from each source and due to the particular sources themselves which were restricted to RTE foods, compartments of the food chain and human clinical cases. A key finding from the phylogenetic analysis of four large CCs (CC8, CC9, CC101, CC121), is that within a CC, clinical isolates are not associated to a specific clade of the tree. The phylogenetic analysis also confirmed recent work by Maury et al. (2016) that CC4 is associated with highly virulent clinical strains.

The third specific objective of this project was to investigate the suitability of WGS as a tool in the investigation of listeriosis outbreaks. This was performed by a retrospective analysis of human and food isolates that had been previously epidemiologically and microbiologically linked. The sequences from each previously defined outbreak were analysed together with all other isolates from this study in the same clonal complex. The CCs to which the outbreak isolates belonged were analysed by two bioinformatic methods, SNP-based analysis and cgMLST and an important finding was that overall the two methods gave concordant results. Nine outbreaks were studied in total of which 6 were typical point source outbreaks. In each of these, previously epidemiologically linked isolates clustered closely together within a maximum 8 SNP pairwise cluster, and separated from other isolates of the same CC that were included in the outbreak analysis. The remaining 3 outbreaks showed more variation although linked isolates could be defined within a maximal 12 SNP pairwise cluster. Two of these outbreaks occurred over an extended time period and the variation seen may reflect diversity within the source over a long period. The third outbreak consisted of two separate outbreaks and restricting the SNP cluster threshold to 5 would not have included all the epidemiologically linked cases. Increased diversity within an outbreak may be due to differences in the ecology of outbreaks e.g. the involvement of more complex food distribution networks.

One of the outbreaks (5) demonstrated the potential impact of SNPs/alleles being acquired in a single event (e.g. phage) on outbreak analyses. Whilst in this instance removing the particular region did not influence the overall interpretation of the phylogenetic analysis, it demonstrates that knowledge of where in the genome SNPs are occurring can sometimes be very important and should be taken in to consideration when using SNP and gene by gene approaches (Wang et al, 2015). There were no additional food isolates included as part of outbreak analyses that fell within the cluster of human isolates for any of the outbreaks. However, in 4 outbreaks 1 or 2 human isolates submitted as sporadic isolates did cluster together with the outbreak isolates. In all four cases these isolates originated from the same country as the outbreak. This study therefore demonstrates the potential of WGS analysis to detect more cases as being part of an outbreak than previous typing methods. Whilst there was not an international aspect to the outbreaks that were analysed in this study, it does demonstrate the ease with which WGS can accurately rule isolates in or out of outbreaks and how valuable the method would be for international surveillance. This study included a limited number of previously identified outbreaks and that were predominantly restricted in time and diversity. In order to fully assess the usefulness for WGS analyses for outbreak investigation more diverse outbreaks including those involving multiple strains and those across more than one Member State need to be examined.

This study shows that WGS analysis clearly separates outbreak isolates from background isolates within the same CC and thus WGS is very well suited for detecting and defining outbreaks. The results also illustrate that when applying WGS analysis every outbreak should be considered in its own context and that there should not be a single universal cut off value for separating outbreak and background isolates. Analysis and interpretation of WGS clusters requires expert knowledge and collaborative input from bioinformaticians, epidemiologists and microbiologists.

Specific objective 2 was to analyse the WGS data of the selected *L. monocytogenes* isolates and thereby explore genetic diversity, epidemiological relationships and investigate putative markers of survival and pathogenicity. Exploring the genetic diversity of *L. monocytogenes* within and between different sources including those of human origin was accomplished using Simpson's Diversity index and Rarefaction. The genetic distance between each source was investigated using Nei's genetic distance (Nei, 1975). Simpson's index indicated high genetic diversity (>0.8) within all the sources investigated for both 7 locus MLST and 30 locus rMLST. Rarefaction demonstrated that only a small proportion of the diversity had been sampled. Whilst we have sampled representatively across the population structure we have not sampled deep into the diversity of the species. Isolates from clinical cases were found to be more diverse than isolates from other sources and this may not be unexpected as humans are most likely to be exposed to greater variety of sources. Other contributing factors are likely to be involved, including that the food isolates were restricted to mainly fish and meat and that the number of isolates were limited. Whilst isolates from all other sources were different to those from humans at all levels explored, those from the bovine source were found to be the closest genetically by Nei's genetic distance. However, whether this is a robust finding or an artefact of the sampling for this study needs to be verified using additional isolates to see whether this pattern continues. It is important, when considering the findings here, to note that isolates for each source came from different points in the food chain with those closest to retail being at a greater chance of cross contamination from another source. Thus, further work using isolates more widely distributed across the food chain is required to provide more robust data. However, because there was not a random genetic distribution between the different sources this study demonstrates that source attribution based on WGS has the potential to produce useful results.

The second part of specific objective 2, to assess the epidemiological relationship of *L. monocytogenes* from the different sources and of human origin considering the genomic information and the metadata available for each isolate, was investigated in two ways. Firstly by exploring source attribution and secondly by analysing WGS data in conjunction with isolate metadata, to investigate any potential relationships between circulating strains in the EU from 2010-2012. All of the source attribution models showed bovine reservoir to be the main source of human disease, however, other sources also contributed and, for most models, confidence intervals were high. For all sources, isolates from different parts of the food chain were combined to produce a sufficient dataset to perform source attribution. It is possible that the genetic distribution of isolates associated with a particular source may change along the food chain and that this could affect the source attribution results. This area merits further investigation and increasing sample size would improve the robustness of the results and reduce biases. It was found that increasing the number of loci did not improve source attribution for all of the models. However, the Aberdeen method tended to perform better (i.e. be more reliable) with larger numbers of loci. New approaches need to be developed for source attribution using the information that is available across the genome. This is because a number of the loci/SNPs are not informative about the source and appear to add noise to the attribution results. There is also the potential for future research to link the source attribution results to the risk from consuming a meal and quantitative risk assessment.

Establishing links between clinical isolates and food isolates is essential for controlling listeriosis and preventing outbreaks and the second way epidemiological relationships were investigated was by analysing WGS data in conjunction with isolate metadata. This was conducted using a SNP-based approach as currently this provides the highest level of strain discrimination. In this study numerous consistent genetic linkages between *a priori*, unlinked strains were identified, some of which involved isolates from multiple countries. A total of 151 clusters were detected including 124 novel clusters that had not been detected previously. Of these, 48 included one or more sporadic human isolates of

which 17 contained only human isolates, and were thus not linked to any of the food isolates included as part of this study. The analysis also revealed sporadic cases that were genetically related to some of the known outbreaks investigated in Section 5, demonstrating the potential of WGS analysis to identify previously undetected cases. The additional cases identified in this study were from the same country in which the outbreak originally occurred and may possibly also represent earlier cases caused by a strain that went on to cause an outbreak. For 27 novel clusters there was at least one food isolate; potentially relating human cases to contemporary food isolates circulating in the EU. Approximately half of all novel clusters detected contained food isolates only and the analysis revealed that strains were circulating in several different EU countries. Whilst this was particularly evident for smoked and gravad fish isolates there were far fewer isolates from meat and soft cheese.

This study illustrates clearly the discriminatory power of WGS, demonstrating its ability to completely change the paradigm of outbreak investigation. WGS comparisons based on SNPs or cgMLST result in the detection of specific and sensitive potential links between human cases and/or foods that merit further epidemiological investigation. Epidemiological information is essential to support the genetically defined links in outbreak investigations but data to support the epidemiological links between the genetically related strains was not available in this retrospective study.

The analysis showed that sporadic cases can be related (i.e., putative outbreaks) and/or associated to food isolates, even links between sporadic human cases in one country and food in another country were identified. Possible links identified in this way would require full epidemiological investigation in order to support the genetic data. Although this project analysed >1,100 genomes, we did not cover all European countries and all relevant food and clinical isolates. If European wide, real time surveillance is set up in the future, it is likely that more outbreaks will be recognised and investigated. Earlier identification of outbreaks and possible sources will allow for more rapid interventions and the possible prevention of more cases.

One of the many advantages of WGS is that as well as affording high resolution typing and phylogenetic context it provides immediate access to a wealth of additional data. The third part of specific objective 2 was to identify the presence of putative markers conferring the potential to survive/multiply in the food chain and/or cause disease in humans. The LISEQ *L. monocytogenes* genomes were mined to identify genes, or other genetic markers, known to be implicated in antimicrobial resistance and in virulence, and also to identify genes which may play a role in persistence and survival and in the host-specificity of different strains. In terms of antimicrobial resistance there was remarkable low presence of tetracycline (<0.1%) and penicillin (1%) resistance genes. Resistance to detergents and antiseptics via efflux activity was significant with mechanisms detected at a prevalence approaching 20%. Whilst it is encouraging that the isolates in this study show low levels of antimicrobial resistance it is important to remain vigil for emerging resistance. Whole genome sequencing allows antimicrobial resistance monitoring to be done as a rapid cost-neutral activity if WGS is part of routine microbial surveillance and therefore allows this potential threat to be reviewed going forward.

WGS data were also assessed for the presence of 115 putative markers of virulence. More than 80% of markers were present in more than 95% of the isolates suggesting that most putative markers described in the literature are ubiquitous across *L. monocytogenes* lineages I and II. The majority of markers not present in all isolates were over-represented in food and/or lineage II isolates with markers associated with stress survival or cell wall modification being particularly enriched. Conversely, the recently discovered *Listeria* pathogenicity island 3 and the surface protein VIP were more likely to be found in clinical and/or lineage I isolates. Although most virulence markers were present in all strains it is not known if the genes are expressed. Further work is needed including the determination of truncation and non-sense mutations which have been shown to be associated with changes in virulence particularly in the internalin genes (Maury et al., 2016). Several truncations were identified in virulence genes across the dataset with some having an increased propensity for truncation dependent on lineage.

The WGS LISEQ data, for isolates collected over long periods of time from food factories or processing environment, was screened to determine the presence of putative markers conferring the potential to

survive and multiply in the food chain. The presence or absence of genes thought to promote persistence was not found to be useful for predicting persistent phenotype neither was the presence of mutations in these genes. It may be that persistent phenotype is determined by gene expression rather than presence or absence of specific genes (Rychli et al., 2016). It is also possible that genes or markers other than those selected in this study are important in persistence or survival including ones in the accessory genome (Fagerlund et al., 2016). Whilst unable to demonstrate differences in persistent gene markers in the isolates in this study, it was shown that WGS SNP-based analysis is well suited and valuable for investigating persistence and contamination routes within food processing facilities and within the food chain.

WGS data from human and different animal sources was used to identify host specific markers that might be valuable for source attribution by comparing four different genotyping techniques (7-locus MLST, rMLST, cgMLST, cgSNP). Aggregating across all hosts identified how many different loci contributed to host specificity. For 7-locus MLST and cgMLST all loci, and for rMLST most loci contributed towards host differentiation. In contrast, cgSNP, is the only genotyping scheme where the loci comprise individual polymorphisms, and was the only genotyping scheme which identified a subset of the markers across all hosts which differentiated between human and other host sources. Whilst the work here constitutes a small study it suggests that cgSNP, and by extension SNP in the accessory genome, are likely to be the most fruitful source of host-associated polymorphisms, which may be of use in refining molecular attribution models. It is important to note that this study did not investigate the accessory genome, which by definition comprises genes that are not present ubiquitously across the population. Such genes should be a rich source for the discovery of polymorphisms associated with host association, and indeed many other features and deserve to be fully explored.

In conclusion, this study carried out WGS of a large unique collection of *L. monocytogenes* isolates from foods, food processing environments and clinical cases from a large number of European countries. The collection included isolates from foods that were part of the EU-wide baseline survey (BLS) on the prevalence of *L. monocytogenes* in certain RTE foods and highlight the value of revisiting well-structured surveys. This study has demonstrated one of the major benefits of WGS, which is the ability to address a wide range of questions including those on virulence, antimicrobial resistance, source attribution, surveillance and outbreak detection and investigation, in a single experiment. The WGS data generated is now available for additional analysis to address a wide range of questions and thus represents a valuable resource for further studies. The LISEQ isolates have all been typed using current molecular methods and thus can be used to demonstrate the back compatibility of WGS with historical data and also to assess bioinformatic programmes that are able to predict such typing results from WGS data.

This study illustrates one of the major strengths of WGS in comparison to conventional molecular typing methods, which is its ability to provide high quality, unambiguous data. WGS analysis such as cgMLST and cgSNP based typing approaches have been shown to have unparalleled strain typing resolution and it has been demonstrated here how WGS is able to link previously undetected cases to outbreaks and detect clusters of cases that were previously undetected. It has also been shown, however, that as well as cgMLST and cgSNP approaches, that knowledge of the accessory genome can contribute to the interpretation of strain relatedness. The limitations of WGS are less to do with the actual sequencing and the analyses themselves but more dependent on representative sampling of isolates and requirement for good epidemiological data to further investigate genetically linked by WGS. This study supports the use of WGS for *L. monocytogenes* outbreak investigations although analysis of more complex outbreaks would be valuable. However, is difficult to recreate outbreak investigations accurately retrospectively and in order to maximise the advantages of using WGS for outbreak detection it would be highly valuable to use WGS prospectively for the surveillance of listeriosis across Europe.

11. Additional supporting information

Annex A - Excel file: LISEQ_DB.xlsx - Supplementary isolate list with metadata: "Characteristics and descriptive epidemiological information for all *L. monocytogenes* isolates included in the database of the LISEQ tender (WGS tender analysis of *L. monocytogenes* from food and human sources)"

Annex A can be found in the online version of this output ('Supporting information' section: <http://onlinelibrary.wiley.com/doi/10.2903/sp.efsa.2017.EN-1151/abstract>)

References

- Autio T, Keto-Timonen R, Lunden J, Bjorkroth J and Korkeala H, 2003. Characterisation of persistent and sporadic *Listeria monocytogenes* strains by pulsed-field gel electrophoresis (PFGE) and amplified fragment length polymorphism (AFLP). *System of Applied Microbiology*, 26, 539-545.
- Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S and Pribelski AD, 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, 19, 455-477.
- Barbosa J, Magalhães R, Santos I, Ferreira V, Brandão TR, Silva J, Almeida G and Teixeira P, 2013. Evaluation of antibiotic resistance patterns of food and clinical *Listeria monocytogenes* isolates in Portugal. *Foodborne pathogens and disease*, 10, 861-866.
- Batz MB, Hoffmann S and Morris Jr JG, 2012. Ranking the disease burden of 14 pathogens in food sources in the united states using attribution data from outbreak investigations and expert elicitation. *Journal of Food Protection*, 75, 1278-1291.
- Bertrand S, Huys G, Yde M, D'Haene K, Tardy F, Vrints M, Swings J and Collard J-M, 2005. Detection and characterization of tet (M) in tetracycline-resistant *Listeria* strains from human and food-processing origins in Belgium and France. *Journal of medical microbiology*, 54, 1151-1156.
- Bolger AM, Lohse M and Usadel B, 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, btu170.
- Bonsaglia E, Silva N, Júnior AF, Júnior JA, Tsunemi M and Rall V, 2014. Production of biofilm by *Listeria monocytogenes* in different materials and temperatures. *Food Control*, 35, 386-391.
- Bostock M, Patrick E, Russell K and Tarr G, 2016. Circle Plot with Bundled Edges. Package 'edgebundleR'. Available at <https://cran.r-project.org/web/packages/edgebundleR/edgebundleR.pdf>.
- Cabanes D, Dussurget O, Dehoux P and Cossart P, 2004. Auto, a surface associated autolysin of *Listeria monocytogenes* required for entry into eukaryotic cells and virulence. *Molecular microbiology*, 51, 1601-1614.
- Cabanes D, Sousa S, Cebriá A, Lecuit M, García-del Portillo F and Cossart P, 2005. Gp96 is a receptor for a novel *Listeria monocytogenes* virulence factor, Vip, a surface protein. *The EMBO journal*, 24, 2827-2838.
- Camargo AC, Woodward JJ and Nero LA, 2016. The Continuous Challenge of Characterizing the Foodborne Pathogen *Listeria monocytogenes*. *Foodborne pathogens and disease*.
- Camejo A, Carvalho F, Reis O, Leitão E, Sousa S and Cabanes D, 2011. The arsenal of virulence factors deployed by *Listeria monocytogenes* to promote its cell infection cycle. *Virulence*, 2, 379-394.
- Carpentier B and Cerf O, 2011. Review - Persistence of *Listeria monocytogenes* in food industry equipment and premises. *International Journal of Food Microbiology*, 145, 1-8.
- Carvalho F, Atilano ML, Pombinho R, Covas G, Gallo RL, Filipe SR, Sousa S and Cabanes D, 2015. L-Rhamnosylation of *Listeria monocytogenes* wall teichoic acids promotes resistance to antimicrobial peptides by delaying interaction with the membrane. *PLoS Pathog*, 11, e1004919.
- Charpentier E and Courvalin P, 1999. Antibiotic Resistance in *Listeria* spp. *Antimicrobial Agents and Chemotherapy*, 43, 2103-2108.
- Chen J, Luo X, Jiang L, Jin P, Wei W, Liu D and Fang W, 2009. Molecular characteristics and virulence potential of *Listeria monocytogenes* isolates from Chinese food systems. *Food microbiology*, 26, 103-111.

- Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J and Harris SR, 2014. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic acids research*, gku1196.
- Csardi G and Nepusz T, 2006. The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695, 1-9.
- de Castro V, Escudero J, Rodriguez J, Muniozgueren N, Uribarri J, Saez D and Vazquez J, 2012. Listeriosis outbreak caused by Latin-style fresh cheese, Bizkaia, Spain, August 2012. *Euro Surveill*, 17.
- Donovan S, 2015. Listeriosis: a Rare but Deadly Disease. *Clinical Microbiology Newsletter*, 37, 135-140.
- Doumith M, Buchrieser C, Glaser P, Jacquet C and Martin P, 2004. Differentiation of the major *Listeria monocytogenes* serovars by multiplex PCR. *Journal of clinical microbiology*, 42, 3819-3822.
- EFSA, 2013. Analysis of the baseline survey on the prevalence of *Listeria monocytogenes* in certain ready-to-eat (RTE) foods in the EU, 2010-2011 Part A: *Listeria monocytogenes* prevalence estimates. *EFSA Journal*, 11, 75 pp.
- EFSA, 2014. Technical specifications for the pilot on the collection of data on molecular testing of food-borne pathogens from food, feed and animal samples. *EFSA Supporting Publications*, 11, 712E-.
- EFSA and ECDC, 2015. The European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2014. *EFSA Journal*, 13, 4329, doi:4310.2903/j.efsa.2015.4329.
- Engberg J, 2006. Contributions to the epidemiology of *Campylobacter* infections. *Dan Med Bull*, 53, 361-389.
- European Commission, 2003. Directive 2003/99/EC of The European Parliament and of the Council of 17 November 2003 on the monitoring of zoonoses and zoonotic agents, amending Council Decision 90/424/EEC and repealing Council Directive 92/117/EEC. *Official Journal of the European Union L*, 325, 31-40.
- European Food Safety Authority, 2010. Panel on Biological Hazards (BIOHAZ); Scientific opinion on quantification of the risk posed by broiler meat to human campylobacteriosis in the EU. *EFSA Journal*, 8, 1437.
- Fagerlund A, Langsrud S, Schirmer BC, Møretrø T and Heir E, 2016. Genome Analysis of *Listeria monocytogenes* Sequence Type 8 Strains Persisting in Salmon and Poultry Processing Environments and Comparison with Related Strains. *PLoS one*, 11, e0151117.
- Félix B, Danan C, Van Walle I, Lailier R, Texier T, Lombard B, Brisabois A and Roussel S, 2014. Building a molecular *Listeria monocytogenes* database to centralize and share PFGE typing data from food, environmental and animal strains throughout Europe. *Journal of microbiological methods*, 104, 1-8.
- Felix B, Mariet J, Maillet A, Firmesse O, Radomski N, Felten A, Touzain F, Mistou M and Roussel S, 2015. Genomic insight to understand the persistence of *Listeria monocytogenes* strains in processing environments of pork products. *Proceedings of the Safe Pork*.
- Franz E, Delaquis P, Morabito S, Beutin L, Gobius K, Rasko DA, Bono J, French N, Osek J and Lindstedt B-A, 2014. Exploiting the explosion of information associated with whole genome sequencing to tackle Shiga toxin-producing *Escherichia coli* (STEC) in global food production systems. *International Journal of Food Microbiology*, 187, 57-72.
- Gallagher D, Pouillot R, Hoelzer K, Tang J, Dennis SB and Kause JR, 2016. *Listeria monocytogenes* in Retail Delicatessens: An Interagency Risk Assessment—Risk Mitigations. *Journal of Food Protection*, 79, 1076-1088.

- Gandhi M and Chikindas ML, 2007. *Listeria*: a foodborne pathogen that knows how to survive. *International Journal of Food Microbiology*, 113, 1-15.
- Gelman A and Rubin DB, 1992. Inference from iterative simulation using multiple sequences. *Statistical science*, 457-472.
- Gillesberg Lassen S, Ethelberg S, Björkman JT, Jensen T, Sørensen G, Kvistholm Jensen A, Müller L, Nielsen EM and Mølbak K, 2016. Two listeria outbreaks caused by smoked fish consumption—using whole-genome sequencing for outbreak investigations. *Clinical Microbiology and Infection*.
- Granier SA, Moubareck C, Colaneri C, Lemire A, Roussel S, Dao T-T, Courvalin P and Brisabois A, 2011. Antimicrobial resistance of *Listeria monocytogenes* isolates from food and the environment in France over a 10-year period. *Applied and environmental microbiology*, 77, 2788-2790.
- Graves LM 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.
- Graves LM, Swaminathan B and Hunter SB, 2007. *Listeria*, listeriosis, and food safety. In: *Food science and technology*. Eds Ryser ET and Marth EH. 3rd, CRC Press, Boca Raton, 873 p.
- Greig JD and Ravel A, 2009. Analysis of foodborne outbreak data reported internationally for source attribution. *International Journal of Food Microbiology*, 130, 77-87.
- Hald T, Vose D, Wegener HC and Koupeev T, 2004. A Bayesian Approach to Quantify the Contribution of Animal-Food Sources to Human Salmonellosis. *Risk Analysis*, 24, 255-269.
- Hayward MR, Petrovska L, Jansen VA and Woodward MJ, 2016. Population structure and associated phenotypes of *Salmonella enterica* serovars Derby and Mbandaka overlap with host range. *BMC microbiology*, 16, 1.
- Heck KL, van Belle G and Simberloff D, 1975. Explicit calculation of the rarefaction diversity measurement and the determination of sufficient sample size. *Ecology*, 56, 1459-1461.
- Heiman KE, Garalde VB, Gronostaj M, Jackson KA, Beam S, Joseph L, Saupe A, Ricotta E, Waechter H and Wellman A, 2015. Multistate outbreak of listeriosis caused by imported cheese and evidence of cross-contamination of other cheeses, USA, 2012. *Epidemiology and Infection*, 1-11.
- Hingston PA, Piercey MJ and Truelstrup Hansen L, 2015. Genes associated with desiccation and osmotic stress in *Listeria monocytogenes* as revealed by insertional mutagenesis. *Applied and environmental microbiology*, 81, 5350-5362.
- Holch A, Webb K, Lukjancenko O, Ussery D, Rosenthal BM and Gram L, 2013. Genome sequencing identifies two nearly unchanged strains of persistent *Listeria monocytogenes* isolated at two different fish processing plants sampled 6 years apart. *Applied and environmental microbiology*, 79, 2944-2951.
- Jackson BR, Tarr C, Strain E, Jackson KA, Conrad A, Carleton H, Katz LS, Stroika S, Gould LH and Mody RK, 2016. Implementation of Nationwide Real-Time Whole-Genome Sequencing to Enhance Listeriosis Outbreak Detection and Investigation. *Clinical Infectious Diseases*, ciw242.
- Jamali H, Paydar M, Ismail S, Looi CY, Wong WF, Radmehr B and Abedini A, 2015. Prevalence, antimicrobial susceptibility and virulotyping of *Listeria* species and *Listeria monocytogenes* isolated from open-air fish markets. *BMC microbiology*, 15, 1.
- Jessen B and Lammert L, 2003. Biofilm and disinfection in meat processing plants. *International Biodeterioration & Biodegradation - Hygiene and Disinfection*, 51, 265-269.
- Jolley KA and Maiden MC, 2010. BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC bioinformatics*, 11, 1.

- Khen B, Lynch O, Carroll J, McDowell D and Duffy G, 2015. Occurrence, antibiotic resistance and molecular characterization of *Listeria monocytogenes* in the beef chain in the Republic of Ireland. *Zoonoses and public health*, 62, 11-17.
- Kvistholm Jensen A, Nielsen EM, Björkman JT, Jensen T, Müller L, Persson S, Bjerager G, Perge A, Krause TG, Kiil K, Sørensen G, Andersen JK, Mølbak K and Ethelberg S, 2016. Whole-genome Sequencing Used to Investigate a Nationwide Outbreak of Listeriosis Caused by Ready-to-eat Delicatessen Meat, Denmark, 2014. *Clinical Infectious Diseases*, 63, 64-70.
- Kwong JC, Stafford R, Strain E, Stinear TP, Seemann T and Howden BP, 2016. Sharing is caring: international sharing of data enhances genomic surveillance of *Listeria monocytogenes*. *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America*.
- Lamden KH, Fox AJ, Amar CF and Little CL, 2013. A case of foodborne listeriosis linked to a contaminated food-production process. *J Med Microbiol*, 62, 1614-1616.
- Langmead B and Salzberg SL, 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9, 357-359.
- Li H and Durbin R, 2010. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, 26, 589-595.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G and Durbin R, 2009. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25, 2078-2079.
- Little CL, Pires SM, Gillespie IA, Grant K and Nichols GL, 2010. Attribution of human *Listeria monocytogenes* infections in England and Wales to ready-to-eat food sources placed on the market: adaptation of the Hald Salmonella source attribution model. *Foodborne pathogens and disease*, 7, 749-756.
- Lourenço A, de Las Heras A, Scotti M, Vazquez-Boland J, Frank JF and Brito L, 2013. Comparison of *Listeria monocytogenes* exoproteomes from biofilm and planktonic state: Lmo2504, a protein associated with biofilms. *Applied and environmental microbiology*, 79, 6075-6082.
- Lundén JM, Autio TJ and Korkeala HJ, 2002. Transfer of persistent *Listeria monocytogenes* contamination between food-processing plants associated with a dicing machine. *Journal of Food Protection*, 65, 1129-1133.
- Lungu B, O'Bryan CA, Muthaiyan A, Milillo SR, Johnson MG, Crandall PG and Ricke SC, 2011. *Listeria monocytogenes*: antibiotic resistance in food production. *Foodborne pathogens and disease*, 8, 569-578.
- Manly BF, 2007. Randomization, bootstrap and Monte Carlo methods in biology. CRC Press, pp.
- Maury MM, Tsai Y-H, Charlier C, Touchon M, Chenal-Francisque V, Leclercq A, Criscuolo A, Gaultier C, Roussel S and Brisabois A, 2016. Uncovering *Listeria monocytogenes* hypervirulence by harnessing its biodiversity. *Nature genetics*, 48, 308–313.
- Mazza R, Mazzette R, McAuliffe O, Jordan K and Fox EM, 2015. Differential gene expression of three gene targets among persistent and nonpersistent *Listeria monocytogenes* strains in the presence or absence of benzethonium chloride. *Journal of Food Protection*, 78, 1569-1573.
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S and Daly M, 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, 20, 1297-1303.
- Melo J, Andrew PW and Faleiro ML, 2015. *Listeria monocytogenes* in cheese and the dairy environment remains a food safety challenge: The role of stress responses. *Food Research International*, 67, 75-90.
- Mettler E and Carpentier B, 1999. Hygienic quality of floors in relation to surface texture. *Transaction of the Institute of Chemical Engineers*, 77, 90-96.

- Morganti M, Scaltriti E, Cozzolino P, Bolzoni L, Casadei G, Pierantoni M, Foni E and Pongolini S, 2016. Processing-dependent and clonal contamination patterns of *Listeria monocytogenes* in the cured ham food chain revealed by genetic analysis. *Applied and environmental microbiology*, 82, 822-831.
- Morvan A, Moubareck C, Leclercq A, Herve-Bazin M, Bremont S, Lecuit M, Courvalin P and Le Monnier A, 2010. Antimicrobial resistance of *Listeria monocytogenes* strains isolated from humans in France. *Antimicrobial Agents and Chemotherapy*, 54, 2728-2731.
- Moura A, Criscuolo A, Pouseele H, Maury MM, Leclercq A, Tarr C, Björkman JT, Dallman T, Reimer A, Enouf V, Larsonneur E, Carleton H, Bracq-Dieye H, Katz LS, Jones L, Touchon M, Tourdjman M, Walker M, Stroika S, Cantinelli T, Chenal-Francisque V, Kucerova Z, Rocha EPC, Nadon C, Grant K, Nielsen EM, Pot B, Gerner-Smidt P, Lecuit M and Brisse S, 2016. Whole genome-based population biology and epidemiological surveillance of *Listeria monocytogenes*. *Nature Microbiology*, 2, 16185.
- Mughini-Gras L and van Pelt W, 2014. Salmonella source attribution based on microbial subtyping: Does including data on food consumption matter? *International Journal of Food Microbiology*, 191, 109-115.
- Muhterem-Uyar M, Dalmasso M, Bolocan AS, Hernandez M, Kapetanakou AE, Kuchta T, Manios SG, Melero B, Minarovičová J and Nicolau AI, 2015. Environmental sampling for *Listeria monocytogenes* control in food processing facilities reveals three contamination scenarios. *Food Control*, 51, 94-107.
- Mullner P, Spencer SE, Wilson DJ, Jones G, Noble AD, Midwinter AC, Collins-Emerson JM, Carter P, Hathaway S and French NP, 2009. Assigning the source of human campylobacteriosis in New Zealand: a comparative genetic and epidemiological approach. *Infection, Genetics and Evolution*, 9, 1311-1319.
- Nei M, 1975. *Molecular population genetics and evolution*. North-Holland Publishing Company., pp.
- Orsi RH, den Bakker HC and Wiedmann M, 2011. *Listeria monocytogenes* lineages: Genomics, evolution, ecology, and phenotypic characteristics. *International Journal of Medical Microbiology*, 301, 79-96.
- Paquet C, Coulombier D, Kaiser R and Ciotti M, 2005. Epidemic intelligence: a new framework for strengthening disease surveillance in Europe. *Euro surveillance: bulletin europeen sur les maladies transmissibles= European communicable disease bulletin*, 11, 212-214.
- Perez-Trallero E, Zigorraga C, Artieda J, Alkorta M and Marimon JM, 2014. Two outbreaks of *Listeria monocytogenes* infection, Northern Spain. *Emerg Infect Dis*, 20, 2155-2157.
- Pires SM, Evers EG, van Pelt W, Ayers T, Scallan E, Angulo FJ, Havelaar A and Hald T, 2009. Attributing the human disease burden of foodborne infections to specific sources. *Foodborne pathogens and disease*, 6, 417-424.
- Pouillot R, Gallagher D, Tang J, Hoelzer K, Kause J and Dennis SB, 2015. *Listeria monocytogenes* in Retail Delicatessens: An Interagency Risk Assessment—Model and Baseline Results. *Journal of Food Protection®*, 78, 134-145.
- Pouillot R, Goulet V, Delignette-Muller ML, Mahé A and Cornu M, 2009. Quantitative risk assessment of *Listeria monocytogenes* in french cold-smoked Salmon: II. Risk characterization. *Risk Analysis*, 29, 806-819.
- Pritchard JK, Stephens M and Donnelly P, 2000. Inference of population structure using multilocus genotype data. *Genetics*, 155, 945-959.
- Promadej N, Fiedler F, Cossart P, Dramsi S and Kathariou S, 1999. Cell Wall Teichoic Acid Glycosylation in *Listeria monocytogenes* Serotype 4b Requires *gtcA*, a Novel, Serogroup-Specific Gene. *Journal of bacteriology*, 181, 418-425.

- 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.
- Reij M, Den Aantrekker E and ILSI Europe Risk Analysis in Microbiology Task Force, 2004. Recontamination as a source of pathogens in processed foods. *International Journal of Food Microbiology*, 91, 1-11.
- Renier S, Hébraud M and Desvaux M, 2011. Molecular biology of surface colonization by *Listeria monocytogenes*: an additional facet of an opportunistic Gram-positive foodborne pathogen. *Environmental microbiology*, 13, 835-850.
- Ryan S, Begley M, Hill C and Gahan C, 2010. A five-gene stress survival islet (SSI-1) that contributes to the growth of *Listeria monocytogenes* in suboptimal conditions. *Journal of Applied Microbiology*, 109, 984-995.
- Rychli K, Grunert T, Ciolacu L, Zaiser A, Razzazi-Fazeli E, Schmitz-Esser S, Ehling-Schulz M and Wagner M, 2016. Exoproteome analysis reveals higher abundance of proteins linked to alkaline stress in persistent *Listeria monocytogenes* strains. *International Journal of Food Microbiology*, 218, 17-26.
- Rychli K, Müller A, Zaiser A, Schoder D, Allerberger F, Wagner M and Schmitz-Esser S, 2014. Genome sequencing of *Listeria monocytogenes* "Quargel" listeriosis outbreak strains reveals two different strains with distinct in vitro virulence potential. *PLoS one*, 9, e89964.
- Rückerl I, Muhterem-Uyar M, Muri-Klinger S, Wagner K-H, Wagner M and Stessl B, 2014. *L. monocytogenes* in a cheese processing facility: Learning from contamination scenarios over three years of sampling. *International Journal of Food Microbiology*, 189, 98-105.
- Sahl JW, Lemmer D, Travis J, Schupp J, Gillece J, Aziz M, Driebe E, Drees K, Hicks N and Williamson C, 2016. The Northern Arizona SNP Pipeline (NASP): accurate, flexible, and rapid identification of SNPs in WGS datasets. *bioRxiv*, 037267.
- Schmid B, Klumpp J, Raimann E, Loessner MJ, Stephan R and Tasara T, 2009. Role of cold shock proteins in growth of *Listeria monocytogenes* under cold and osmotic stress conditions. *Applied and environmental microbiology*, 75, 1621-1627.
- Seemann T, 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, btu153.
- Sheppard SK, Dallas JF, Strachan NJ, MacRae M, McCarthy ND, Wilson DJ, Gormley FJ, Falush D, Ogden ID and Maiden MC, 2009. *Campylobacter* genotyping to determine the source of human infection. *Clinical Infectious Diseases*, 48, 1072-1078.
- Sheppard SK, Didelot X, Méric G, Torralbo A, Jolley KA, Kelly DJ, Bentley SD, Maiden MC, Parkhill J and Falush D, 2013. Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in *Campylobacter*. *Proceedings of the National Academy of Sciences*, 110, 11923-11927.
- Simpson EH, 1949. Measurement of diversity. *Nature*, 163, 688-688.
- Smid JH, Gras LM, de Boer AG, French NP, Havelaar AH, Wagenaar JA and van Pelt W, 2013. Practicalities of using non-local or non-recent multilocus sequence typing data for source attribution in space and time of human campylobacteriosis. *PLoS one*, 8, e55029.
- Stamatakis A, 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30, 1312-1313.
- Stasiewicz MJ, Oliver HF, Wiedmann M and den Bakker HC, 2015. Whole-genome sequencing allows for improved identification of persistent *Listeria monocytogenes* in food-associated environments. *Applied and environmental microbiology*, 81, 6024-6037.

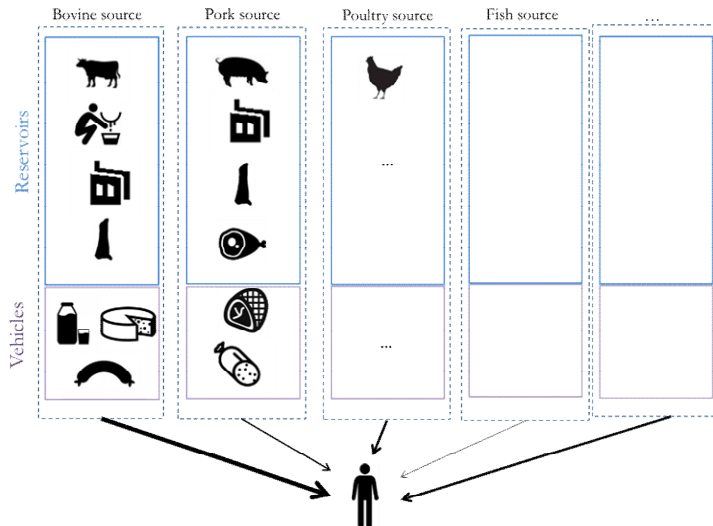
- Strachan NJ, Rotariu O, MacRae M, Sheppard SK, Smith-Palmer A, Cowden J, Maiden MC and Forbes KJ, 2013. Operationalising factors that explain the emergence of infectious diseases: a case study of the human campylobacteriosis epidemic. *PLoS one*, 8, e79331.
- Swaminathan B, Gerner-Smidt P, Ng L-K, Lukinmaa S, Kam K-M, Rolando S, Gutiérrez EP and Binsztein N, 2006. Building PulseNet International: an interconnected system of laboratory networks to facilitate timely public health recognition and response to foodborne disease outbreaks and emerging foodborne diseases. *Foodborne Pathogens & Disease*, 3, 36-50.
- Tamura K, Stecher G, Peterson D, Filipski A and Kumar S, 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. *Molecular biology and evolution*, 30, 2725-2729.
- Tewelde R, Dallman T, Schaefer U, Sheppard CL, Ashton P, Pichon B, Ellington M, Swift C, Green J and Underwood A (PeerJ Preprints), 2016. MOST: A modified MLST typing tool based on short read sequencing. 2167-9843.
- Thévenot D, Dernburg A and Vernozy-Rozand C, 2006. An updated review of *Listeria monocytogenes* in the pork meat industry and its products. *Journal of Applied Microbiology*, 101, 7-17.
- Tocmo R, Krizman K, Khoo WJ, Phua LK, Kim M and Yuk HG, 2014. *Listeria monocytogenes* in Vacuum-Packed Smoked Fish Products: Occurrence, Routes of Contamination, and Potential Intervention Measures. *Comprehensive Reviews in Food Science and Food Safety*, 13, 172-189.
- Toledo-Arana A, Dussurget O, Nikitas G, Sesto N, Guet-Revillet H, Balestrino D, Loh E, Gripenland J, Tiensuu T and Vaitkevicius K, 2009. The *Listeria* transcriptional landscape from saprophytism to virulence. *Nature*, 459, 950-956.
- Toma B, Vaillancourt J-P, Dufour B, Eloit M, Moutou F, Marsh W, Bénet J, Sanaa M and Michel P, 1999. *Dictionary of veterinary epidemiology*. Iowa State University Press, pp.
- Tourdjman M, Leroux B, Leclercq A, Laurent E, Chenal-Francisque V, King L, Loyer S, Vaillant V, Donguy M-P, Lecuit M and de Valk H, 2014. Épidémie d'infections à *Listeria monocytogenes* liée à la consommation de brie au lait cru – France, 2012. *Institut de veille sanitaire*, 15 p.
- Travier L, Guadagnini S, Gouin E, Dufour A, Chenal-Francisque V, Cossart P, Olivo-Marin J-C, Ghigo J-M, Disson O and Lecuit M, 2013. ActA promotes *Listeria monocytogenes* aggregation, intestinal colonization and carriage. *PLoS Pathog*, 9, e1003131.
- Treangen TJ, Ondov BD, Koren S and Phillippy AM, 2014. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome biology*, 15, 1.
- Tremoulet F, Duche O, Namane A, Martinie B, Labadie J and Consortium ELG, 2002. Comparison of protein patterns of *Listeria monocytogenes* grown in biofilm or in planktonic mode by proteomic analysis. *FEMS microbiology letters*, 210, 25-31.
- Wang Q, Holmes N, Martinez E, Howard P, Hill-Cawthorne G and Sintchenko V, 2015. It Is Not All about Single Nucleotide Polymorphisms: Comparison of Mobile Genetic Elements and Deletions in *Listeria monocytogenes* Genomes Links Cases of Hospital-Acquired Listeriosis to the Environmental Source. *J Clin Microbiol*, 53, 3492-3500.
- Vázquez-Boland JA, Kuhn M, Berche P, Chakraborty T, Domínguez-Bernal G, Goebel W, González-Zorn B, Wehland J and Kreft J, 2001. *Listeria* pathogenesis and molecular virulence determinants. *Clinical microbiology reviews*, 14, 584-640.
- Wieczorek K, Dmowska K and Osek J, 2012. Prevalence, characterization, and antimicrobial resistance of *Listeria monocytogenes* isolates from bovine hides and carcasses. *Applied and environmental microbiology*, 78, 2043-2045.
- Wilson DJ, Gabriel E, Leatherbarrow AJ, Cheesbrough J, Gee S, Bolton E, Fox A, Fearnhead P, Hart CA and Diggle PJ, 2008. Tracing the source of campylobacteriosis. *PLoS Genet*, 4, e1000203.

Wulff G, Gram L, Ahrens P and Fonnesbech Vogel B, 2006. One group of genetically similar *Listeria monocytogenes* strains frequently dominates and persists in several fish slaughter- and smokehouses. *Applied and environmental microbiology*, 72, 4313-4322.

Yde M, Naranjo M, Mattheus W, Stragier P, Pochet B, Beulens K, De Schrijver K, Van den Branden D, Laisnez V and Flipse W, 2012. Usefulness of the European Epidemic Intelligence Information System in the management of an outbreak of listeriosis, Belgium, 2011. *Eurosurveillance*, 17.

Glossary

Food vehicle: Food involved in transmitting a pathogen to a receptive host (RTE foods for *L. monocytogenes*).



Food-borne outbreak: Incidence, observed under given circumstances, of two or more human cases of the same disease and/or infection, or a situation in which the observed number of cases exceeds the expected number and where the cases are linked, or are probably linked, to the same food source (European Commission, 2003).

Reservoir: An animate (humans, animals, insects etc.) or inanimate object (plant, raw milk, soil, surface in contact with food, etc.) or any combination of these serving as a habitat of a pathogen that produces itself in such a way as to be transmitted to a susceptible host (Toma et al., 1999; European Food Safety Authority, 2010).

Source: Origin of the pathogen causing infection, including reservoirs and food vehicles.

Source attribution: Partitioning of the human disease burden of one or more foodborne infections to specific sources, where the term source includes animal reservoirs and vehicles (e.g. foods) (Pires et al., 2009)

Sporadic case: Case that has not been associated with known outbreaks (Engberg, 2006)

LISEQ database glossary

Best date: gives the closest date from sampling available.

Clinical symptoms: gives the symptoms presented by the patient at the time of isolation.

Context: describes in which framework the sampling was carried out, which may be an outbreak investigation or a specific research project.

Context level 1: gives a linear numbering of the nine outbreak strains used in this project.

Conventional serotyping: gives the serotyped obtained from serologic agglutination according to the method of Seeliger and Höhne (1979).

Date of sampling: gives the date when the sample was taken from which listeria was isolated.

EFSA_Code.MTX.mapping: gives the mapping with the EFSA standard sample description 2 with specific distinction between Food products isolates and Food processing environment isolates (EFSA, 2013a).

EFSA_Complete.MTX.mapping: gives the mapping with the EFSA Foodex 2 language (EFSA, 2013a).

Food matrix: the major food categories were defined according to the classification of EFSA risk-food matrices (EFSA, 2013).

Food origin: describes the type of animal or vegetable that is the main component of the food product. Composite food products including more than one animal or vegetable, "Mixed sources" is specified. It refers to source as described in the glossary under Section 12.

Food origin level 1: specifies the type of fish species when documented.

Food product: specific category describing the type of products further obtained from a given "food matrix". These definitions follow the EFSA guidance on listeria risk (EFSA, 2012)

Geographic information: provide information on the geographic area of sampling.

Molecular serotyping: gives the serotype according to the method of the EURL for Lm. This method respects the international reference method established by Doumith et al. (2004).

Reception date: gives the date when the strain or the sample was received in the laboratory. It may be different from the sampling date.

Sample type: describes the isolation of the strains from food product (EFSA code: S019A Food sample) or from food processing environment (EFSA code: S027A Environmental sample). Food processing environments include all types of samples (food contact surfaces or non-food contact surfaces) obtained from the place where the food product is processed.

Sampling stage: gives the level in the food chain where the sample was taken.

Sector: distinguishes clinical samples isolated from human pathology and non-human isolates.

List of abbreviations used in the report

7-MLST: 7 locus multi-locus sequence typing

BLS: base line survey

CC: clonal complex

cgMLST: core genome multi-locus sequence typing

EFSA: European Food Safety Agency

ECDC: European Centre of Disease Prevention and Control

fAFLP: fluorescent amplified-fragment length polymorphism

GWAS: genome-wide association study

IP scheme: Institute Pasteur scheme

MLST: multi-locus sequence typing

MOST: Metric oriented sequence typer

rMLST: ribosomal multi locus sequence typing

PFGE: pulsed-field gel electrophoresis

SNP: single-nucleotide polymorphism

ST: sequence type

TESSy: the European surveillance system

Appendix 1: Isolates from the EU-wide baseline survey on prevalence of *L. monocytogenes* in certain RTE foods conducted in 2010-2012

RL_number	Country	Year	Sample type	Description	MLST Clonal complexes	MLST sequence types
RL15000006	C	2010	Food products	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-Smoked processing not specified	CC121	121
RL15000007	C	2010	Food products	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-Smoked processing not specified	CC204	204
RL15000008	C	2011	Food products	Fish and fishery products-Fish-Clupea harengus (Herring, kipper)-Wild fish-Smoked processing not specified	CC14	14
RL15000009	C	2011	Food products	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-Smoked processing not specified	CC121	121
RL15000010	C	2011	Food products	Meat and meat products-Mixed sources--Deli products - Pate-Cooked	CC121	121
RL15000011	C	2011	Food products	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-	CC121	121
RL15000012	C	2011	Food products	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-	CC9	9
RL15000013	C	2011	Food products	Milk and milk products-Bovine--Soft cheese-Not specified	CC20	20
RL15000014	C	2011	Food products	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-Smoked processing not specified	CC121	121
RL15000015	C	2011	Food products	Fish and fishery products-Fish-Clupea harengus (Herring, kipper)-Wild fish-Smoked processing not specified	CC121	121
RL15000016	C	2011	Food products	Fish and fishery products-Fish-Oncorhynchus mykiss, Salmo trutta (Trout)-Fish origin not specified-Smoked processing not specified	CC101	101
RL15000017	C	2011	Food products	Fish and fishery products-Fish-Clupea harengus (Herring, kipper)-Wild fish-Smoked processing not specified	CC8	8
RL15000018	C	2011	Food products	Fish and fishery products-Fish-Clupea harengus (Herring, kipper)-Wild fish-Filet	CC121	121
RL15000019	C	2011	Food products	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-Smoked processing not specified	CC121	121
RL15000020	C	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked processing not specified	CC204	204
RL15000021	C	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked processing not specified	CC204	204
RL15000022	C	2011	Food products	Meat and meat products-Mixed sources--Deli products - Pate-	CC204	204
RL15000023	C	2011	Food products	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-Smoked processing not specified	CC121	121
RL15000024	C	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked processing not specified	CC204	204
RL15000025	R	2011	Food products	Meat and meat products-Unspecified--Deli product - Sliced-Cooked	CC5	5
RL15000026	R	2011	Food products	Meat and meat products-Unspecified--Deli products - Pate-	CC121	121
RL15000027	R	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC155	155
RL15000028	R	2011	Food products	Meat and meat products-Unspecified--Deli product - Other product-Other stabilization	CC9	9

RL_number	Country	Year	Sample type	Description	MLST Clonal complexes	MLST sequence types
RL1500029	F	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC8	8
RL1500030	F	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Warm smoked	CC3	3
RL1500031	F	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC121	121
RL1500032	F	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Gravad/slightly salted	CC3	3
RL1500033	F	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC20	20
RL1500034	F	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC20	20
RL1500035	F	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC9	9
RL1500036	F	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC2	2
RL1500037	F	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC20	20
RL1500038	F	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Gravad/slightly salted	CC121	121
RL1500039	F	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC9	9
RL1500040	F	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC155	155
RL1500041	F	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC3	3
RL1500042	F	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC121	121
RL1500043	F	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC101	101
RL1500044	G	2011	Food products	Meat and meat products-Swine--Deli product - Sliced-Cooked	CC204	204
RL1500045	G	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked processing not specified	CC9	9
RL1500046	G	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Gravad/slightly salted	ST124	124
RL1500047	G	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked processing not specified	ST124	124
RL1500048	N	2011	Food products	Fish and fishery products-Fish--Wild fish-Warm smoked	CC6	6
RL1500049	N	2011	Food products	Fish and fishery products-Fish--Wild fish-Cold smoked	CC155	155
RL1500050	N	2011	Food products	Fish and fishery products-Fish--Wild fish-Warm smoked	CC155	155
RL1500051	N	2011	Food products	Fish and fishery products-Fish--Wild fish-Cold smoked	CC121	121
RL1500052	N	2011	Food products	Fish and fishery products-Fish--Wild fish-Warm smoked	CC8	8
RL1500053	N	2011	Food products	Fish and fishery products-Fish--Wild fish-Warm smoked	CC155	155
RL1500054	N	2011	Food products	Fish and fishery products-Fish--Wild fish-Cold smoked	CC9	9
RL1500055	N	2011	Food products	Fish and fishery products-Fish--Wild fish-Warm smoked	CC121	121
RL1500056	N	2011	Food products	Fish and fishery products-Fish--Wild fish-Cold smoked	CC6	6
RL1500057	S	2011	Food products	Fish and fishery products-Fish--Farmed fish-Cold smoked	CC8	8
RL1500058	S	2011	Food products	Fish and fishery products-Fish--Wild fish-Cold smoked	CC2	2
RL1500059	S	2011	Food products	Fish and fishery products-Fish--Farmed fish-Cold smoked	CC8	8

RL_number	Country	Year	Sample type	Description	MLST Clonal complexes	MLST sequence types
RL15000060	S	2011	Food products	Fish and fishery products-Fish--Wild fish-Gravad/slightly salted	CC121	121
RL15000061	B	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked processing not specified	CC121	121
RL15000062	B	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked processing not specified	CC121	121
RL15000063	B	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Gravad/slightly salted	CC121	121
RL15000064	B	2011	Food products	Fish and fishery products-Fish--Wild fish-Smoked processing not specified	CC31	31
RL15000065	U	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Warm smoked	CC3	3
RL15000066	U	2010	Food products	Meat and meat products-Turkeys--Deli product - Sliced-Cooked	CC7	12
RL15000067	U	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked processing not specified	CC8	551
RL15000068	U	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked processing not specified	CC9	9
RL15000069	U	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked processing not specified	CC1	1
RL15000070	U	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked processing not specified	CC9	9
RL15000071	U	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC14	14
RL15000072	U	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked processing not specified	CC7	12
RL15000073	U	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Gravad/slightly salted	CC403	403
RL15000074	U	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC9	9
RL15000076	U	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Warm smoked	CC193	193
RL15000077	U	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC9	9
RL15000078	U	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC9	9
RL15000079	U	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Warm smoked	CC3	3
RL15000080	U	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked processing not specified	CC155	155
RL15000081	U	2011	Food products	Meat and meat products-Turkeys--Deli product - Sliced-Cooked	CC3	3
RL15000082	U	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked processing not specified	CC3	3
RL15000083	U	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked processing not specified	CC14	14
RL15000084	U	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked processing not specified	CC8	8
RL15000085	U	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked processing not specified	CC121	121
RL15000086	U	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked processing not specified	CC3	3
RL15000087	U	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC9	9
RL15000088	U	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC8	8
RL15000089	U	2011	Food products	Meat and meat products-Swine--Deli product - Sliced-Cooked	CC7	12
RL15000090	U	2011	Food products	Meat and meat products-Mixed sources--Deli product - Sliced-Cooked	CC87	87
RL15000091	U	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC121	121

RL_number	Country	Year	Sample type	Description	MLST Clonal complexes	MLST sequence types
RL1500092	U	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC155	155
RL1500093	U	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC121	121
RL1500094	U	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked processing not specified	CC8	8
RL1500095	U	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked processing not specified	CC8	8
RL1500096	U	2011	Food products	Milk and milk products-Unspecified--Semi soft cheese-Made from pasteurized milk	CC14	91
RL1500097	U	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked processing not specified	CC8	8
RL1500098	U	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Warm smoked	CC121	121
RL1500099	U	2011	Food products	Meat and meat products-Swine--Deli product - Sliced-Cooked	CC7	12
RL1500100	U	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked processing not specified	CC8	8
RL1500101	U	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked processing not specified	CC8	8
RL1500102	U	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Gravad/slightly salted	CC14	14
RL1500103	U	2011	Food products	Meat and meat products-Gallus gallus (fowl)--Deli products - Pate-Other stabilization	CC2	145
RL1500104	U	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC155	155
RL1500105	U	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked processing not specified	CC8	8
RL1500106	U	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked processing not specified	CC8	8
RL1500107	U	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC9	9
RL1500108	U	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked processing not specified	CC3	3
RL1500109	U	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked processing not specified	CC8	8
RL1500110	U	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC155	155
RL1500111	U	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked processing not specified	CC155	155
RL1500112	U	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC155	155
RL1500113	U	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Warm smoked	CC403	403
RL1500114	U	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked processing not specified	CC204	204
RL1500115	U	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC8	8
RL1500116	U	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked processing not specified	CC403	403
RL1500117	U	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC14	14
RL1500118	U	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC9	9
RL1500119	U	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked processing not specified	CC9	9
RL1500120	U	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked processing not specified	CC8	8
RL1500121	U	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC121	121
RL1500122	U	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC155	155

RL_number	Country	Year	Sample type	Description	MLST Clonal complexes	MLST sequence types
RL15000123	U	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC155	155
RL15000124	U	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC9	9
RL15000125	U	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC155	155
RL15000126	U	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC8	8
RL15000127	U	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked processing not specified	CC6	6
RL15000128	E	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-	CC121	121
RL15000129	E	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-	CC121	121
RL15000130	E	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-	CC121	121
RL15000131	E	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-	CC204	204
RL15000132	E	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-	CC8	8
RL15000133	E	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-	CC101	101
RL15000134	Z	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC121	121
RL15000135	Z	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC6	6
RL15000136	Z	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC9	9
RL15000137	Z	2011	Food products	Meat and meat products-Swine--Deli product - Sliced-Cooked	CC9	9
RL15000138	Z	2011	Food products	Meat and meat products-Swine--Deli product - Sliced-Cooked	CC9	9
RL15000139	Z	2011	Food products	Meat and meat products-Swine--Deli product - Sliced-Cooked	CC9	9
RL15000140	Z	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC6	6
RL15000141	Z	2011	Food products	Meat and meat products-Swine--Deli product - Sliced-Cooked	CC8	120
RL15000142	Z	2011	Food products	Fish and Fishery products-Fish--Fish origin not specified-Cold smoked	CC121	121
RL15000143	Z	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC9	9
RL15000144	Z	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC19	173
RL15000145	Z	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked processing not specified	CC121	121
RL15000146	Z	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC9	9
RL15000147	Z	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC9	9
RL15000148	Z	2011	Food products	Fish and fishery products-Fish--Wild fish-Gravad/slightly salted	CC121	121
RL15000149	C	2010	Food products	Fish and fishery products-Fish-Clupea harengus (Kipper)-Fish origin not specified-Filet	CC8	16
RL15000150	C	2010	Food products	Fish and fishery products-Fish-Salmo spp. (Salmon)-Farmed fish-	CC5	5
RL15000151	C	2010	Food products	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-Raw	CC9	9
RL15000152	A	2010	Food products	Fish and fishery products-Fish--Wild fish-Smoked processing not specified	CC101	101
RL15000153	A	2010	Food products	Meat and meat products-Unspecified--Deli product - Sliced-Cooked	CC31	31

RL_number	Country	Year	Sample type	Description	MLST Clonal complexes	MLST sequence types
RL15000154	A	2010	Food products	Fish and fishery products-Fish--Wild fish-Smoked processing not specified	CC121	121
RL15000155	A	2010	Food products	Fish and fishery products-Fish--Wild fish-Smoked processing not specified	CC9	9
RL15000156	A	2011	Food products	Fish and fishery products-Fish--Wild fish-Smoked processing not specified	CC121	121
RL15000157	A	2011	Food products	Fish and fishery products-Fish--Wild fish-Warm smoked	CC6	6
RL15000158	P	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC121	121
RL15000159	P	2011	Food products	Meat and meat products-Swine--Deli product - Other product-Cooked	CC1	1
RL15000160	P	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked processing not specified	CC121	121
RL15000161	C	2011	Food products	Meat and meat products-Gallus gallus (fowl)--Fresh meat - Cut-Raw	CC121	121
RL15000162	C	2011	Food products	Fish and fishery products-Fish-Clupea harengus (Herring, kipper)-Wild fish-Raw	CC7	7
RL15000163	C	2011	Food products	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-Raw	CC121	121
RL15000164	C	2011	Food products	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-Raw	CC9	9
RL15000165	C	2011	Food products	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-	CC121	121
RL15000166	C	2011	Food products	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-Raw	CC121	121
RL15000167	C	2011	Food products	Fish and fishery products-Fish-Oncorhynchus mykiss, Salmo trutta (Trout)-Fish origin not specified-Raw	CC121	121
RL15000168	C	2011	Food products	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-Raw	CC121	121
RL15000169	C	2011	Food products	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-Raw	CC121	121
RL15000170	C	2011	Food products	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-Raw	CC121	121
RL15000171	C	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked processing not specified	CC8	8
RL15000172	J	2010	Food products	Fish and fishery products-Fish--Wild fish-Cold smoked	CC193	193
RL15000173	J	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC121	121
RL15000174	J	2011	Food products	Fish and fishery products-Fish--Wild fish-Warm smoked	CC121	121
RL15000175	J	2011	Food products	Meat and meat products-Gallus gallus (fowl)--Deli product - Other product-Other stabilization	CC155	155
RL15000176	J	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC9	9
RL15000177	J	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Warm smoked	CC155	155
RL15000178	M	2011	Food products	Meat and meat products-Fish-Salmo spp. (Salmon)-Deli product - Sliced-Cooked	CC121	121
RL15000179	M	2011	Food products	Meat and meat products-Swine--Deli product - Sliced-Cooked	CC8	8
RL15000180	V	2011	Food products	Fish and fishery products-Fish--Farmed fish-Cold smoked	CC6	6
RL15000181	V	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Warm smoked	CC121	121
RL15000182	V	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC31	31
RL15000183	V	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC9	9

RL_number	Country	Year	Sample type	Description	MLST Clonal complexes	MLST sequence types
RL15000184	V	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC193	193
RL15000185	V	2010	Food products	Fish and fishery products-Fish--Farmed fish-Smoked processing not specified	CC193	193
RL15000186	A	2011	Food products	Fish and fishery products-Fish--Farmed fish-Gravad/slightly salted	CC121	121
RL15000187	A	2011	Food products	Fish and fishery products-Fish--Farmed fish-Smoked processing not specified	CC155	155
RL15000188	A	2011	Food products	Fish and fishery products-Fish--Farmed fish-Smoked processing not specified	CC121	121
RL15000189	A	2011	Food products	Milk and milk products-Bovine--Semi soft cheese-Made from pasteurized milk	CC31	325
RL15000190	W	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC121	121
RL15000191	W	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Gravad/slightly salted	CC155	155
RL15000192	W	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Warm smoked	CC8	8
RL15000193	W	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Warm smoked	CC8	8
RL15000194	W	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked processing not specified	CC8	8
RL15000195	W	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked processing not specified	CC403	403
RL15000196	W	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked processing not specified	CC2	145
RL15000197	L	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC121	121
RL15000198	L	2010	Food products	Meat and meat products-Mixed sources--Deli product - Sausage-Cooked	CC6	6
RL15000199	L	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC9	9
RL15000200	L	2010	Food products	Milk and milk products-Bovine--Semi soft cheese-Made from raw milk	CC1	1
RL15000201	L	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked processing not specified	CC121	121
RL15000202	L	2010	Food products	Meat and meat products-Swine--Deli products - Pate-Cooked	CC121	236
RL15000203	L	2010	Food products	Meat and meat products-Swine--Deli products - Pate-Cooked	CC121	236
RL15000204	L	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Gravad/slightly salted	CC9	9
RL15000205	L	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC9	9
RL15000206	L	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC121	121
RL15000207	L	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Gravad/slightly salted	CC9	9
RL15000208	L	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC121	121
RL15000209	L	2010	Food products	Meat and meat products-Swine--Deli product - Other product-Cooked	CC3	3
RL15000210	L	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC8	8
RL15000211	L	2010	Food products	Meat and meat products-Swine--Deli product - Sausage-Cooked	CC2	2
RL15000212	L	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked processing not specified	CC101	101
RL15000213	L	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Warm smoked	CC101	101
RL15000214	L	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC121	121

RL_number	Country	Year	Sample type	Description	MLST Clonal complexes	MLST sequence types
RL15000215	L	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Warm smoked	CC9	9
RL15000216	L	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC204	204
RL15000217	L	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Warm smoked	CC121	121
RL15000218	L	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC121	121
RL15000219	L	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked processing not specified	CC121	121
RL15000220	L	2010	Food products	Meat and meat products-Swine--Deli product - Sausage-Cooked	CC8	8
RL15000221	L	2010	Food products	Meat and meat products-Mixed sources--Deli product - Sausage-Cooked	CC2	2
RL15000222	L	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Gravad/slightly salted	CC9	9
RL15000223	L	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC121	121
RL15000224	L	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Warm smoked	CC121	121
RL15000225	L	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Warm smoked	CC121	121
RL15000226	L	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC121	121
RL15000227	L	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC9	9
RL15000228	L	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC19	173
RL15000229	L	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC121	121
RL15000230	L	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC121	121
RL15000231	L	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Warm smoked	CC121	121
RL15000232	L	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Warm smoked	CC19	173
RL15000233	L	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC121	121
RL15000234	L	2010	Food products	Meat and meat products-Swine--Deli product - Other product-Cooked	CC9	9
RL15000235	L	2010	Food products	Meat and meat products-Swine--Deli product - Sausage-Cooked	CC177	177
RL15000236	L	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked processing not specified	CC121	121
RL15000237	L	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC121	121
RL15000238	L	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Gravad/slightly salted	CC8	8
RL15000239	L	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked processing not specified	CC121	121
RL15000240	L	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC121	121
RL15000241	L	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Warm smoked	CC121	121
RL15000242	L	2011	Food products	Meat and meat products-Swine--Deli product - Sausage-Cooked	CC121	236
RL15000243	L	2011	Food products	Meat and meat products-Swine--Deli product - Sausage-Cooked	CC121	121
RL15000244	L	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC121	121
RL15000245	Y	2010	Food products	Fish and fishery products-Fish--Farmed fish-Gravad/slightly salted	CC8	120

RL_number	Country	Year	Sample type	Description	MLST Clonal complexes	MLST sequence types
RL15000246	Y	2010	Food products	Fish and fishery products-Fish--Farmed fish-Gravad/slightly salted	CC121	121
RL15000247	Q	2011	Food products	Fish and fishery products-Fish--Farmed fish-Smoked (process not specify)	CC59	59
RL15000248	Q	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked (process not specify)	CC59	59
RL15000249	Q	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked (process not specify)	CC87	87
RL15000250	Q	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-	CC7	7
RL15000251	Y	2010	Food products	Fish and fishery products-Fish--Farmed fish-Gravad/slightly salted	CC121	121
RL15000252	Q	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked (process not specify)	CC121	121
RL15000253	Q	2011	Food products	Fish and fishery products-Fish--Farmed fish-Smoked (process not specify)	CC121	121
RL15000254	Q	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked (process not specify)	CC59	59
RL15000255	Q	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked (process not specify)	CC59	59
RL15000256	Q	2011	Food products	Fish and fishery products-Fish--Farmed fish-Smoked (process not specify)	CC87	87
RL15000257	Q	2011	Food products	Fish and fishery products-Fish--Farmed fish-Smoked (process not specify)	CC31	31
RL15000258	Q	2011	Food products	Fish and fishery products-Fish--Farmed fish-Smoked (process not specify)	CC87	87
RL15000259	Q	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-	CC59	59
RL15000260	Q	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked (process not specify)	CC7	7
RL15000261	Q	2011	Food products	Fish and fishery products-Fish--Farmed fish-Smoked (process not specify)	CC87	87
RL15000262	Q	2011	Food products	Fish and fishery products-Fish--Farmed fish-Smoked (process not specify)	CC121	121
RL15000263	Q	2011	Food products	Fish and fishery products-Fish--Farmed fish-Smoked (process not specify)	CC155	155
RL15000264	Q	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked (process not specify)	CC87	87
RL15000265	Q	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked (process not specify)	CC87	87
RL15000266	Q	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked (process not specify)	CC20	20
RL15000267	Q	2011	Food products	Fish and fishery products-Fish--Farmed fish-Smoked (process not specify)	CC8	8
RL15000268	Y	2010	Food products	Fish and fishery products-Fish--Farmed fish-Warm smoked	CC31	31
RL15000269	Y	2010	Food products	Fish and fishery products-Fish--Farmed fish-Gravad/slightly salted	CC121	121
RL15000270	Y	2010	Food products	Fish and fishery products-Fish--Farmed fish-Gravad/slightly salted	CC8	8
RL15000271	Q	2011	Food products	Milk and milk products-Unspecified--Semi soft cheese-	CC31	325
RL15000272	Q	2011	Food products	Milk and milk products-Unspecified--Semi soft cheese-	CC31	325
RL15000273	Q	2011	Food products	Fish and fishery products-Fish--Farmed fish-Smoked (process not specify)	CC7	732
RL15000274	Q	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-	CC8	8
RL15000275	Q	2011	Food products	Fish and fishery products-Fish--Farmed fish-Smoked (process not specify)	CC7	7
RL15000276	Q	2011	Food products	Fish and fishery products-Fish--Farmed fish-Smoked (process not specify)	CC121	121

RL_number	Country	Year	Sample type	Description	MLST Clonal complexes	MLST sequence types
RL15000277	Y	2010	Food products	Fish and fishery products-Fish--Farmed fish-Gravad/slightly salted	CC121	121
RL15000278	Q	2011	Food products	Fish and fishery products-Fish--Wild fish-Cold smoked	CC9	9
RL15000279	Q	2011	Food products	Fish and fishery products-Fish--Farmed fish-Smoked (process not specify)	CC7	7
RL15000280	Q	2011	Food products	Fish and fishery products-Fish--Farmed fish-Cold smoked	CC6	6
RL15000281	Q	2011	Food products	Fish and fishery products-Fish--Farmed fish-Smoked (process not specify)	CC8	8
RL15000282	Q	2011	Food products	Fish and fishery products-Fish--Farmed fish-Warm smoked	CC121	121
RL15000284	Q	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-	CC121	121
RL15000286	T	2010	Food products	Fish and fishery products-Fish--Fish and fishery products-Cold smoked	CC3	44
RL15000287	T	2010	Food products	Fish and fishery products-Fish--Fish and fishery products-Cold smoked	CC121	121
RL15000288	T	2010	Food products	Fish and fishery products-Fish--Fish and fishery products-Cold smoked	CC8	8
RL15000289	T	2010	Food products	Fish and fishery products-Fish--Fish and fishery products-Cold smoked	CC121	121
RL15000290	Y	2010	Food products	Fish and fishery products-Fish--Farmed fish-Gravad/slightly salted	CC8	8
RL15000291	K	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC8	8
RL15000292	K	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC8	8
RL15000293	K	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC8	8
RL15000294	K	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC8	8
RL15000295	K	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Gravad/slightly salted	CC155	155
RL15000296	K	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Warm smoked	CC121	121
RL15000297	K	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Gravad/slightly salted	CC155	155
RL15000298	K	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Gravad/slightly salted	CC8	8
RL15000299	K	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Gravad/slightly salted	CC8	8
RL15000300	K	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC8	8
RL15000301	K	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Gravad/slightly salted	CC121	121
RL15000302	L	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked processing not specified	CC88	296
RL15000303	C	2011	Food products	Fish and fishery products-Fish-Clupea harengus (Herring, kipper)-Wild fish-Raw	CC204	204
RL15000304	C	2011	Food products	Fish and fishery products-Fish-Clupea harengus (Herring, kipper)-Wild fish-Raw	CC7	624
RL15000305	Q	2011	Food products	Meat and meat products-Swine--Deli product - Other meat products-	CC8	8
RL15000329	L	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	-	*
RL15000330	L	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC8	8
RL15000331	L	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Warm smoked	CC59	59
RL15000332	L	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-Warm smoked	CC8	8

RL_number	Country	Year	Sample type	Description	MLST Clonal complexes	MLST sequence types
RL15000333	L	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Warm smoked	CC121	121
RL15000334	K	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-	CC8	8
RL15000335	K	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-	CC8	8
RL15000336	K	2010	Food products	Fish and fishery products-Fish--Fish origin not specified-	CC121	121
RL15000337	A	2010	Food products	Fish and fishery products-Fish--Farmed fish-Gravad/slightly salted	CC8	8
RL15000338	J	2011	Food products	Fish and fishery products-Fish--Wild fish-Cold smoked	CC9	9
RL15000339	J	2011	Food products	Fish and fishery products-Fish--Wild fish-Cold smoked	CC101	101
RL15000340	J	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC155	155
RL15000341	J	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Cold smoked	CC8	8
RL15000658	X	2010	Food products	Milk and milk products-Goat--Soft cheese-Made from pasteurized milk	CC7	7
RL15000659	X	2010	Food products	Meat and meat products-Turkeys--Deli product - Sliced-Cooked	CC9	9
RL15000660	X	2010	Food products	Meat and meat products-Swine--Deli product - Ham-Smoked processing not specified	CC121	121
RL15000661	X	2010	Food products	Meat and meat products-Swine--Deli product - Ham-Smoked processing not specified	CC121	121
RL15000662	X	2010	Food products	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-Smoked processing not specified	CC9	9
RL15000663	X	2010	Food products	Fish and fishery products-Fish-Oncorhynchus mykiss, Salmo trutta (Trout)-Fish origin not specified-Smoked processing not specified	CC8	16
RL15000664	X	2011	Food products	Meat and meat products-Gallus gallus (fowl)--Deli product - Other product-	CC121	121
RL15000665	X	2011	Food products	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-Smoked processing not specified	CC6	6
RL15000666	X	2011	Food products	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-Smoked processing not specified	CC8	8
RL15000667	X	2011	Food products	Meat and meat products-Swine--Deli product - Sliced-Cooked	CC121	121
RL15000668	X	2011	Food products	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-Smoked processing not specified	CC121	121
RL15000669	X	2011	Food products	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-Smoked processing not specified	CC9	9
RL15000670	X	2011	Food products	Meat and meat products-Gallus gallus (fowl)--Deli product - Sliced-Cooked	CC2	2
RL15000671	X	2011	Food products	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-Smoked processing not specified	CC121	121
RL15000672	X	2011	Food products	Meat and meat products-Gallus gallus (fowl)--Deli product - Sliced-Cooked	CC2	2
RL15000673	X	2011	Food products	Meat and meat products-Mixed sources--Deli product - Sliced-Cooked	CC2	2
RL15000674	X	2011	Food products	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-Smoked processing not specified	CC121	121
RL15000675	X	2011	Food products	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-Smoked	CC121	121

RL_number	Country	Year	Sample type	Description	MLST Clonal complexes	MLST sequence types
				processing not specified		
RL15000676	X	2011	Food products	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-Smoked processing not specified	CC9	9
RL15000677	X	2011	Food products	Meat and meat products-Swine--Deli product - Sliced-Cooked	CC9	9
RL15000678	X	2011	Food products	Meat and meat products-Swine--Deli product - Sliced-Cooked	CC9	9
RL15000679	X	2011	Food products	Meat and meat products-Gallus gallus (fowl)--Deli product - Sliced-Cooked	CC3	3
RL15000680	X	2011	Food products	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-Smoked processing not specified	CC21	21
RL15000681	X	2011	Food products	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-Smoked processing not specified	CC9	9
RL15000682	X	2011	Food products	Meat and meat products-Geese--Deli product - Sliced-Cooked	CC121	121
RL15000683	X	2011	Food products	Fish and fishery products-Fish-Oncorhynchus mykiss, Salmo trutta (Trout)-Fish origin not specified-Smoked processing not specified	CC177	177
RL15000684	X	2011	Food products	Fish and fishery products-Fish-Oncorhynchus mykiss, Salmo trutta (Trout)-Fish origin not specified-Smoked processing not specified	CC6	6
RL15000685	X	2011	Food products	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-Smoked processing not specified	CC121	121
RL15000686	X	2011	Food products	Meat and meat products-Swine--Deli product - Sliced-Cooked	CC3	3
RL15000687	X	2011	Food products	Meat and meat products-Swine--Deli product - Sliced-Cooked	CC121	121
RL15000688	X	2011	Food products	Meat and meat products-Swine--Deli products - Pate-Cooked	CC6	6
RL15000689	X	2011	Food products	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-Smoked processing not specified	CC121	121
RL15000690	X	2011	Food products	Fish and fishery products-Fish-Oncorhynchus mykiss, Salmo trutta (Trout)-Fish origin not specified-Smoked processing not specified	CC155	155
RL15000691	X	2011	Food products	Fish and fishery products-Fish-Oncorhynchus mykiss, Salmo trutta (Trout)-Fish origin not specified-Smoked processing not specified	CC31	325
RL15000692	X	2011	Food products	Fish and fishery products-Fish--Fish origin not specified-Smoked processing not specified	CC121	121
RL15000693	X	2011	Food products	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-Gravad/slightly salted	CC121	121
RL15000694	X	2011	Food products	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-Smoked processing not specified	CC2	2
RL15000695	X	2011	Food products	Meat and meat products-Swine--Deli product - Sliced-Cooked	CC8	8
RL15000730	H	2010	Food products	Milk and milk products-Unspecified--Cheese category not specified-	CC204	204
RL15000731	H	2010	Food products	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-Smoked processing not specified	CC121	121
RL15000732	H	2010	Food products	Milk and milk products-Unspecified--Cheese category not specified-	CC204	204
RL15000733	H	2010	Food products	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-Smoked processing not specified	CC155	155

RL_number	Country	Year	Sample type	Description	MLST Clonal complexes	MLST sequence types
RL15000734	H	2010	Food products	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-Smoked processing not specified	CC155	155

*NOVEL allele. Cannot determine closest ST (SLV).

Appendix 2: Isolates other food, ready-to-eat meat and cheese

RL_number	Country	Year	Sample type	Description	MLST Clonal complexes	MLST sequence types
RL15000342	V	2010	Food products	Meat and meat products-Unspecified---	CC5	5
RL15000343	V	2010	Food products	Meat and meat products-Unspecified---	CC5	5
RL15000344	Z	2014	Food products	Meat and meat products-Bovine--Meat - Sausage-	CC77	77
RL15000345	Z	2014	Food products	Meat and meat products-Bovine--Meat - Sausage-	CC9	9
RL15000346	G	2011	Food products	Meat and meat products-Swine--Deli product - Ham-	CC31	31
RL15000347	G	2011	Food products	Meat and meat products-Swine--Deli product - Ham-	CC121	121
RL15000348	G	2011	Food products	Meat and meat products-Swine--Deli product - Ham-	CC204	204
RL15000349	Z	2014	Food products	Meat and meat products-Bovine--Meat - Minced-	CC87	87
RL15000350	G	2012	Food products	Meat and meat products-Swine--Deli product - Ham-	CC31	31
RL15000351	Z	2014	Food products	Elaborated food products combining several food categories-Mixed sources--Sandwich-	CC7	7
RL15000352	Z	2014	Food products	Meat and meat products-Bovine--Meat - Minced-	CC20	20
RL15000353	Z	2014	Food products	Meat and meat products-Bovine--Meat - Minced-	CC29	29
RL15000354	Z	2013	Food products	Meat and meat products-Bovine--Meat - Sausage-	CC9	9
RL15000355	Z	2013	Food products	Meat and meat products-Swine--Deli product - Ham-Smoked processing not specified	CC101	101
RL15000356	Z	2013	Food products	Elaborated food products combining several food categories-Mixed sources--Sandwich-	CC9	9
RL15000357	Z	2013	Food products	Meat and meat products-Gallus gallus (fowl)--Meat - Cut-Smoked processing not specified	CC9	9
RL15000358	Z	2013	Food products	Meat and meat products-Unspecified--Deli product - Sausage-	CC11	451
RL15000359	Z	2013	Food products	Meat and meat products-Bovine--Meat - Minced-	CC8	8
RL15000360	Z	2012	Food products	Meat and meat products-Gallus gallus (fowl)--Meat - Cut-	CC121	121
RL15000394	C	2010	Food products	Milk and milk products-Bovine--Semi soft cheese-	CC155	155
RL15000395	C	2010	Food products	Milk and milk products-Bovine--Hard cheese-	CC77	77
RL15000396	C	2010	Food products	Milk and milk products-Unspecified--Cheese category not specified-	CC207	207
RL15000397	C	2010	Food products	Milk and milk products-Bovine--Hard cheese-	CC21	21
RL15000398	C	2010	Food products	Milk and milk products-Bovine--Cream-	CC101	101
RL15000399	C	2010	Food products	Milk and milk products-Bovine--Soft cheese-	CC54	54
RL15000400	C	2010	Food products	Milk and milk products-Bovine--Cream-	CC101	101
RL15000401	C	2010	Food products	Milk and milk products-Bovine--Cream-	CC59	59
RL15000402	C	2010	Food products	Milk and milk products-Goat--Fresh cheese-	CC2	2

RL_number	Country	Year	Sample type	Description	MLST Clonal complexes	MLST sequence types
RL15000403	C	2010	Food products	Milk and milk products-Bovine--Cheese category not specified-	CC217	217
RL15000404	C	2010	Food products	Milk and milk products-Bovine--Cheese category not specified-	CC5	5
RL15000405	C	2010	Food products	Milk and milk products-Bovine--Hard cheese-	CC6	6
RL15000406	C	2010	Food products	Milk and milk products-Bovine--Semi soft cheese-	CC6	6
RL15000407	C	2011	Food products	Milk and milk products-Bovine--Cheese category not specified-	CC6	6
RL15000408	P	2011	Food products	Milk and milk products-Bovine--Cheese category not specified-	CC9	9
RL15000409	P	2011	Food products	Milk and milk products-Bovine--Cream-	CC54	54
RL15000410	P	2011	Food products	Milk and milk products-Bovine--Butter-	CC3	174
RL15000411	P	2011	Food products	Milk and milk products-Bovine--Cheese category not specified-	CC121	121
RL15000412	C	2011	Food products	Milk and milk products-Bovine--Cheese category not specified-	CC21	21
RL15000413	C	2011	Food products	Milk and milk products-Unspecified--Melted cheese-	CC121	121
RL15000414	C	2011	Food products	Milk and milk products-Bovine--Soft cheese-	CC20	20
RL15000415	C	2011	Food products	Milk and milk products-Unspecified--Cheese category not specified-	CC6	6
RL15000416	C	2011	Food products	Milk and milk products-Bovine--Hard cheese-	CC1	1
RL15000417	C	2012	Food products	Milk and milk products-Bovine--Semi soft cheese-	CC21	21
RL15000418	C	2012	Food products	Milk and milk products-Bovine--Hard cheese-	CC37	37
RL15000419	C	2012	Food products	Milk and milk products-Bovine--Cheese category not specified-	CC31	325
RL15000420	C	2012	Food products	Milk and milk products-Bovine--Soft cheese-	CC11	451
RL15000421	C	2012	Food products	Milk and milk products-Bovine--Hard cheese-	CC6	175
RL15000422	C	2012	Food products	Milk and milk products-Bovine--Semi soft cheese-	CC21	21
RL15000423	C	2012	Food products	Milk and milk products-Bovine--Semi soft cheese-	CC8	8
RL15000424	C	2012	Food products	Milk and milk products-Bovine--Semi soft cheese-	CC8	8
RL15000425	C	2012	Food products	Milk and milk products-Bovine--Semi soft cheese-	CC8	8
RL15000426	C	2012	Food products	Milk and milk products-Bovine--Semi soft cheese-	CC8	8
RL15000427	C	2012	Food products	Milk and milk products-Bovine--Semi soft cheese-	CC155	155
RL15000428	C	2012	Food products	Milk and milk products-Bovine--Semi soft cheese-	CC8	8
RL15000429	C	2012	Food products	Milk and milk products-Bovine--Semi soft cheese-	CC8	8
RL15000430	C	2012	Food products	Milk and milk products-Bovine--Semi soft cheese-	CC6	6
RL15000431	C	2012	Food products	Milk and milk products-Bovine--Semi soft cheese-	CC6	6
RL15000432	C	2012	Food products	Milk and milk products-Bovine--Semi soft cheese-	CC6	6
RL15000433	C	2012	Food products	Milk and milk products-Bovine--Cheese category not specified-	CC6	6
RL15000434	C	2012	Food products	Milk and milk products-Bovine--Semi soft cheese-	CC6	6

RL_number	Country	Year	Sample type	Description	MLST Clonal complexes	MLST sequence types
RL15000435	C	2012	Food products	Milk and milk products-Bovine--Semi soft cheese-	CC6	6
RL15000436	A	2012	Food products	Milk and milk products-Bovine--Soft cheese-	CC8	8
RL15000437	V	2010	Food products	Milk and milk products-Unspecified--Cheese category not specified-	CC14	14
RL15000438	V	2010	Food products	Meat and meat products-Unspecified--Meat - Cut-Smoked processing not specified	CC37	37
RL15000439	V	2010	Food products	Milk and milk products-Sheep--Cheese category not specified-	CC21	21
RL15000440	V	2010	Food products	Milk and milk products-Sheep--Cheese category not specified-	CC14	91
RL15000441	V	2010	Food products	Milk and milk products-Sheep--Cheese category not specified-	CC14	91
RL15000442	V	2010	Food products	Milk and milk products-Sheep--Cheese category not specified-	CC14	91
RL15000443	V	2010	Food products	Milk and milk products-Sheep--Cheese category not specified-	CC21	21
RL15000444	V	2010	Food products	Milk and milk products-Sheep--Cheese category not specified-	CC21	21
RL15000445	V	2010	Food products	Milk and milk products-Unspecified--Cheese category not specified-	CC7	511
RL15000446	V	2010	Food products	Milk and milk products-Sheep--Cheese category not specified-	CC21	21
RL15000447	V	2010	Food products	Milk and milk products-Sheep--Cheese category not specified-	CC14	91
RL15000448	V	2010	Food products	Milk and milk products-Sheep--Cheese category not specified-	CC21	21
RL15000449	V	2011	Food products	Milk and milk products-Unspecified--Cheese category not specified-	CC14	14
RL15000450	V	2011	Food products	Milk and milk products-Sheep--Cheese category not specified-	CC7	7
RL15000451	V	2011	Food products	Milk and milk products-Sheep--Cheese category not specified-	CC37	37
RL15000452	V	2011	Food products	Milk and milk products-Unspecified--Hard cheese-	CC21	21
RL15000453	V	2011	Food products	Milk and milk products-Sheep--Cheese category not specified-	CC31	325
RL15000454	V	2012	Food products	Milk and milk products-Unspecified--Cheese category not specified-	CC1	1
RL15000455	V	2012	Food products	Milk and milk products-Sheep--Cheese category not specified-	CC20	20
RL15000456	V	2012	Food products	Milk and milk products-Sheep--Cheese category not specified-	CC121	121
RL15000457	V	2012	Food products	Milk and milk products-Unspecified--Cheese category not specified-	CC1	1
RL15000458	V	2012	Food products	Milk and milk products-Sheep--Cheese category not specified-	CC37	37
RL15000459	V	2012	Food products	Milk and milk products-Sheep--Cheese category not specified-	ST200	200
RL15000460	V	2012	Food products	Milk and milk products-Unspecified--Cheese category not specified-	CC1	1
RL15000461	C	2010	Food products	Meat and meat products-Swine--Deli product - Sausage-	CC9	9
RL15000462	C	2011	Food products	Meat and meat products-Unspecified--Deli product - Sausage-	CC121	121
RL15000463	C	2011	Food products	Meat and meat products-Swine--Deli product - Other product-	CC9	9
RL15000464	C	2011	Food products	Meat and meat products-Swine--Deli product - Sausage-	CC9	9
RL15000465	C	2011	Food products	Meat and meat products-Swine--Deli product - Other product-	CC8	8

RL_number	Country	Year	Sample type	Description	MLST Clonal complexes	MLST sequence types
RL15000466	C	2011	Food products	Meat and meat products-Poultry not specified--Deli product - Other product-	CC121	121
RL15000467	C	2011	Food products	Meat and meat products-Ducks--Deli product - Other product-	CC204	204
RL15000468	C	2011	Food products	Meat and meat products-Ducks--Deli product - Other product-	CC204	204
RL15000469	C	2011	Food products	Meat and meat products-Swine--Deli product - Other product-	CC121	121
RL15000470	C	2011	Food products	Meat and meat products-Swine--Deli product - Other product-	CC121	121
RL15000471	C	2011	Food products	Meat and meat products-Swine--Deli product - Other product-	CC121	121
RL15000472	C	2011	Food products	Meat and meat products-Swine--Deli product - Other product-	CC121	121
RL15000473	C	2011	Food products	Meat and meat products-Unspecified--Deli product - Sausage-	CC6	6
RL15000474	C	2011	Food products	Meat and meat products-Swine--Deli product - Other product-	CC31	31
RL15000475	C	2011	Food products	Meat and meat products-Swine--Deli product - Other product-	CC31	31
RL15000476	C	2011	Food products	Meat and meat products-Swine--Deli product - Other product-	CC31	31
RL15000477	C	2011	Food products	Meat and meat products-Swine--Deli product - Other product-	CC31	31
RL15000478	C	2011	Food products	Meat and meat products-Swine--Deli product - Other product-	CC31	31
RL15000479	C	2011	Food products	Meat and meat products-Unspecified--Deli product - Sausage-	CC9	9
RL15000480	C	2011	Food products	Meat and meat products-Unspecified--Deli product - Sausage-	CC9	9
RL15000481	C	2011	Food products	Meat and meat products-Ducks--Deli product - Other product-	CC204	204
RL15000482	C	2011	Food products	Meat and meat products-Mixed sources--Deli product - Sausage-	CC6	6
RL15000483	C	2011	Food products	Meat and meat products-Swine--Deli product - Other product-	CC31	31
RL15000484	C	2011	Food products	Meat and meat products-Swine--Deli product - Sausage-	CC121	121
RL15000485	C	2011	Food products	Meat and meat products-Gallus gallus (fowl)--Deli products - Pate-	CC121	121
RL15000486	C	2011	Food products	Meat and meat products-Swine--Deli product - Other product-Sliced	CC2	2
RL15000487	C	2011	Food products	Meat and meat products-Unspecified--Deli product - Sausage-	CC9	9
RL15000488	C	2011	Food products	Meat and meat products-Unspecified--Deli products - Pate-	CC121	121
RL15000489	C	2012	Food products	Meat and meat products-Unspecified--Deli product - Sausage-	CC37	37
RL15000490	C	2012	Food products	Meat and meat products-Unspecified--Deli product - Sausage-	CC2	2
RL15000491	C	2012	Food products	Meat and meat products-Unspecified--Meat - Minced-Raw	CC9	9
RL15000492	C	2012	Food products	Meat and meat products-Swine--Deli products - Pate-	CC5	5
RL15000493	C	2012	Food products	Meat and meat products-Swine--Deli product - Other product-	CC31	31
RL15000494	C	2012	Food products	Meat and meat products-Unspecified--Deli product - Sausage-	CC121	121
RL15000495	C	2012	Food products	Meat and meat products-Swine--Deli product - Sausage-	CC121	121
RL15000496	C	2012	Food products	Meat and meat products-Unspecified--Deli products - Pate-	CC121	121

RL_number	Country	Year	Sample type	Description	MLST Clonal complexes	MLST sequence types
RL15000497	C	2012	Food products	Meat and meat products-Swine--Deli products - Pate-	CC2	2
RL15000498	C	2012	Food products	Meat and meat products-Unspecified--Deli products - Pate-	CC121	121
RL15000499	C	2012	Food products	Meat and meat products-Unspecified--Deli product - Sausage-	CC121	176
RL15000500	C	2012	Food products	Meat and meat products-Swine--Deli product - Other product-	CC3	3
RL15000501	C	2012	Food products	Meat and meat products-Unspecified--Deli product - Sausage-	CC121	121
RL15000502	C	2012	Food products	Meat and meat products-Unspecified--Deli product - Sausage-	CC59	59
RL15000503	C	2012	Food products	Meat and meat products-Swine--Deli product - Other product-	CC9	9
RL15000504	C	2012	Food products	Meat and meat products-Swine--Deli product - Sausage-	CC6	6
RL15000505	C	2012	Food products	Meat and meat products-Unspecified--Deli product - Sausage-	CC37	37
RL15000506	A	2010	Food products	Meat and meat products-Gallus gallus (fowl)--Meat - Cut-	CC8	16
RL15000507	A	2010	Food products	Meat and meat products-Swine--Deli product - Ham-Sliced	CC2	2
RL15000508	A	2010	Food products	Meat and meat products-Unspecified--Meat - Minced-Cooked	CC9	9
RL15000509	A	2010	Food products	Meat and meat products-Swine--Deli product - Ham-Smoked processing not specified	CC9	9
RL15000510	A	2010	Food products	Meat and meat products-Swine--Meat - Minced-Smoked processing not specified	CC9	9
RL15000511	A	2010	Food products	Meat and meat products-Swine--Meat - Other product-	CC14	91
RL15000512	A	2010	Food products	Meat and meat products-Swine--Meat - Other product-	CC21	21
RL15000513	A	2010	Food products	Meat and meat products-Sheep--Meat - Cut-Smoked processing not specified	CC9	9
RL15000514	A	2011	Food products	Elaborated food products combining several food categories-Mixed sources--Soups-	CC9	9
RL15000515	A	2011	Food products	Meat and meat products-Swine--Meat - Other product-Cooked	CC2	2
RL15000516	A	2011	Food products	Meat and meat products-Unspecified--Meat - Cut-Cooked	CC121	121
RL15000517	A	2011	Food products	Meat and meat products-Unspecified--Meat - Cut-Cooked	CC121	121
RL15000518	A	2011	Food products	Meat and meat products-Unspecified--Meat - Cut-Cooked	CC11	451
RL15000519	A	2011	Food products	Meat and meat products-Unspecified--Meat - Cut-Cooked	CC11	451
RL15000520	A	2011	Food products	Meat and meat products-Unspecified--Meat - Cut-	CC6	6
RL15000521	A	2011	Food products	Meat and meat products-Unspecified--Meat - Cut-Sliced	CC31	31
RL15000522	A	2011	Food products	Meat and meat products-Unspecified--Meat - Cut-	CC9	9
RL15000523	A	2011	Food products	Meat and meat products-Unspecified--Meat - Cut-Sliced	CC9	9
RL15000524	A	2012	Food products	Meat and meat products-Gallus gallus (fowl)--Meat - Cut-	CC9	9
RL15000525	A	2012	Food products	Meat and meat products-Gallus gallus (fowl)--Meat - Cut-	CC9	9
RL15000526	A	2012	Food products	Meat and meat products-Unspecified--Meat - Cut-Sliced	CC121	121

RL_number	Country	Year	Sample type	Description	MLST Clonal complexes	MLST sequence types
RL15000527	A	2012	Food products	Meat and meat products-Swine--Meat - Cut-Cooked	CC8	8
RL15000528	A	2012	Food products	Meat and meat products-Unspecified--Deli product - Other product-Smoked processing not specified	CC9	9
RL15000529	A	2012	Food products	Meat and meat products-Unspecified--Deli product - Other product-Smoked processing not specified	CC9	9
RL15000530	A	2012	Food products	Meat and meat products-Swine--Deli product - Ham-	CC9	9
RL15000531	A	2011	Food products	Meat and meat products-Bovine--Meat - Cut-Cooked	CC9	9
RL15000532	A	2011	Food products	Meat and meat products-Unspecified--Meat - Cut-Sliced	CC121	121
RL15000533	A	2011	Food products	Meat and meat products-Bovine--Meat - Cut-	CC9	9
RL15000534	V	2010	Food products	Meat and meat products-Unspecified---	CC87	87
RL15000535	V	2010	Food products	Meat and meat products-Unspecified---	CC7	*7
RL15000696	X	2010	Food products	Milk and milk products-Unspecified--Hard cheese-	CC1	328
RL15000697	X	2010	Food products	Milk and milk products-Unspecified--Cheese category not specified-	CC1	1
RL15000698	X	2010	Food products	Milk and milk products-Unspecified--Cheese category not specified-	CC1	1
RL15000699	X	2011	Food products	Milk and milk products-Unspecified--Hard cheese-	ST382	183
RL15000700	X	2011	Food products	Milk and milk products-Unspecified--Hard cheese-	CC3	3
RL15000701	X	2010	Food products	Meat and meat products-Swine--Meat - Cut-	CC9	9
RL15000702	X	2010	Food products	Meat and meat products-Swine--Deli product - Sausage-	CC31	31
RL15000703	X	2010	Food products	Meat and meat products-Swine--Deli product - Sausage-	CC31	31
RL15000704	X	2010	Food products	Meat and meat products-Swine--Meat - Sausage-	CC4	4
RL15000705	X	2010	Food products	Meat and meat products-Gallus gallus (fowl)--Meat - Cut-Cooked	CC121	121
RL15000706	X	2010	Food products	Meat and meat products-Swine--Deli product - Sausage-	CC9	9
RL15000707	X	2010	Food products	Elaborated food products combining several food categories-Mixed sources--Sandwich-	CC121	121
RL15000708	X	2010	Food products	Meat and meat products-Swine--Deli product - Sausage-	CC2	2
RL15000709	X	2010	Food products	Elaborated food products combining several food categories-Mixed sources--Ready made meal-	CC121	121
RL15000710	X	2010	Food products	Meat and meat products-Ducks--Deli product - Ham-	CC5	5
RL15000711	X	2010	Food products	Meat and meat products-Swine--Meat - Sausage-	CC26	26
RL15000712	X	2010	Food products	Meat and meat products-Swine--Deli product - Sausage-	CC121	121
RL15000713	X	2010	Food products	Meat and meat products-Swine--Deli product - Sausage-	CC121	121
RL15000714	X	2010	Food products	Meat and meat products-Swine--Deli product - Sausage-	CC87	87
RL15000715	X	2010	Food products	Meat and meat products-Swine--Meat - Sausage-	CC9	9
RL15000716	X	2010	Food products	Meat and meat products-Swine--Deli product - Sausage-	CC8	8

RL_number	Country	Year	Sample type	Description	MLST Clonal complexes	MLST sequence types
RL15000717	X	2010	Food products	Meat and meat products-Swine--Deli product - Sausage-	CC9	9
RL15000718	X	2010	Food products	Elaborated food products combining several food categories-Mixed sources--Ready made meal-	CC121	121
RL15000719	X	2010	Food products	Meat and meat products-Bovine--Meat - Cut-	CC3	3
RL15000720	X	2011	Food products	Meat and meat products-Unspecified--Meat - Minced-	CC121	121
RL15000721	X	2011	Food products	Meat and meat products-Unspecified--Meat - Cut-Cooked	CC2	2
RL15000722	X	2011	Food products	Meat and meat products-Swine--Deli product - Sausage-	CC9	9
RL15000723	X	2011	Food products	Meat and meat products-Swine--Deli product - Sausage-	CC9	9
RL15000724	X	2011	Food products	Elaborated food products combining several food categories-Mixed sources--Ready to eat salad-	CC121	121
RL15000725	X	2011	Food products	Elaborated food products combining several food categories-Mixed sources--Ready to eat salad-	CC155	155
RL15000726	X	2011	Food products	Elaborated food products combining several food categories-Mixed sources--Ready to eat salad-	CC155	155
RL15000727	X	2011	Food products	Elaborated food products combining several food categories-Mixed sources--Ready to eat salad-	CC155	155
RL15000728	X	2011	Food products	Elaborated food products combining several food categories-Mixed sources--Ready to eat salad-	CC155	155
RL15000729	X	2011	Food products	Elaborated food products combining several food categories-Mixed sources--Ready to eat salad-	CC155	155
RL15001282	B	2010	Food products	Milk and milk products-Bovine--Cream-	CC8	8
RL15001292	B	2010	Food products	Milk and milk products-Unspecified--Cheese category not specified-	CC37	37
RL15001293	B	2010	Food products	Meat and meat products-Swine--Meat - Cut-	CC59	59
RL15001294	B	2010	Food products	Meat and meat products-Swine--Meat - Cut-Cooked	CC21	21
RL15001295	B	2010	Food products	Meat and meat products-Unspecified--Meat - Cut-	CC3	3
RL15001299	B	2010	Food products	Meat and meat products-Bovine--Meat - Cut-	CC121	121
RL15001300	B	2010	Food products	Meat and meat products-Poultry not specified--Deli product - Other product-	CC121	121
RL15001301	B	2011	Food products	Meat and meat products-Swine--Deli product - Ham-	CC121	121
RL15001311	B	2011	Food products	Milk and milk products-Unspecified--Ice cream-	CC218	218
RL15001312	B	2011	Food products	Milk and milk products-Bovine--Semi soft cheese-	CC224	224
RL15001332	B	2011	Food products	Milk and milk products-Bovine--Cream-	CC8	8
RL15001333	B	2011	Food products	Milk and milk products-Bovine--Hard cheese-	CC29	427
RL15001334	B	2011	Food products	Meat and meat products-Gallus gallus (fowl)--Meat - Cut-	CC9	9
RL15001335	B	2011	Food products	Meat and meat products-Swine--Deli product - Sausage-	CC8	16
RL15001346	B	2011	Food products	Milk and milk products-Unspecified--Milk-	CC37	37

RL_number	Country	Year	Sample type	Description	MLST Clonal complexes	MLST sequence types
RL15001347	B	2011	Food products	Meat and meat products-Bovine--Meat - Cut-Sliced	CC121	121
RL15001348	B	2011	Food products	Meat and meat products-Swine--Deli product - Ham-	CC6	6
RL15001349	B	2011	Food products	Meat and meat products-Unspecified--Meat - Cut-	CC9	9
RL15001352	B	2011	Food products	Meat and meat products-Bovine--Deli product - Other product-	CC6	6
RL15001353	B	2011	Food products	Meat and meat products-Bovine--Meat - Cut-Sliced	CC204	204
RL15001383	B	2010	Food products	Milk and milk products-Goat--Soft cheese-	CC59	59
RL15001384	B	2010	Food products	Milk and milk products-Bovine--Semi soft cheese-	CC37	37
RL15001385	B	2010	Food products	Meat and meat products-Swine--Deli product - Ham-	CC9	9

Appendix 3: Isolates from other food, fruits and vegetables

RL_number	Country	Year	Sample type	Description	MLST Clonal complexes	MLST sequence types
RL15001981	B	2010	Food products	Fruit, vegetables, cereals and herbs-Vegetal-Vegetable-	CC121	121
RL15001982	B	2010	Food products	Fruit, vegetables, cereals and herbs-Vegetal-Fruit-	CC2	2
RL15001984	B	2011	Food products	Fruit, vegetables, cereals and herbs-Vegetal-Vegetable-	CC6	6
RL15001985	B	2011	Food products	Fruit, vegetables, cereals and herbs-Vegetal-Vegetable-	ST839	839
RL15001986	B	2011	Food products	Fruit, vegetables, cereals and herbs-Vegetal-Fruit-	CC31	325

Appendix 4: Isolates from the food production chain

RL_number	Country	Year	Sample type	Description	MLST Clonal complexes	MLST sequence types
RL10000011	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC9	9
RL10000012	Q		Food products	Milk and milk products-Sheep--Fresh cheese-	CC18	18
RL10000013	Q		Food products	Milk and milk products-Sheep--Fresh cheese-	CC18	18
RL10000014	Q		Food products	Milk and milk products-Sheep--Fresh cheese-	CC2	171
RL10000015	Q		Food products	Milk and milk products-Sheep--Fresh cheese-	CC18	18
RL10000016	Q		Food products	Milk and milk products-Sheep--Fresh cheese-	CC193	193
RL10000018	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC101	101
RL10000019	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC121	121
RL1000002	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC18	18
RL10000020	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC121	121
RL10000021	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC121	121
RL10000022	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC121	121
RL10000023	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC121	121
RL10000024	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC121	121
RL10000025	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC121	121
RL10000026	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC101	101
RL10000027	Q		Food products	Milk and milk products-Sheep--Fresh cheese-	CC121	121
RL10000028	Q		Food products	Milk and milk products-Sheep--Fresh cheese-	CC121	121
RL10000029	Q		Food products	Milk and milk products-Sheep--Fresh cheese-	CC121	121
RL1000003	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC2	2
RL10000030	Q		Food products	Milk and milk products-Sheep--Fresh cheese-	CC121	121
RL10000031	Q		Food products	Milk and milk products-Sheep--Fresh cheese-	CC121	121
RL10000032	Q		Food products	Milk and milk products-Sheep--Fresh cheese-	CC121	121
RL10000033	Q		Food products	Milk and milk products-Sheep--Fresh cheese-	CC121	121
RL10000034	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC1	1
RL10000035	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC1	1
RL10000036	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC101	101
RL10000037	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC101	101
RL10000038	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC101	101

RL_number	Country	Year	Sample type	Description	MLST Clonal complexes	MLST sequence types
RL1000039	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC101	101
RL1000004	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC2	2
RL10000040	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC101	101
RL10000041	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC101	101
RL10000042	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC101	101
RL10000043	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC101	101
RL10000044	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC101	101
RL10000045	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC101	101
RL10000046	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC101	101
RL10000047	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC101	101
RL10000048	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC5	5
RL10000049	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC5	5
RL1000005	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC2	2
RL10000050	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC5	5
RL10000051	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC2	2
RL10000052	Q		Food processing environment	Milk and milk products-Sheep--Milk-	CC29	29
RL10000053	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC3	3
RL10000054	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC204	204
RL10000055	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC3	3
RL10000056	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC3	3
RL10000058	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC101	101
RL10000059	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC101	101
RL1000006	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC2	2
RL10000060	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC101	101
RL10000061	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC101	101
RL10000062	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC101	101
RL10000063	Q		Food processing environment	Milk and milk products-Sheep--Milk-	CC7	7
RL10000064	Q		Food processing environment	Milk and milk products-Sheep--Milk-	CC155	155
RL10000065	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC2	2
RL10000066	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC2	2
RL10000067	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC2	2

RL_number	Country	Year	Sample type	Description	MLST Clonal complexes	MLST sequence types
RL1000068	Q		Food processing environment	Milk and milk products-Sheep-- Fresh cheese-	CC2	2
RL1000069	Q		Food processing environment	Milk and milk products-Sheep-- Fresh cheese-	CC2	2
RL1000070	Q		Food processing environment	Milk and milk products-Sheep-- Fresh cheese-	CC2	2
RL1000071	Q		Food processing environment	Milk and milk products-Sheep-- Fresh cheese-	CC2	2
RL1000072	Q		Food processing environment	Milk and milk products-Sheep-- Fresh cheese-	CC9	9
RL1000073	Q		Food processing environment	Milk and milk products-Sheep-- Fresh cheese-	CC9	9
RL1000074	Q		Food processing environment	Milk and milk products-Sheep-- Fresh cheese-	CC2	2
RL1000075	Q		Food processing environment	Milk and milk products-Sheep-- Fresh cheese-	CC9	9
RL1000076	Q		Food processing environment	Milk and milk products-Sheep-- Fresh cheese-	CC2	2
RL1000077	Q		Food processing environment	Milk and milk products-Sheep-- Fresh cheese-	CC3	3
RL1000078	Q		Food processing environment	Milk and milk products-Sheep-- Fresh cheese-	CC2	2
RL1000079	Q		Food processing environment	Milk and milk products-Sheep-- Fresh cheese-	CC2	2
RL1000080	Q		Food processing environment	Milk and milk products-Sheep-- Fresh cheese-	CC2	2
RL1000081	Q		Food processing environment	Milk and milk products-Sheep-- Fresh cheese-	CC3	3
RL1000082	Q		Food processing environment	Milk and milk products-Sheep-- Fresh cheese-	CC2	2
RL1000083	Q		Food processing environment	Milk and milk products-Sheep-- Fresh cheese-	CC3	3
RL1000084	Q		Food processing environment	Milk and milk products-Sheep-- Fresh cheese-	CC2	2
RL1000085	Q		Food processing environment	Milk and milk products-Sheep-- Fresh cheese-	CC2	2
RL1000086	Q		Food processing environment	Milk and milk products-Sheep-- Fresh cheese-	CC9	9
RL1000087	Q		Food processing environment	Milk and milk products-Sheep-- Fresh cheese-	CC9	9
RL1000088	Q		Food processing environment	Milk and milk products-Sheep-- Fresh cheese-	CC9	9
RL1000089	Q		Food processing environment	Milk and milk products-Sheep-- Fresh cheese-	CC101	101
RL1000090	Q		Food processing environment	Milk and milk products-Sheep-- Fresh cheese-	CC2	2
RL1000091	Q		Food processing environment	Milk and milk products-Sheep-- Fresh cheese-	CC101	101
RL1000092	Q		Food processing environment	Milk and milk products-Sheep-- Fresh cheese-	CC101	101
RL1000093	Q		Food processing environment	Milk and milk products-Sheep-- Fresh cheese-	CC101	101
RL1000094	Q		Food processing environment	Milk and milk products-Sheep-- Fresh cheese-	CC101	101
RL1000095	Q		Food processing environment	Milk and milk products-Sheep-- Fresh cheese-	CC101	101
RL1000096	Q		Food processing environment	Milk and milk products-Sheep-- Fresh cheese-	CC101	101

RL_number	Country	Year	Sample type	Description	MLST Clonal complexes	MLST sequence types
RL1000097	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC101	101
RL1000098	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC101	101
RL1000010	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC9	9
RL14000001	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC18	18
RL14000008	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC9	9
RL14000017	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC101	101
RL14000057	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC3	3
RL14000099	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC101	101
RL14000100	Q		Food processing environment	Milk and milk products-Sheep--Fresh cheese-	CC101	101
RL15000361	C	2009	Food products	Meat and meat products-Swine--Meat - Cut-	CC2	2
RL15000362	C	2010	Food products	Meat and meat products-Unspecified--Meat - Cut-	CC6	6
RL15000363	C	2010	Food products	Meat and meat products-Swine--Meat - Cut-	CC5	5
RL15000364	C	2010	Food products	Meat and meat products-Swine--Meat - Cut-	CC2	2
RL15000365	C	2010	Food products	Meat and meat products-Swine--Deli product - Other product-	CC77	77
RL15000366	C	2010	Food products	Meat and meat products-Swine--Meat - Cut-	CC37	37
RL15000367	C	2010	Food products	Meat and meat products-Swine--Deli product - Other product-	CC37	37
RL15000368	C	2010	Food products	Meat and meat products-Swine--Meat - Cut-	CC18	18
RL15000370	C	2010	Food products	Meat and meat products-Swine--Deli product - Ham-	CC121	121
RL15000371	C	2010	Food products	Meat and meat products-Swine--Meat - Cut-	ST191	191
RL15000372	C	2010	Food products	Meat and meat products-Unspecified--Meat - Cut-	CC121	121
RL15000373	C	2011	Food products	Meat and meat products-Unspecified--Meat - Sausage-	CC9	9
RL15000374	C	2011	Food products	Meat and meat products-Swine--Meat - Cut-	ST191	191
RL15000375	C	2011	Food products	Meat and meat products-Swine--Deli product - Ham-	CC9	9
RL15000376	C	2011	Food products	Meat and meat products-Swine--Meat - Cut-	CC5	5
RL15000377	C	2011	Food products	Meat and meat products-Unspecified--Meat - Cut-	CC121	121
RL15000378	C	2011	Food products	Meat and meat products-Swine--Meat - Cut-	CC121	121
RL15000387	C	2008	Food processing environment	Meat and meat products-Swine---	CC8	8
RL15000388	C	2008	Food processing environment	Meat and meat products-Swine---	ST602	602
RL15000389	C	2008	Food processing environment	Meat and meat products-Swine---	CC7	691
RL15000390	C	2004	Food processing environment	Meat and meat products-Swine---	CC121	121
RL15000391	C	2005	Food processing environment	Meat and meat products-Swine---	CC121	121

RL_number	Country	Year	Sample type	Description	MLST Clonal complexes	MLST sequence types
RL15000392	C	2006	Food processing environment	Meat and meat products-Swine---	CC121	121
RL15000393	C	2005	Food processing environment	Meat and meat products-Swine---	CC1	1
RL15000536	C	2011	Food products	Meat and meat products-Swine-- Meat - Cut-	CC1	1
RL15000537	C	2011	Food products	Meat and meat products-Swine-- Deli product - Ham-	CC4	4
RL15000538	C	2008	Food processing environment	Meat and meat products-Swine---	CC101	775
RL15000539	C	2008	Food processing environment	Meat and meat products-Swine---	CC31	31
RL15000540	C	2008	Food processing environment	Meat and meat products-Swine---	CC31	31
RL15000541	C	2008	Food processing environment	Meat and meat products-Swine---	CC5	5
RL15000542	C	2008	Food processing environment	Meat and meat products-Swine---	CC9	9
RL15000543	C	2008	Food processing environment	Meat and meat products-Swine---	CC7	691
RL15000619	B	2002	Food products	Fish and fishery products-Fish- Salmo spp. (Salmon)-Fish origin not specified-	CC121	121
RL15000620	B	2011	Food products	Fish and fishery products-Fish- Salmo spp. (Salmon)-Fish origin not specified-	CC121	121
RL15000621	B	2011	Food processing environment	Fish and fishery products-Fish- Salmo spp. (Salmon)-Fish origin not specified-	CC121	121
RL15000622	B	2011	Food processing environment	Fish and fishery products-Fish- Salmo spp. (Salmon)-Fish origin not specified-	CC121	121
RL15000623	B	2011	Food products	Fish and fishery products-Fish- Salmo spp. (Salmon)-Fish origin not specified-	CC121	121
RL15000624	B	2002	Food products	Fish and fishery products-Fish- Salmo spp. (Salmon)-Fish origin not specified-	CC121	121
RL15000625	B	2012	Food products	Fish and fishery products-Fish- Salmo spp. (Salmon)-Fish origin not specified-	CC121	121
RL15000626	B	2012	Food products	Fish and fishery products-Fish- Salmo spp. (Salmon)-Fish origin not specified-	CC121	121
RL15000627	B	2011	Food products	Fish and fishery products-Fish- Salmo spp. (Salmon)-Fish origin not specified-	CC101	775
RL15000628	B	2011	Food products	Fish and fishery products-Fish- Salmo spp. (Salmon)-Fish origin not specified-	CC101	775
RL15000629	B	2011	Food products	Fish and fishery products-Fish- Salmo spp. (Salmon)-Fish origin not specified-	CC121	121
RL15000630	B	2013	Food products	Milk and milk products-Unspecified-- Cheese category not specified-	CC37	37
RL15000631	B	2013	Food products	Milk and milk products-Unspecified-- Cheese category not specified-	CC37	37
RL15000632	B	2013	Food products	Milk and milk products-Unspecified-- Cheese category not specified-	CC37	37
RL15000633	B	2013	Food products	Milk and milk products-Unspecified-- Cheese category not specified-	CC37	37
RL15000634	B	2013	Food products	Milk and milk products-Unspecified-- Cheese category not specified-	CC37	37

RL_number	Country	Year	Sample type	Description	MLST Clonal complexes	MLST sequence types
RL15000635	B	2013	Food products	Milk and milk products-Unspecified--Cheese category not specified-	CC37	37
RL15000636	B	2013	Food products	Milk and milk products-Unspecified--Cheese category not specified-	CC37	37
RL15000637	B	2013	Food products	Milk and milk products-Unspecified--Cheese category not specified-	CC37	37
RL15000638	B	2013	Food products	Milk and milk products-Unspecified--Cheese category not specified-	CC37	37
RL15000639	B	2013	Food products	Milk and milk products-Unspecified--Cheese category not specified-	CC121	121
RL15000640	B	2013	Food products	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-	CC101	775
RL15000641	B	2013	Food processing environment	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-	CC101	775
RL15000642	B	2013	Food processing environment	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-	CC101	775
RL15000643	B	2013	Food products	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-	CC101	775
RL15000644	B	2013	Food products	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-	CC101	775
RL15000645	B	2013	Food processing environment	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-	CC121	121
RL15000646	B	2013	Food processing environment	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-	CC6	6
RL15000647	B	2013	Food products	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-	CC101	775
RL15000648	B	2013	Food products	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-	CC121	121
RL15000649	B	2013	Food products	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-	CC14	399
RL15000650	B	2013	Food products	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-	CC121	121
RL15000651	B	2014	Food processing environment	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-	CC415	394
RL15000652	B	2014	Food products	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-	CC31	325
RL15000653	B	2014	Food processing environment	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-	CC31	325
RL15000654	B	2014	Food products	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-	CC7	7
RL15000655	B	2014	Food products	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-	CC31	325
RL15000656	B	2014	Food products	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-	CC89	391

RL_number	Country	Year	Sample type	Description	MLST Clonal complexes	MLST sequence types
RL1500657	B	2014	Food products	Fish and fishery products-Fish-Salmo spp. (Salmon)-Fish origin not specified-	CC8	8
RL15001296	B	2010	Food processing environment	Meat and meat products-Unspecified---	CC9	9
RL15001297	B	2010	Food processing environment	Meat and meat products-Mixed animal source---	CC121	121
RL15001298	B	2010	Food processing environment	Milk and milk products-Unspecified-- Cheese category not specified-	CC31	31
RL15001302	B	2010	Food processing environment	Meat and meat products-Unspecified---	CC121	121
RL15001303	B	2010	Food processing environment	Meat and meat products-Unspecified---	CC2	2
RL15001304	B	2010	Food processing environment	Meat and meat products-Unspecified---	CC9	9
RL15001305	B	2010	Food processing environment	Elaborated food products combining several food categories-Mixed sources--Sandwich-	CC2	2
RL15001306	B	2011	Food processing environment	Meat and meat products-Unspecified---	CC9	9
RL15001307	B	2011	Food processing environment	Meat and meat products-Unspecified---	CC6	6
RL15001308	B	2011	Food processing environment	Meat and meat products-Unspecified---	CC20	20
RL15001309	B	2011	Food processing environment	Meat and meat products-Unspecified---	CC31	31
RL15001314	B	2011	Food processing environment	Meat and meat products-Mixed animal source---	CC415	394
RL15001315	B	2011	Food processing environment	Elaborated food products combining several food categories-Mixed sources--Sandwich-	CC121	121
RL15001316	B	2011	Food processing environment	Elaborated food products combining several food categories-Mixed sources--Sandwich-	CC20	20
RL15001317	B	2011	Food processing environment	Elaborated food products combining several food categories-Mixed sources--Sandwich-	CC2	2
RL15001318	B	2011	Food processing environment	Meat and meat products-Unspecified---	CC204	204
RL15001319	B	2011	Food processing environment	Milk and milk products-Unspecified-- Ice cream-	CC220	220
RL15001320	B	2011	Food processing environment	Milk and milk products-Unspecified-- Cheese category not specified-	CC37	37
RL15001321	B	2011	Food processing environment	Milk and milk products-Unspecified-- Cheese category not specified-	CC8	8
RL15001322	B	2011	Food processing environment	Milk and milk products-Unspecified-- Cheese category not specified-	CC37	37
RL15001323	B	2011	Food processing environment	Milk and milk products-Unspecified-- Cheese category not specified-	CC8	8
RL15001324	B	2011	Food processing environment	Milk and milk products-Unspecified-- Cheese category not specified-	CC37	37
RL15001325	B	2011	Food processing environment	Milk and milk products-Unspecified-- Cheese category not specified-	CC37	37
RL15001326	B	2011	Food processing environment	Milk and milk products-Unspecified-- Cheese category not specified-	CC224	224
RL15001327	B	2011	Food processing environment	Milk and milk products-Unspecified-- Ice cream-	CC8	8
RL15001350	B	2011	Food processing environment	Meat and meat products-Unspecified---	CC14	91
RL15001351	B	2011	Food processing environment	Meat and meat products-Mixed animal source---	CC2	2

RL_number	Country	Year	Sample type	Description	MLST Clonal complexes	MLST sequence types
RL15001354	B	2011	Food processing environment	Elaborated food products combining several food categories-Mixed sources--Sandwich-	CC204	204
RL15001386	B	2010	Food processing environment	Milk and milk products-Unspecified--Cheese category not specified-	CC403	403

Appendix 5: Isolates from sporadic clinical cases

RL_number	Country	Year	Clinical symptoms	MLST Clonal complexes	MLST sequence types
RL15000306	Q	2011	Unknown	CC101	101
RL15000307	Q	2010	Unknown	CC101	38
RL15000308	Q	2010	Unknown	CC101	38
RL15000309	Q	2010	Unknown	CC6	6
RL15000310	Q	2010	Unknown	CC101	38
RL15000311	Q	2010	Unknown	CC398	398
RL15000312	Q	2010	Unknown	CC101	38
RL15000313	Q	2010	Unknown	CC1	1
RL15000314	Q	2010	Unknown	CC9	9
RL15000315	Q	2010	Unknown	CC8	8
RL15000316	Q	2010	Unknown	CC101	38
RL15000317	Q	2010	Unknown	CC101	38
RL15000318	Q	2010	Unknown	ST560	560
RL15000319	Q	2010	Unknown	CC101	38
RL15000320	Q	2010	Unknown	CC7	7
RL15000321	Q	2010	Unknown	CC1	1
RL15000322	Q	2010	Unknown	CC6	6
RL15000323	Q	2010	Unknown	CC8	8
RL15000324	Q	2010	Unknown	CC5	5
RL15000325	Q	2010	Unknown	CC3	287
RL15000326	Q	2010	Unknown	CC1	1
RL15000327	Q	2010	Unknown	CC101	38
RL15000328	Q	2010	Unknown	CC37	37
RL15001278	B	2010	Unknown	CC6	178
RL15001279	B	2010	Bacteremia	CC9	9
RL15001280	B	2010	Pregnancy related	CC6	6
RL15001284	B	2010	Meningitis	CC1	1
RL15001285	B	2010	Bacteremia	CC1	1
RL15001286	B	2010	Unknown	CC1	1
RL15001287	B	2010	Meningitis	CC1	1
RL15001288	B	2010	Other	ST736	736
RL15001289	B	2010	Meningitis	CC1	1
RL15001290	B	2010	Other	CC8	16
RL15001291	B	2010	Bacteremia	CC7	7
RL15001310	B	2011	Bacteremia	CC54	54
RL15001328	B	2011	Unknown	CC220	220
RL15001329	B	2011	Bacteremia	CC2	2

RL_number	Country	Year	Clinical symptoms	MLST Clonal complexes	MLST sequence types
RL15001330	B	2011	Bacteremia	CC1	1
RL15001331	B	2011	Unknown	CC220	220
RL15001338	B	2011	Bacteremia	CC4	4
RL15001339	B	2011	Unknown	CC6	6
RL15001340	B	2011	Other	CC1	1
RL15001341	B	2011	Bacteremia	CC1	1
RL15001342	B	2011	Meningitis	CC20	20
RL15001343	B	2011	Bacteremia	CC6	6
RL15001344	B	2011	Meningitis	CC2	2
RL15001345	B	2011	Bacteremia	CC6	6
RL15001374	B	2009	Bacteremia	CC121	121
RL15001375	B	2011	Meningitis	CC2	2
RL15001376	B	2011	Bacteremia	CC37	37
RL15001378	B	2009	Bacteremia	CC2	2
RL15001379	B	2010	Bacteremia	CC59	59
RL15001380	B	2010	Meningitis	CC7	7
RL15001381	B	2010	Meningitis	ST392	392
RL15001414	A	2010	Bacteremia	CC1	1
RL15001415	A	2010	Bacteremia	CC155	155
RL15001416	A	2010	Bacteremia	CC155	155
RL15001417	A	2010	Bacteremia	CC6	6
RL15001418	A	2010	Bacteremia	CC11	451
RL15001419	A	2010	Bacteremia	CC18	18
RL15001420	A	2010	Bacteremia	CC6	6
RL15001421	A	2010	Bacteremia	CC1	1
RL15001422	A	2010	Bacteremia	CC1	1
RL15001423	A	2010	Bacteremia	CC6	6
RL15001424	A	2011	Bacteremia	CC9	9
RL15001425	A	2011	Other	CC9	9
RL15001426	A	2011	Other	CC1	1
RL15001427	A	2011	Meningitis	CC101	101
RL15001522	A	2011	Bacteremia	CC6	179
RL15001523	A	2011	Bacteremia	CC9	9
RL15001524	A	2011	Bacteremia	CC37	37
RL15001525	A	2011	Bacteremia	CC1	1
RL15001526	A	2011	Bacteremia	CC1	1
RL15001527	F	2010	Other	CC5	5
RL15001528	F	2010	Pregnancy related	CC101	101
RL15001529	F	2010	Meningitis	CC2	2
RL15001530	F	2010	Unknown	CC3	3
RL15001531	F	2011	Meningitis	CC155	155

RL_number	Country	Year	Clinical symptoms	MLST Clonal complexes	MLST sequence types
RL15001532	F	2011	Unknown	CC2	2
RL15001533	F	2011	Meningitis	CC1	1
RL15001534	F	2011	Unknown	CC3	3
RL15001535	W	2011	Bacteremia	CC3	3
RL15001536	W	2010	Meningitis	ST570	570
RL15001537	W	2010	Bacteremia	CC9	9
RL15001538	W	2010	Bacteremia	CC9	9
RL15001539	W	2010	Meningitis	CC7	7
RL15001540	W	2010	Other	CC6	6
RL15001541	W	2011	Other	ST32	32
RL15001542	W	2010	Meningitis	CC26	26
RL15001543	W	2010	Meningitis	CC21	21
RL15001544	W	2011	Bacteremia	CC4	4
RL15001545	W	2011	Bacteremia	CC8	8
RL15001546	W	2010	Meningitis	ST184	184
RL15001547	W	2010	Meningitis	CC2	2
RL15001548	W	2010	Other	CC2	2
RL15001549	W	2010	Bacteremia	CC37	37
RL15001550	Y	2011	Unknown	CC1	1
RL15001551	D	2011	Meningitis	CC21	21
RL15001552	Y	2011	Unknown	CC1	1
RL15001553	Y	2011	Unknown	CC8	8
RL15001554	Y	2011	Unknown	CC8	8
RL15001555	Y	2011	Unknown	CC18	18
RL15001556	Y	2011	Unknown	CC398	398
RL15001557	Y	2011	Unknown	CC59	59
RL15001558	Y	2011	Unknown	CC14	399
RL15001559	Y	2011	Unknown	CC14	14
RL15001560	Y	2011	Unknown	CC155	155
RL15001561	Y	2011	Unknown	CC8	8
RL15001562	Y	2011	Unknown	CC11	451
RL15001563	T	2010	Bacteremia	CC31	31
RL15001564	T	2010	Bacteremia	CC1	1
RL15001565	T	2010	Bacteremia	CC20	20
RL15001566	T	2010	Bacteremia	CC7	7
RL15001567	T	2010	Bacteremia	CC101	101
RL15001568	T	2010	Bacteremia	CC7	7
RL15001569	T	2010	Bacteremia	CC1	1
RL15001570	T	2010	Bacteremia	CC14	399
RL15001571	T	2010	Bacteremia	CC14	399
RL15001572	T	2011	Bacteremia	CC7	7

RL_number	Country	Year	Clinical symptoms	MLST Clonal complexes	MLST sequence types
RL15001573	T	2011	Bacteremia	CC121	121
RL15001574	T	2011	Unknown	CC177	180
RL15001575	T	2011	Bacteremia	CC14	399
RL15001576	T	2011	Bacteremia	CC8	8
RL15001577	T	2011	Bacteremia	CC14	91
RL15001578	T	2011	Meningitis	CC7	7
RL15001579	T	2011	Bacteremia	CC9	9
RL15001580	T	2011	Bacteremia	CC8	8
RL15001581	T	2011	Bacteremia	CC121	121
RL15001582	T	2011	Bacteremia	CC8	8
RL15001583	X	2010	Other	CC155	155
RL15001584	X	2010	Unknown	CC1	1
RL15001585	X	2010	Bacteremia	CC1	1
RL15001586	X	2010	Meningitis	CC1	181
RL15001587	X	2010	Bacteremia	CC6	6
RL15001588	X	2010	Meningitis	CC5	5
RL15001589	X	2010	Pregnancy related	CC14	14
RL15001590	X	2010	Other	CC6	6
RL15001591	X	2010	Other	CC1	1
RL15001592	X	2010	Meningitis	CC155	155
RL15001593	X	2010	Bacteremia	CC8	16
RL15001594	X	2010	Bacteremia	CC87	87
RL15001595	X	2010	Bacteremia	CC8	8
RL15001596	X	2010	Bacteremia	CC6	6
RL15001597	X	2011	Other	CC37	37
RL15001598	X	2011	Meningitis	CC1	1
RL15001599	X	2011	Meningitis	CC1	1
RL15001600	X	2011	Bacteremia	CC9	9
RL15001601	X	2011	Bacteremia	CC1	1
RL15001602	X	2011	Pregnancy related	CC388	388
RL15001603	X	2011	Pregnancy related	CC87	87
RL15001604	X	2011	Bacteremia	CC7	7
RL15001605	X	2011	Bacteremia	CC8	8
RL15001606	X	2011	Bacteremia	CC54	54
RL15001607	X	2011	Other	CC204	204
RL15001608	X	2011	Bacteremia	CC1	1
RL15001609	X	2011	Bacteremia	CC1	1
RL15001610	X	2011	Meningitis	CC1	1
RL15001611	X	2011	Other	CC7	7
RL15001612	X	2011	Bacteremia	CC7	7
RL15001613	X	2011	Bacteremia	CC87	87

RL_number	Country	Year	Clinical symptoms	MLST Clonal complexes	MLST sequence types
RL15001614	X	2011	Meningitis	CC7	7
RL15001615	X	2011	Meningitis	CC1	1
RL15001616	X	2011	Meningitis	CC9	9
RL15001617	X	2011	Other	CC1	1
RL15001618	D	2010	Pregnancy related	CC7	7
RL15001619	D	2010	Bacteremia	CC9	9
RL15001620	D	2010	Bacteremia	CC101	101
RL15001621	D	2010	Meningitis	CC14	91
RL15001622	D	2010	Pregnancy related	CC2	2
RL15001623	D	2010	Bacteremia	CC4	4
RL15001624	D	2010	Bacteremia	CC11	451
RL15001625	D	2010	Pregnancy related	CC2	2
RL15001626	D	2010	Bacteremia	CC3	3
RL15001627	D	2010	Bacteremia	CC121	121
RL15001628	D	2011	Bacteremia	CC4	4
RL15001629	D	2011	Bacteremia	CC8	8
RL15001630	D	2011	Other	CC379	182
RL15001631	D	2011	Bacteremia	CC7	7
RL15001632	D	2011	Bacteremia	CC11	451
RL15001633	D	2011	Meningitis	CC3	3
RL15001634	D	2011	Other	CC1	1
RL15001635	D	2011	Bacteremia	CC398	398
RL15001636	Z	2010	Other	CC1	1
RL15001637	Z	2010	Pregnancy related	CC2	2
RL15001638	Z	2010	Unknown	CC8	8
RL15001639	Z	2010	Unknown	CC1	515
RL15001640	Z	2010	Meningitis	CC8	8
RL15001641	Z	2010	Meningitis	CC1	1
RL15001642	Z	2010	Unknown	CC8	8
RL15001643	Z	2010	Unknown	CC6	6
RL15001644	Z	2010	Unknown	CC6	6
RL15001645	Z	2010	Unknown	CC2	2
RL15001646	Z	2011	Other	CC379	808
RL15001647	Z	2011	Meningitis	CC2	2
RL15001648	Z	2011	Meningitis	CC87	87
RL15001649	Z	2011	Meningitis	CC4	4
RL15001650	Z	2011	Unknown	CC9	9
RL15001651	Z	2011	Bacteremia	CC224	224
RL15001652	Z	2011	Bacteremia	CC6	6
RL15001653	Z	2011	Meningitis	CC8	120
RL15001654	Z	2011	Unknown	ST773	773

RL_number	Country	Year	Clinical symptoms	MLST Clonal complexes	MLST sequence types
RL15001655	Z	2011	Meningitis	CC6	6
RL15001737	D	2011	Other	CC6	6
RL15002422	C	2010	Bacteremia	CC101	101
RL15002423	C	2010	Meningitis	CC1	1
RL15002424	C	2010	Bacteremia	CC224	224
RL15002425	C	2010	Meningitis	CC1	1
RL15002426	C	2010	Bacteremia	CC59	59
RL15002427	C	2010	Pregnancy related	CC1	1
RL15002428	C	2010	Pregnancy related	CC6	6
RL15002429	C	2010	Bacteremia	CC2	2
RL15002430	C	2010	Bacteremia	CC101	101
RL15002431	C	2010	Bacteremia	CC6	6
RL15002432	C	2010	Meningitis	CC8	8
RL15002433	C	2010	Bacteremia	CC4	4
RL15002434	C	2010	Bacteremia	CC4	4
RL15002435	C	2010	Bacteremia	CC5	5
RL15002436	C	2010	Bacteremia	CC155	155
RL15002437	C	2010	Bacteremia	CC2	2
RL15002438	C	2010	Pregnancy related	CC1	1
RL15002439	C	2010	Meningitis	CC6	6
RL15002440	C	2010	Bacteremia	CC9	9
RL15002441	C	2010	Meningitis	CC1	1
RL15002442	C	2011	Meningitis	CC54	54
RL15002443	C	2011	Meningitis	CC4	4
RL15002444	C	2011	Bacteremia	CC224	224
RL15002445	C	2011	Bacteremia	CC1	1
RL15002446	C	2011	Meningitis	CC2	2
RL15002447	C	2011	Bacteremia	CC18	18
RL15002448	C	2011	Pregnancy related	CC1	1
RL15002449	C	2011	Meningitis	CC6	6
RL15002450	C	2011	Bacteremia	CC6	6
RL15002451	C	2011	Bacteremia	CC4	4
RL15002452	C	2011	Meningitis	CC6	6
RL15002453	C	2011	Bacteremia	CC121	121
RL15002454	C	2011	Bacteremia	CC6	6
RL15002455	C	2011	Pregnancy related	CC1	1
RL15002456	C	2011	Meningitis	CC1	1
RL15003001	Y	2010	Bacteremia	CC8	8
RL15003002	Y	2010	Unknown	CC7	7
RL15003003	Y	2010	Unknown	CC1	1
RL15003004	Y	2010	Unknown	CC4	4

RL_number	Country	Year	Clinical symptoms	MLST Clonal complexes	MLST sequence types
RL15003005	Y	2010	Unknown	CC101	101
RL15003006	Y	2010	Unknown	CC398	398
RL15003007	Y	2010	Unknown	CC7	7
RL15003008	Y	2010	Unknown	CC8	8
RL15003015	A	2010	Bacteremia	CC8	8
RL15003016	A	2010	Bacteremia	CC59	59
RL15003017	A	2010	Bacteremia	CC9	9
RL15003018	A	2010	Meningitis	CC2	2
RL15003019	A	2010	Bacteremia	CC121	121
RL15003020	A	2010	Meningitis	CC2	2
RL15003021	A	2010	Pregnancy related	CC2	2
RL15003022	A	2010	Pregnancy related	CC2	2
RL15003023	A	2011	Bacteremia	CC155	155
RL15003024	A	2011	Bacteremia	CC14	399
RL15003025	A	2011	Other	CC8	8
RL15003026	A	2011	Meningitis	CC1	1
RL15003027	A	2011	Bacteremia	CC1	1
RL15003028	A	2011	Bacteremia	CC8	8
RL15003029	A	2011	Meningitis	CC6	*6
RL15003030	A	2011	Bacteremia	CC18	18

Appendix 6: Isolates from outbreaks

RL_number	Country	Year	Food/Human /Environment	Clinical symptoms ^a	Outbreak number	MLST Clonal complexes	MLST sequence types
RL15001313	B	2011	Food	Not applicable	Outbreak 4	CC59	59
RL15001336	B	2011	Human	Bacteremia	Outbreak 4	CC59	59
RL15001337	B	2011	Food	Not applicable	Outbreak 4	CC59	59
RL15001355	B	2012	Human	Other	Outbreak 1	CC155	155
RL15001356	B	2012	Human	Other	Outbreak 1	CC155	155
RL15001357	B	2012	Food	Not applicable	Outbreak 1	CC155	155
RL15001358	B	2012	Human	Other	Outbreak 1	CC155	155
RL15001359	B	2012	Food	Not applicable	Outbreak 1	CC155	155
RL15001360	B	2012	Human	Bacteremia	Outbreak 1	CC155	155
RL15001361	B	2012	Food	Not applicable	Outbreak 1	CC155	155
RL15001362	B	2012	Environment	Not applicable	Outbreak 1	CC155	155
RL15001363	B	2012	Environment	Not applicable	Outbreak 1	CC155	155
RL15001364	B	2012	Food	Not applicable	Outbreak 1	CC155	155
RL15001366	B	2012	Environment	Not applicable	Outbreak 1	CC155	155
RL15001367	B	2013	Human	Bacteremia	Outbreak 1	CC155	155
RL15001368	B	2014	Human	Unknown	Outbreak 5	CC475	504
RL15001369	B	2013	Food	Not applicable	Outbreak 1	CC155	155
RL15001370	B	2014	Food	Not applicable	Outbreak 3	CC7	7
RL15001371	B	2014	Human	Other	Outbreak 5	CC415	394
RL15001372	B	2014	Food	Not applicable	Outbreak 3	CC7	7
RL15001377	B	2009	Human	Bacteremia	Outbreak 4	CC59	59
RL15001390	B	2007	Human	Bacteremia	Outbreak 2	CC1	746
RL15001391	B	2008	Human	Other	Outbreak 2	CC1	1
RL15001392	B	2011	Human	Unknown	Outbreak 2	CC1	1
RL15001393	B	2011	Food	Not applicable	Outbreak 2	CC1	1
RL15001394	B	2012	Food	Not applicable	Outbreak 5	CC415	394
RL15001395	B	2013	Food	Not applicable	Outbreak 3	CC7	*7
RL15001396	B	2013	Human	Bacteremia	Outbreak 3	CC7	7
RL15001397	B	2013	Human	Bacteremia	Outbreak 2	CC1	1
RL15001398	B	2013	Food	Not applicable	Outbreak 3	CC7	7
RL15001399	B	2013	Human	Unknown	Outbreak 2	CC1	1
RL15001400	B	2013	Human	Unknown	Outbreak 5	CC415	394
RL15001401	B	2013	Food	Not applicable	Outbreak 2	CC1	1
RL15001402	B	2013	Food	Not applicable	Outbreak 2	CC1	1
RL15001403	B	2013	Human	Bacteremia	Outbreak 5	CC415	394
RL15001404	B	2013	Human	Bacteremia	Outbreak 5	CC415	394
RL15001405	B	2013	Human	Unknown	Outbreak 5	CC415	394

RL_number	Country	Year	Food/Human /Environment	Clinical symptoms ^a	Outbreak number	MLST Clonal complexes	MLST sequence types
RL15001406	B	2013	Human	Bacteremia	Outbreak 5	CC415	394
RL15001407	B	2013	Human	Unknown	Outbreak 5	CC415	394
RL15001408	B	2014	Human	Other	Outbreak 3	CC7	7
RL15001409	B	2014	Human	Other	Outbreak 5	CC415	394
RL15001410	B	2014	Human	Unknown	Outbreak 3	CC7	7
RL15001411	B	2014	Human	Unknown	Outbreak 3	CC7	7
RL15001656	X	2012	Human	Pregnancy related	Outbreak 8	CC14	14
RL15001657	X	2012	Human	Pregnancy related	Outbreak 8	CC14	14
RL15001658	X	2012	Human	Meningitis	Outbreak 8	CC14	14
RL15001659	X	2012	Human	Pregnancy related	Outbreak 8	CC14	14
RL15001660	X	2012	Food	Not applicable	Outbreak 8	CC14	14
RL15001661	X	2012	Food	Not applicable	Outbreak 8	CC14	14
RL15001662	X	2012	Food	Not applicable	Outbreak 8	CC14	14
RL15001663	X	2012	Food	Not applicable	Outbreak 8	CC14	14
RL15001664	X	2012	Food	Not applicable	Outbreak 8	CC14	14
RL15001665	X	2012	Food	Not applicable	Outbreak 8	CC14	14
RL15001666	X	2012	Food	Not applicable	Outbreak 8	CC14	14
RL15001667	X	2012	Food	Not applicable	Outbreak 8	CC14	14
RL15001668	X	2012	Food	Not applicable	Outbreak 8	CC14	14
RL15001669	X	2013	Human	Pregnancy related	Outbreak 7	CC87	87
RL15001670	X	2013	Human	Pregnancy related	Outbreak 7	CC87	87
RL15001671	X	2013	Human	Pregnancy related	Outbreak 7	CC87	87
RL15001672	X	2013	Human	Bacteremia	Outbreak 7	CC87	87
RL15001673	X	2013	Human	Meningitis	Outbreak 7	CC87	87
RL15001674	X	2013	Human	Pregnancy related	Outbreak 7	CC87	87
RL15001675	X	2013	Human	Pregnancy related	Outbreak 7	CC87	87
RL15001676	X	2014	Human	Bacteremia	Outbreak 7	CC87	87
RL15001677	X	2013	Human	Bacteremia	Outbreak 7	CC87	87
RL15001678	X	2013	Human	Bacteremia	Outbreak 7	CC87	87
RL15001679	X	2014	Human	Pregnancy related	Outbreak 7	CC87	87
RL15001680	X	2014	Human	Pregnancy related	Outbreak 7	CC87	87
RL15001681	X	2014	Human	Pregnancy related	Outbreak 7	CC87	87
RL15001682	X	2014	Food	Not applicable	Outbreak 7	CC87	87
RL15001683	X	2014	Food	Not applicable	Outbreak 7	CC87	87
RL15001684	X	2014	Food	Not applicable	Outbreak 7	CC87	87
RL15001685	X	2014	Food	Not applicable	Outbreak 7	CC87	87
RL15001686	X	2014	Food	Not applicable	Outbreak 7	CC87	87

RL_number	Country	Year	Food/Human /Environment	Clinical symptoms ^a	Outbreak number	MLST Clonal complexes	MLST sequence types
RL15001687	X	2014	Food	Not applicable	Outbreak 7	CC87	87
RL15002457	C	2012	Human	Other	Outbreak 9	CC4	4
RL15002458	C	2012	Human	Meningitis	Outbreak 9	CC4	4
RL15002459	C	2012	Human	Meningitis	Outbreak 9	CC4	4
RL15002460	C	2012	Human	Other	Outbreak 9	CC4	4
RL15002461	C	2012	Human	Bacteremia	Outbreak 9	CC4	4
RL15002462	C	2012	Human	Meningitis	Outbreak 9	CC4	4
RL15002463	C	2012	Human	Pregnancy related	Outbreak 9	CC4	4
RL15002464	C	2012	Human	Pregnancy related	Outbreak 9	CC4	4
RL15002465	C	2012	Human	Pregnancy related	Outbreak 9	CC4	4
RL15002466	C	2012	Human	Bacteremia	Outbreak 9	CC4	4
RL15002467	C	2012	Human	Bacteremia	Outbreak 9	CC4	4
RL15002468	C	2012	Human	Meningitis	Outbreak 9	CC4	4
RL15002469	C	2012	Human	Meningitis	Outbreak 9	CC217	217
RL15002470	C	2012	Human	Pregnancy related	Outbreak 9	CC4	4
RL15002471	C	2012	Human	Bacteremia	Outbreak 9	CC4	4
RL15002472	C	2012	Human	Pregnancy related	Outbreak 9	CC4	4
RL15002473	C	2012	Human	Pregnancy related	Outbreak 9	CC4	170
RL15002474	C	2012	Human	Bacteremia	Outbreak 9	CC4	4
RL15002475	C	2012	Human	Meningitis	Outbreak 9	CC4	4
RL15002476	C	2012	Human	Pregnancy related	Outbreak 9	CC4	4
RL15002477	C	2012	Human	Bacteremia	Outbreak 9	CC4	4
RL15002478	C	2012	Human	Meningitis	Outbreak 9	CC4	4
RL15002479	C	2012	Human	Bacteremia	Outbreak 9	CC4	4
RL15002480	C	2012	Human	Meningitis	Outbreak 9	CC4	4
RL15002481	C	2012	Human	Meningitis	Outbreak 9	CC4	4
RL15003010	T	2013	Human	Meningitis	Outbreak 6	CC398	802
RL15003011	T	2013	Food	Not applicable	Outbreak 6	CC398	802
RL15003012	T	2013	Human	Other	Outbreak 6	CC398	802
RL15003013	T	2013	Human	Bacteremia	Outbreak 6	CC398	802
RL15003014	T	2013	Human	Meningitis	Outbreak 6	CC398	802

Clinical symptoms are not relevant for food or environmental isolates.

Appendix 7: Rarefaction and Simpson's diversity index of 7 locus MLST clinical data stratified by age

Simpson's diversity index was determined for isolates from humans <60 years or >60 years old, respectively (Figure 21.1). Both age groups were equally diverse (Simpson's index = 0.881 ± 0.058 for <60 year old and 0.923 ± 0.032 for >60 years old, respectively) ($P > 0.05$).

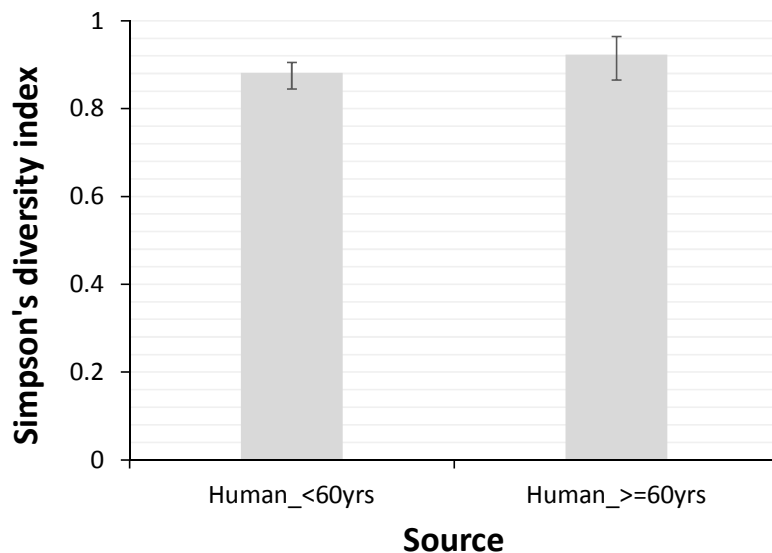


Figure 21.1.: Simpson's diversity index of human isolates based on 7 locus MLST data stratified by age

Rarefaction curves of clinical isolates stratified by age were plotted for 7 locus MLST data (Figure 21.2). The number of new STs per genome is similar (i.e. the two rarefaction curves appear to be virtually identical) for both age groups (either ≥ 60 or < 60 years old).

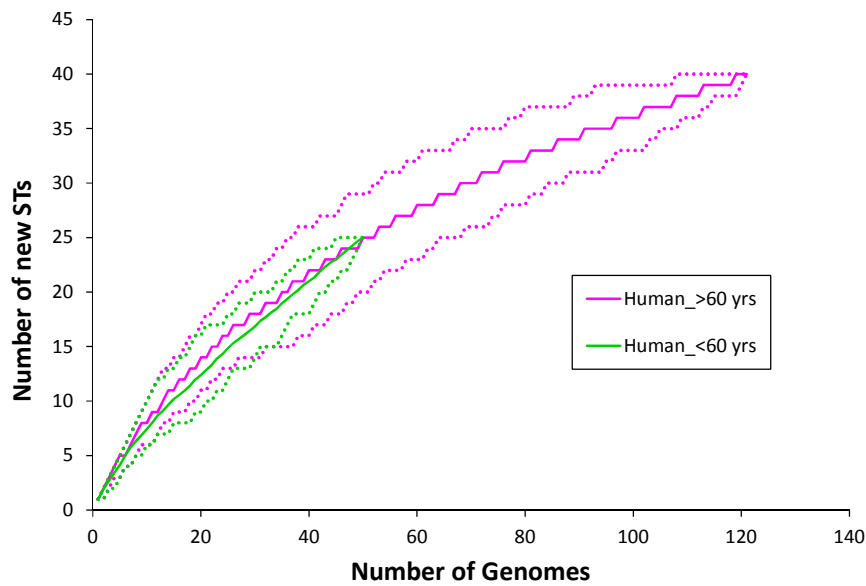


Figure 21.2.: Rarefaction of human isolates based on 7 locus MLST data stratified by age

Nei's genetic distance was determined between humans < 60 and ≥ 60 years of age. It was found to be 0.28 (95% CI's 0.22-0.35). The distance between the two age groups was not significant by randomisation test ($P=0.141$).

Appendix 8: Links to Attribution model software

There were five attribution models that were used in the study. The github link to the software for these programs is:

<https://github.com/lguillier/LISEQ-codes/tree/master/Chapter7>

Listed below are brief instructions on how to access and utilise these.

The Dutch model

An example of the Dutch model for 30 locus rMLST, 864 isolates and 10,000 runs, is given in

"DutchModelrMLST_v5_SampleSizeCorrection_Attribution_Human.7z"

This needs to be extracted using ZIP software. This will produce a .xlsm Excel file.

The program runs under VBA Excel. The input data have to be placed in spreadsheet "Program", starting with column "X" in the format given in the example.

The attribution scores will be displayed in the columns J,K,L,... depending on the number of sources.

These scores have to be copied to the spreadsheet "Results" where the attribution graphic is displayed.

The modified Hald model

The modified Hald model runs under WinBugs which can be downloaded from

<http://www.mrc-bsu.cam.ac.uk/software/bugs/the-bugs-project-winbugs/>

The prevalence sub-model,

"Attribution_Listeria_Prevalence_Hald_EFSA.odc"

has to run first and the results have to be fed into the main model,

"Attribution_Listeria_RealModel_Hald_EFSA.odc"

STRUCTURE model

The program can be downloaded from:

<http://pritchardlab.stanford.edu/structure.html>

The Asymmetric Island (AI) model (iSource)

The program can be downloaded from:

<http://www.danielwilson.me.uk/iSource.html>

The Aberdeen model

The model was implemented in Mathematica and it can be run in any Linux system with Mathematica installed. An example of the model and associated files ("AbdnAttribution_Mathematica.tgz"). In the following instructions on how to install the files and to run the model are given.

*** Installation:

- Uncompress the file in the directory you wish. On a Unix shell,

```
$ tar -xzf AbdnAttribution_Mathematica.tgz
```

- Enter in the program directory:

```
$ cd AbdnAttribution_Mathematica
```

- The package contains the following:

- AbdnAttribution.m : Mathematica script to run the Aberdeen attribution method. It does not need to be edited. Make sure that AbdnAttribution.m is executable. This can be achieved by executing the following command once in the unix shell:

```
$ chmod a+x AbdnAttribution.m
```

- Input_AbdnAttribution.ini : Editable file containing the setting to run source attribution.

- Directory MLST with the data used for Listeria source attribution based on MLST information. This will be used as a benchmark to illustrate the functioning of the program below.

*** Initialisation:

Open Input_AbdnAttribution.ini. This can be done with any text editor.

The format of the file is:

```
1 "***** Parameters for source attribution with AbdnAttribution.m *****"
2 "*****"
3 "***** Please note: only lines 6, 8 and 10 should be edited! *****"
4 "*****"
5 "---- Data directory/file ----"
6   MLST/MLST_AbdnAttribution
7 "---- Reservoir to be attributed ----"
8   Poultry
9 "---- Number of iterations for sample size correction ----"
10  10000
```

Line 6: indicates [data directory]/[data file], i.e. the directory where the input and output data is stored (MLST in our example) and the name of the file containing the data (MLST_AbdnAttribution).

Line 8: Reservoir whose samples will be attributed. In the Listeria dataset, these can be: Bovine, Fish, Human, Ovine, Poultry, Swine

Line 10: Number of iterations for sample size correction.

*** Input data format (see example in the directory MLST, file MLST_AbdnAttribution.csv)

The input data is stored in a comma-separated-values (csv) file.

Each line of the file contains:

[Reservoir name], loci 1, loci 2,, loci n

The isolates for each reservoir type should be consecutive in the file. However, the particular order of reservoirs does not matter. For instance, in MLST_AbdnAttribution.csv they were grouped in the following order: Bovine, Fish, Human, Ovine, Poultry, Swine

*** Running the code:

```
$ ./AbdnAttribution.m
```

*** Output results:

Results are output to an *.xml file which contains the name of the attributed reservoir (can be directly open with Excel or Libreoffice calc).

The output file is stored in [data directory]/[data file]. In the MLST example, it is MLST_AbdnAttribution_Poultry.xls